OLIY TA'LIM, FAN VA INNOVATSIYALAR VAZIRLIGI

ALISHER NAVOIY NOMIDAGI TOSHKENT DAVLAT O'ZBEK TILI VA ADABIYOTI UNIVERSITETI

"KOMPYUTER LINGVISTIKASI: MUAMMOLAR, YECHIM VA ISTIQBOLLAR"

mavzusidagi V an'anaviy xalqaro ilmiy-amaliy konferensiya materiallari

O'zbek tilining ta'limiy korpusi

Turkiy tillarning lingvodidaktik elektron platformasi

O'zbek tilining morfologik analizatori

2025-yil 24-aprel

UO‘K: 004.8+81’32
KBK: 81.1

Kompyuter lingvistikasi: muammolar, yechim va istiqbollar / Xalqaro ilmiy-amaliy konferensiya to‘plami. Elektron nashr / ebook. – Toshkent: ToshDO‘TAU, 24.04.2025. – 840 b.

Mazkur to‘plamdan kompyuter lingvistikasi sohasiga doir tadqiqot natijalari, filologik dasturiy ta’minotlar, ularning imkoniyatlari va afzalliklari, tabiiy tilni qayta ishlashga oid masalalar, semantik tizim bazalari, til va adabiy korpuslar, lingvistik ta’minot manbalari, mashina tarjimasida lingvistik muammolar, lingvodidaktik dasturlar va platformalar, terminologik lug‘atlarni yaratish masalalariga oid materiallar joy olgan.

Ushbu ilmiy to‘plamdan kompyuter lingvistikasi sohasi vakillari, tadqiqotchilar, mustaqil izlanuvchilar, magistrantlar va kelajak sohalariga qiziquvchi talabalar foydalanishlari mumkin.

Mas’ul muharrir:

Sirojiddinov Sh.S. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti rektori, akademik

Taqrizchilar:

Normamatov S.E. Toshkent davlat o‘zbek tili va adabiyoti universiteti, professor, f.f.d.

Eşref Adalı. Istanbul texnika universiteti, professor, t.f.d.

TASHKILIY QO‘MITA

Sirojiddinov Shuhrat Samariddinovich (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Normuradova Nozliya Zarilovna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Eşref Adalı (Turkiya, Istanbul texnika universiteti)

Elov Botir Boltayevich (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Hamroyeva Shahlo Mirdjonovna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Abdullayeva Oqila Xolmo‘minovna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

Xusainova Zilola Yuldashevna (Toshkent davlat o‘zbek tili va adabiyoti universiteti)

TAHRIR HAY’ATI

Eşref Adalı

Elov Botir Boltayevich

Hamroyeva Shahlo Mirdjonovna

Abdullayeva Oqila Xolmo‘minovna

Xusainova Zilola Yuldashevna

Xudayberganov Nizomaddin Uktambay o‘g‘li

Assalomu alaykum, hurmatli konferensiya ishtirokchilari, qadrli mehmonlar!

Mamlakatimizda amalga oshirilayotgan keng ko‘lamli islohotlar jarayonida davlat tilining hayotimizdagi o‘rni va nufuzi tobora oshib bormoqda. O‘zbek tili siyosiy, ijtimoiy-iqtisodiy, ma’naviy-ma’rifiy jabhalarda faol qo‘llanib, xalqaro minbarlarda baralla yangramoqda. Xorijiy mamlakatlarda tilimizga e’tibor va uni o‘rganishga qiziqish kuchaymoqda. Bugungi globallashuv davrida har bir xalq, har qaysi mustaqil davlat o‘z milliy manfaatlarini ta’minlash, bu borada avvalo o‘z madaniyatini, azaliy qadriyatlarini, ona tilini asrab-avaylash va rivojlantirish masalasiga ustuvor ahamiyat qaratishi tabiiy.

O‘zbekiston Respublikasi Prezidentining 2019 yil 21-oktyabrdagi “O‘zbek tilining davlat tili sifatidagi nufuzi va mavqeini tubdan oshirish chora-tadbirlari to‘g‘risida”gi PF-5850-sonli Farmoni, 2020 yil 20-oktyabrdagi “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida”gi PF-6084-sonli Farmonlari bu vazifalarni amalga oshirishni yanada jadallashtirdi.

Bugungi kunda mamlakatimiz ijtimoiy hayotining barcha sohalarida zamonaviy va innovatsion texnologiyalarning imkoniyatlaridan to‘liq va to‘g‘ri foydalanishga erishish; ta’lim muassasalarida davlat tilini o‘rgatish va uni bilish darajasini baholash tizimini yanada takomillashtirish, davlat tilining ilm-fan tili sifatidagi nufuzini oshirish; davlat tili sofligini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining axborot va kommunikatsiya texnologiyalari, xususan, Internet jahon axborot tarmog‘ida munosib o‘rin egallashini ta’minlash, o‘zbek tilining kompyuter dasturlarini yaratish bo‘yicha bosqichma-bosqich tizimli ishlar amalga oshirilmoqda.

Bugun universitetimizda tashkil etilgan “Kompyuter lingvistikasi: muammolar, yechim va istiqbollar” mavzusidagi V xalqaro konferensiyasi ham davlat tilining axborot va kommunikatsiya texnologiyalari, xususan, Internet jahon axborot tarmog‘ida munosib o‘rin egallashini ta’minlash, o‘zbek tilining kompyuter dasturlarini yaratish kabi vazifalar ijrosiga bag‘ishlangan.

Konferensiyaning barcha ishtirokchilarini tabriklayman, konferensiya ishiga omad tilayman!

Shuhrat Samaraddinovich Sirojiddinov,

ToshDO‘TAU rektori, akademik

KOMPYUTER LINGVISTIKASI TADRIJI ZARURATI

Dunyo kompyuter lingvistikasida oxirgi yillarda milliy korpuslar, avtomatik tarjima platformalari, tezaurus, elektron lug‘atlardan foydalanish, ularni ilmiy-nazariy jihatdan o‘rganish harakati avj oldi. Shu jarayonda korpusni tuzishning dasturiy va lingvistik tamoyillarini asoslash, tabiiy tilni qayta ishlash (NLP), mashina tarjimasi, matnga qayta ishlov berish tizimlari, kompyuter lingvodidaktikasi, kompyuter leksikografiyasi, semantik tizimlar va lingvistik ontologiyalar ustida ishlash va korpus grammatikalarini yaratishga keng imkon yaratadi.

Mamlakatimizda global ilmiy-texnik islohotlar amalga oshirilayotgan bugungi kunda kompyuter lingvistikasi doirasida avtomatik tarjima, o‘zbek tilini tushunish va qayta ishlash, uni boshqa tillar bilan chog‘ishtirishda sun'iy intellektning rolini belgilashga kuchli ehtiyoj sezilmoqda. Shu ma'noda “...davlat tilining sofligini saqlash, uni boyitib borish va aholining nutq madaniyatini oshirish; davlat tilining zamonaviy axborot texnologiyalari va kommunikasiyalariga faol integrasiyalashuvini ta'minlash” kabi ustuvor yo‘nalishlar tahlil qilinayotgan til faktlarini tizimlashtirish, akademik maqsadlarda milliy va xorijiy til ko‘nikmalarini egallash, tilni real tasvirlash, shuningdek, amaliy tilshunoslikda til qo‘llanishi etalonlarini optimallashtirishga keng yo‘l ochadi.

Kompyuter lingvistikasi istiqbolli ilmiy yo‘nalish ekanligini inobatga olgan holda o‘zbek tili milliy korpusini yaratish, lingvistik modellarni tuzish singari masalalarni zamonaviy ilmiy tamoyillar asosida tadqiq etish fanimiz oldida turgan dolzarb vazifa hisoblanadi. Shu nuqtai nazardan o‘zbek tilining formal grammatikasini yaratish, korpus lingvistikasi, tabiiy tilni qayta ishlash (NLP), mashina tarjimasi, matnga qayta ishlov berish tizimlari, kompyuter lingvodidaktikasi, kompyuter leksikografiyasi, semantik tizimlar va lingvistik ontologiyalar, kompyuter lingvistikasida terminologiya va terminografiya, kompyuter lingvistikasida sun'iy intellekt kabi sohalarni rivojlantirish nihoyatda muhim.

Bugun universitetimizda tashkil etilgan “Kompyuter lingvistikasi: muammolar, echim va istiqbollar” mavzusidagi V xalqaro konferensiya mazkur masalalarni mutaxassislar va yosh tadqiqotchilar bilan muhokama qilish, bu sohalarda qilingan ishlar bilan tanishishga imkon beradi.

Nodir Imomqo‘ziyevich Jo‘raqo‘ziyev,
Davlat tilini rivojlantirish departamenti bosh mutaxassisi,
filologiya fanlari bo‘yicha falsafa doktori PhD, dotsent

O‘ZBEK TILINI RAQAMLASHTIRISH YO‘LIDA

Assalomu alaykum, hurmatli anjuman ishtirokchilari!

O‘zbek tilining nufuzini oshirish borasida mamlakatimizda keyingi yillarda olib borilayotgan islohotlar birin-ketin o‘z samarasini ko‘rsatmoqda. O‘zbekiston Respublikasi Prezidenti mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risidagi 6084-sonli farmonining 6-bandida Oliy va o‘rta maxsus ta'lim vazirligi oliy ta'lim muassasalarining filologiya ta'lim yo‘nalishlarida “Kompyuter lingvistikasi”, “Amaliy filologiya” yo‘nalishlarini tashkil qilish, o‘zbek tiliga oid barcha ilmiy, nazariy va amaliy ma'lumotlarni o‘zida jamlagan elektron ko‘rinishdagi o‘zbek tili milliy korpusini yaratish vazifalari qo‘yilgan edi. Buning amaliy natijasi sifatida jamg‘arma tomonidan o‘zbek tili milliy korpusini yaratish bo‘yicha amaliy loyihalar e’lon qilindi.

Bundan tashqari Toshkent davlat o‘zbek tili va adabiyoti universiteti olimlari tomonidan keng miqyosda milliy til korpusini ishlab chiqish ishlari olib borilmoqda. “O‘zbek tilining ta'limiy korpusini yaratish”, “Turkiy tillarning lingvodidaktik platformasi”, “O‘zbek tilining morfologik analizatori” amaliy loyihasi doirasida o‘zbek tilining korpusi, o‘zbek tilining morfologik analizatori, turkiy tillarni o‘rgatishga mo‘ljallangan platforma ishlab chiqildi. Olimlarning bunday tashabbuskorligini biz, albatta, qo‘llab-quvvatlaymiz! Korpus katta hajmli lug‘atlarni tuzish uchun manba vazifasini o‘taydi. Vaqt o‘tishi bilan korpus turli lingvistik yo‘nalishlar uchun ahamiyatli bo‘lishi bilan kuchli informasion resursga aylandi. Korpus asosida kompyuter yordamida lug‘atlar avvalgiga nisbatan tezlik bilan tuziladi va qayta ishlanadi. Shu yo‘l bilan ish boshlanishidan tugash jarayonigacha (nashrgacha) tilni aks ettirib turadi, lug‘at maqolasi “eskirish”ga ulgurmaydi. Tilning zamonaviy holatini aks ettiruvchi korpus turli davrlarda yashab ijod etgan ijodkorlarning mualliflik korpuslari uchun asos bo‘lib xizmat qiladi. Korpusning so‘zshunoslik sohasidagi ahamiyati shundaki, so‘zning qo‘llanish davri va chastotasini aniqlashda hech qanday vosita korpusga tenglasha olmaydi. Korpus lingvistikasining keyingi taraqqiyot bosqichida statistik tadqiq metodi kompyuter tarjimasi, nutqni sintezlash va tanish, aniqlash, orfografik tekshiruv kabi lingvistik amallarni bajarishda qo‘l kela boshladi. Yaqin orada til o‘rganayotgan o‘quvchi yoki uning biror jihatini tadqiq etayotgan kishining bugungi kunda lug‘atga bo‘lgan ehtiyoji korpusga ko‘chishi shubhasiz.

Konferensiya ishiga omad tilayman!

Nozliya Zarilovna Normurodova
ToshDO‘TAU ilmiy ishlar va innovatsiyalar bo‘yicha prorektori,
filologiya fanlari doktori, dotsent

YALPI YIG‘ILISH MA’RUZALARI

DOĞAL DİL İŞLEMEDEKİ GELİŞMELER

Eşref Adalı,
Prof. Dr.
İstanbul Teknik Üniversitesi
Bilgisayar ve Bilişim Fakültesi
Türkiye
adali@itu.edu.tr

1. Doğal Dil İşlemenin Kısa tarihi

Konuştuğumuz dilin (doğal dil) bilgisayar ile işlenmesine Doğal Dil İşleme diyoruz. DDİ çalışmalarının ne zaman başladığı konusu, 1954 yılına kadar gitmektedir. O yıllarda, Rusça bilimsel yayınların İngilizceye bilgisayar ile çevrilmesi amaçlanmıştır. Bu amaçla ALPAC (Automatic Language Processing Advisory Committe) kurulmuştur. Dönemin bilgisayarlarının yetenekleri bu çalışmayı sürdürmeye yetmemiştir. Bu dönemin önemli bir çalışması IBM Georgetown Demonstration (1954) ve aynı yıl ilk dergi (Mechanical Translation- daha sonra Computational Linguistics) sayılır.

DDİ nedir sorusunun yanıtı şöyle verilir:

- İnsan dilini anlamaya, yorumlamaya ve işlemeye yardımcı olan bir Yapay Zeka dalı.
- Doğal dili işlemek ve anlamak için kullanılan araçlar, teknikle, yöntemler ve algoritmalar hakkında her şey.
- Bilgisayarların insan dilini anlamasını ve işlemesini sağlamaya odaklanan Yapay Zeka alt alanı.

Şekil-1: DDİ - Bilgisayar bilimi - Yapay zeka - Doğal dil ilişkisi

Şekil-1’de DDİ ile bilgisayar bilimi, yapay zeka ve doğal dil arasındaki ilişki gösterilmiştir.

2. DDİ’nin Kısa Tarihi

DDİ’nin kısa tarihi aşağıda verilmiştir [1]:

1950'ler

- Erken Makine Çeviri çalışmaları
- Dil bilgisi üzerinde çalışmalar

1960-80'ler

- **Uygulamalar**

- BEYZBOL: Beyzbol oyunlarıyla ilgili bilgileri veri tabanında aramak üzere hazırlanmıştır.
- ELIZA: Çok basit soru-yanıt uygulaması
- SHREDLU: Belli sayıda harf içeren sözcük üretme uygulaması
- Mesaj anlama: Gazetelerde yayımlanan yazılardan terörizmle ilgili olanları bulmak ile ilgili bir çalışma
- Makine çevirisi çalışmaları.

- **Yöntemler**

- ATN (artırılmış geçiş ağları): Bağlamdan bağımsız dilbilgisi yöntemleri ile ilgili çalışmalar.
- Dilbilgisi çalışmaları (etken, nesne, vb.)
- DCG-Belirli Tümce Dilbilgisi
- Bağımlılık dilbilgisi: Öğeler arası bağımlılık araştırması ile ilgili çalışmalar.

1990 ve sonrası

- İstatistiksel yöntemler
- Konuşma tanıma
- Makine Çevirisi Sistemleri
- Soru Yanıtlama
- Büyük Dil Modelleri (BDM-LLMs)

Bazı Güncel DDİ Uygulamaları

Günümüzde çalışılan konulardan bazıları şunlardır:

- Soru Yanıtlama
- Makine Çevirisi
- Duygu Analizi
- Özet Çıkarma
- Metni anlama
- Doğal dil benzeri tümce oluşturma

3. Bir Tümce veya Metni Anlama

Bir tümce veya metni anlama ve sorunları aşağıda kısa açıklanmıştır.

Tümce veya Metni Anlama

- Metin ya da tümceyi (girdi) yararlı gösterime eşleme.
- Dilin farklı yönlerini analiz etme.

Doğal Dil Üretimi

- Doğal dile yakın biçimde, anlamlı ifadeler ve tümceler üretme süreci: Şu konuları içerir:
  - Metni Düzenleme (bilgi kaynağından ilgili içeriği getirme)
  - Tümce Düzenleme (gerekli sözcükleri seçme, anlamlı ifadeler oluşturma, tümcenin tonunu ayarlama)
  - Metni Gerçekleştirme (tümce düzenini tümce yapısına eşleme)

Anlam Belirsizliği

Belirsizlik

- **Sözcesel/Biçim Bilimsel** (Lexical/Morphological): değişim (Eylem, Ad - V,N), eğitim (Eylem, Ad - V,N), çift (Önad, Belirteç -ADJ,ADV)…
- **Söz Dizimsel**: İngilizcede sorun “Helicopter powered by human flies”, Türkçede sorun yok “İnsan yönetiminde uçan helikopter”
- **Anlamsal**: İngilizcede sorun “He saw a man on the hill with a telescope”, Türkçede sorun yok “ Tepedeki bir adamı teleskopla gördü”
- **Söylem**: anafor ,her dilde sorun. Ekmek sözcüğünün Türkçe çok farklı kullanımı var

Çözüm Yöntemleri

- Daha önceki adımda bulunmuş çözümü kullanmak.
- İstatistiksel yöntem kullanmak
- Dikkat vektörü kullanmak
- Sözcük kesesi kullanmak

4. DDİ’nin Temel Ögeleri

DDİ çalışmalarının temel taşlarını oluşturan ögeler Şekil-2’de sıradüzensel biçimde gösterilmiştir [3].

Şekil 2: DDİ'nin temel ögeleri

Şekil-3’te bir DDİ sürecinde işlemlerin izlediği akış (İş Akışı) gösterilmiştir.

Şekil 3: DDİ İş hattı

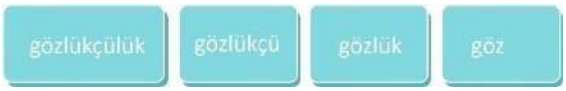
DDİ İş hattında görülen işlem ile ilgili açıklamalar aşağıda verilmiştir.

Lokmalama (Tokenization)

- Karmaşık bir tümcelerine ayırmak
- Tümce açısından her bir sözcüğün önemini belirtmek
- Tümceden yapısal bir açıklama oluşturmak
- Noktalama işaretini göstermek

Sözcüğün Kökünü Bulma (Yapım eklerinden arındırma) (Stemming)

Bir sözcüğün kökünü bulma işlemidir. Türkçe gibi eklemeli dillerde yapım ekiyle türetilmiş sözcük kök sözcüğün temel anlamını taşır.



Sözcüğün Çekimli Biçimleri (Lemmatization)

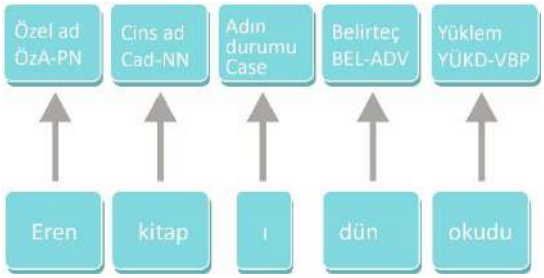
Türkçede sözcüğü çekim eklerinden arındırmak anlamında düşünülebilir. İngilizce gibi kuralsız dillerde anlamı farklıdır. Örneğin “*go, going, went*” ve “*gone*” sözcüklerinin temeli “*go*” dur. Benzer biçimde “*better, best*” sözcüklerinin temeli ”*good*” sözcüğüdür. Türkçe gibi kurallı dillerde bu tür gariplikler görülmez. Türkçe gibi kurallı dillerde sözcüğün (eylem ya da ad) çekim eklerinden arındırılması olarak yorumlanabilir.

Sağır Sözcüklerin Kaldırılması

Sağır sözcükler, makine öğrenmesinde gürültüye neden olabilir bu nedenle metnin işlenmesinden önce veya sonra kaldırılır.

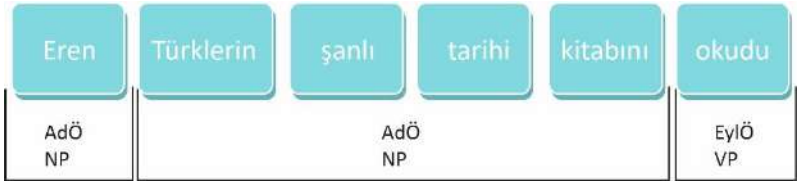
POS

POS, sözcüklerin söz dizimsel bağlamlarına ve rollerine göre atanması sürecidir.



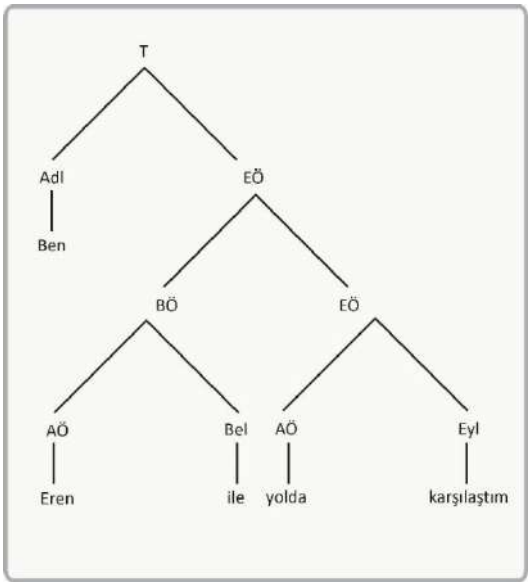
Parçalara ayırma

Bir tümceyi ögelerine ayırma işlemi



Ayrıştırma

Bir cümleyi en küçük bileşenlerine ayırma işlemi



5. Dil Modelleri

Bir dilin özellikleri çıkarabilmek üzere dillerin modelleri çıkarılmaya çalışılmıştır. İlk dönemlerde, bir sözcükten sonra gelebilecek sözcük ya da sözcüklerin nasıl tahmin edileceği üzerinde çalışılmıştır. Bu amaçla n-gram ve Matkov zincirleri geliştirilmiştir. Örneğin "*Her şey çok …… olacak*" tümcesinde boş olan yere "*güzel, kötü, berbat, muhteşem, ucuz, pahalı*" gibi sözcüklerden biri gelebilir. Bu sözcüklerden hangisinin geleceğini o dilin derlemine bakılarak karar verilir. Derlemde en sık geçen sözcüğün bu boş alana yazılması en doğru seçenektir.

Tanım olarak bir derlemin, bir dilde üretilmiş olan tüm yayınları içermesi beklenir ancak bunu sağlamak çok kolay değildir. Bunun yerine, farklı konularda, farklı yazarlar tarafından üretilmiş metinlerden belli boyutlarda örnek metinler alınarak derlem oluşturulabilir. Böylesine hazırlanmış olan bir derleme "**Dengeli Derlem**" adı verilmektedir. Dillerin zaman içinde geliştiği düşünüldüğünde, dengeli derlemde yer alacak metinlerin belli bir zaman kesitinde yayımlanmış olanlardan seçilmesi gerekir.

Bir dilde yayımlanmış yayınları rastgele toplayarak da derlem hazırlanabilir. Bu durumda hazırlanmış olan derleme "**Dengesiz Derlem**" adı verilir.

Derlem bir dildeki tüm yayınları barındıracak biçimde hazırlanmış ise bu tür derlemlere "**Genel Derlem**", yalnızca belli bir alanı kapsayan yayımlardan oluşturulmuş ise "**Özel Derlem**" adı verilmektedir.

Diller arası çevirilerde kullanılmak üzere de derlemler hazırlanmaktadır. "**Koşut Derlem**" olarak anılan bu derlemlerde yer alan tümceler birbirinin karşılığı olacak biçimde yer alırlar [2].

Yakın zaman, dil modelleri üzerindeki çalışmalar yoğunlaşmıştır. N-gram kullanmak yerine Büyük Dil Modelleri (LLMs) geliştirilmeye başlanmıştır. BDM'lerini üretmek için genellikle çok büyük miktarda metin bir araya getirilir; bir anlamda dengesiz derlem oluşturulur. Bu veri kümesi üzerinde, bir sonraki sözcük, bir sonraki tümce hatta bir sonraki metin tahmin edilmeye çalışılır. Bazı veri kümeleri metinlere ek olarak resim de içerir. Böylece bir resim de değerlendirilir. Bir sonraki sözcük, tümce veya metni tahmin etmek üzere sinir ağının gelişmiş bir biçimi olan **Derin Öğrenme** yöntemi kullanılır.

5.1 Büyük Dil Modelleri Nelerdir?

Büyük dil modelleri, çok büyük veri kümeleri üzerinde eğitim sırasında edinilen bilgiye dayanarak metni tanıyabilen, çıkarım yapabilen, özetleyebilen, tahmin edebilen ve üretebilen derin öğrenme algoritmalarıdır. Ayrıca dil modelleri adı verilen daha genel bir teknolojinin alt kümesidir. Dil modellerinin ortak özellikleri doğal dilde yazılmış metinleri işleyebilmeleri ve doğal dilde tümceler kurabilmeleridir. Büyük boyuttaki veri kümesi metin ve resim barındırdığından

büyük veri olarak kabul edilir. BDM’leri, bu büyük veri kümesi üzerinde eğitilirler. Bu eğitim aşaması çok sayıda eğitilebilir katsayı (parametre) içerir. Günümüzde çok bilinen iki BDM’nin veri kümesi boyut Çizelge-1’de verilmiştir [4].

Çizelge-1: MT-NLG ve GPT-3'ün Karşılaştırılması

Büyük Dil Modeli	Katsayı (Parametre) sayısı	Eğitim veri kümesindeki imleç (token) sayısı
NVIDIA (MT-NLG)	530 milyar	270 milyar
OpenAI Model: GPT-3 Davinci Model	175 milyar	499 milyar

Çizelge-1’de görülen dil modelleri çok büyük veri üzerinde eğitilirler ve soruları yanıtlamak, özetleri oluşturmak, tümceleri tamamlamak ve diller arası çeviri yapmak amacıyla kullanılırlar.

Bir modelin başarımı, bir anlamda kalitesi, eğitim verisinin boyutuna bağlıdır. Bu nedenle, daha büyük dil modelleri genellikle daha küçük dil modellerine göre daha doğru başarılıdırlar. Ancak, şunu da hatırlatmakta yarar vardır: büyük dil modellerinin başarımı yalnızca model boyutuna veya veri miktarına bağlı değil; verinin kalitesine de bağlıdır. Örneğin, hakemler tarafından değerlendirilmiş bildiri ya da makaleler veya editör tarafından düzeltilmiş hikaye ve romanlar üzerinde eğitilmiş alan BDM'lerin başarımı sosyal medyadan toplanmış mesajlar, yorumlar veya diğer incelenmemiş içerikler üzerinde eğitilmiş alan BDM'lerden daha başarılı sonuçlar üretirler. Belli bir konuda çalışmak üzere hazırlanan dil modelleri doğal olarak, ilgili alandaki veriler ile oluşturulur. Belli bir alanı kapsayacak biçimde oluşturulan BDM’leri boyut olarak genel BDM’lerinden daha küçük olurlar. Ayrıca, bu veri kümesi üzerinde ince ayar yapılarak daha başarılı sonuçlar üretirler.

5.2 Büyük Dil Modellerinin Evrimi

Yapay zeka (YZ) alanında, geçmişte yapılmış olan çalışmalar verileri işlemek ve analiz etmek amaçlı iken yakın zamandaki çalışmalar anlamak ve doğal dilde tümce kurmak üzerinedir. Böylece Algılayıcı ve Üretken YZ çalışmaları ağırlık kazanmıştır. Günümüzdeki BDM’lerini anlayabilmek için büyük dil modellerinden önceki DDİ modellerini gözden geçirmekte yarar vardır.

1. **Dönüştürücü Kullanmayan DDİ**: Bu çalışmalarda DDİ görevlerini yerine getirmek için makine öğrenmesi yöntemleri yerine insan tarafından işaretlenen modeller kullanılmıştır. Doğal olarak tüm kuralların bilindiği ve modele işlendiği varsayılır. Bu modeller, metin sınıflandırması çalışmaları için yeterli olmuştur. Makine çevirisi uygulamalarında, kaynak ve hedef dil aynı dil ailesinden ise bu modeller yeterince başarılı olmuşlardır ancak farklı dil ailesinde yer alan diller arasındaki çevirilerde başarılı olamamışlardır. Bilinmeyen kurallar söz konusu olduğunda başarısız olurlar. Varlığı bilinmeyen kuralların neden olduğu sorunlar

çözebilmek amacıyla RNN'ler ve LSTM'ler gibi sinir ağlarıyla çözülmeye çalışılmıştır. RNN'ler ve LSTM'ler önceki veriler üzerinden çıkarımlarda bulunarak ya da önceki sonuçları belleyerek tahmin ve sınıflandırma yapabiliyordu. RNN'ler ve LSTM'lerin tahminler kısa geçmişteki verilere dayandığından uzun metinler için tahmin yapamazlardı.

2. **Dönüştürücü Kullanan DDİ**: Dönüştürücü mimarisi 2017'de önerildi ve kullanılmaya başlandı. Dönüştürücüler, o zamana kadar kullanılan RNN ve LSTM'lere oranla daha iyi genelleme yapabiliyor daha fazla bağlam yakalayabiliyor ve daha fazla veriyi işleyebiliyordu. Dönüştürücü kullanımı ile DDİ modellerinin daha uzun veri dizilerini anlaması ve çok daha görev yapabilmesine olanak sağlandı. Dönüştürücü kullanan DDİ uygulamaları ilk aşamada araştırıcılar tarafından kullanılmaya başlandı; ticari olarak alanda uygulanmadılar.

3. **BDM Kullanan DDİ**: Büyük Dil Modellerini kullanan DDİ’ler 2020'de OpenAI'nin GPT-3'ün tanıtılması ile dikkat çekmeye başladı. GPT-3 ve benzeri BDM’leri, önceki dil modellerine göre daha doğru ve kapsamlı sonuçlar üretebiliyordu. Bunun nedeni kullandıkları veri kümesinin çok büyük olması ve bu modelin eğitilebilmesidir. BDM'leri çeşitli DDİ işlevlerini yalnızca doğal dil istemlerini kullanarak çözebilen teknik olmayan kullanıcılar için kolay erişilebilir hale getirdi.

Büyük Dil Modeline geçişin aşamalı biçimde olduğu yukarıdaki açıklamalardan anlaşılmaktadır. Bu gelişmelerde sinir ağlarının özellikle derin öğrenmenin çok etkin olduğu bir gerçektir. Dönüştürücüler bu yöntemler kullanılarak gerçekleştirilebilmiştir. Bu teknik ve yöntemlerin öğrenmekte yarar vardır.

5.3 Yapay Sinir Ağları

Yapay Sinir ağları YSA (SA) (NN) ya da kısaca sinir ağları, insan beyninin çalışma ilkesine temel alarak modellenmiş makine öğrenmesi algoritmalarıdır. İnsan beynine benzetilerek sinir ağları düğümlerden (nöron) oluşur ve bunlar girdiyi işlemekten çıktı üretmeye kadar tüm işlevleri yürütürler. Her sinir ağında en az üç katman bulunur:

- **Giriş katmanı**: Verileri kabul eder ve ağın geri kalanına iletir.
- **Gizli katman veya birden fazla gizli katman**: SA'dan beklenen işlevleri yerine getirir. Bu işlevler, verileri tanımlamayı veya sınıflandırmayı, yeni veri üretmeyi ve belirlenmiş olan DDİ işlevini yerine getiren işlevlerdir.
- **Çıkış katmanı**: Girdiye dayalı bir tahmin veya sınıflandırma sonucunu üretir.

RNN'ler ve LSTM'ler gibi sinir ağlarının, önceki kural tabanlı yöntemlere göre en büyük üstünlükleri, çok az ya da hiç insan katkısı olmadan kendi başlarına

öğrenebilmeleriydi. Bu yeni yöntemler kuralları öğrenip daha sonra verilere uygulamak yerine, kendi kurallarını oluşturmak için verileri analiz edebiliyorlardır. Çalışma biçimleri aşağıdaki örnekte görülmektedir:

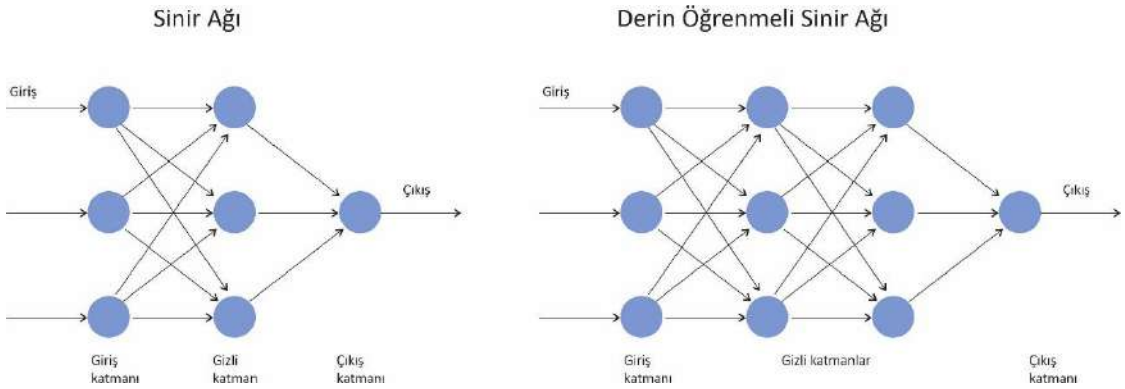
“kedi,” “kediler,” kuş,” “kuşlar”

Bu örnekte görülen lokmalar analiz ettikten sonra, model aşağıdaki biçimde modeli eğitir:

Çoğul adlar “-lAr” eki alırlar.

Model daha sonra bu modeli kullanarak başka adların çoğullarını bulabilir veya üretebilir. Örneğin, ”masa” dan “masalar”, “ev” den “evler”i türetebilir. Benzer biçimde diğer dil bilgisi kurallarını çıkarabilir. Bu tür özellik çıkarma, diğer adıyla gösterim çıkarma Türkçe gibi kurallı diller için çok kullanışlıdır. Sinir ağları ne kadar çok nörona sahipse, o kadar karmaşık gösterimler çıkarabilirler. Bu nedenle, BDM’leri birden fazla gizli katmana ve dolayısıyla daha fazla nörona sahip derin öğrenme sinir ağlarını kullanır [7].

Şekil-4’te tek katmanlı bir sinir ağı ile derin öğrenme sinir ağı çizilmiştir.



Şekil-4:Tek Katmanlı ve Derin Öğrenmeli Sinir Ağı

Günümüzde çok sayıda gizli katman içeren sinir ağlarının kullanılabilmesi Grafik İşlemci Birimlerinin (GPU) GPU’ların 1999’dan sonra kullanılmaya başlanması ve 2012 de yaygınlaşması ile sağlanmıştır. GPU’lar, sinir ağında bulunan katmanlar ile işlemleri paralel biçimde işleyebilme yeteneğine sahiptirler. Böylece işlemlerin süresi kısalmıştır. GPU'ların kullanımındaki başarının bir sonucu olarak dönüştürücüler gündeme gelmiştir.

5.4 Dönüştürücüler

Daha önce açıklandığı gibi RNN'ler ve LSTM'ler diller arası çevirilerde yetersiz kalıyorlardı. Bu eksiklikleri gidermek üzere dönüştürücü adı verilen yeni, özel sinir

ağları tasarlanmıştır. Dönüştürücüler ilk olarak 2017 yılında Vaswani ve arkadaşları tarafından "***Dikkat İhtiyacınız Olan Tek Şey***" başlıklı bir makalede tanıtılmıştır [5]. Bu makalede, dönüştürücülerin temel bileşeni haline gelecek olan “**dikkat mekanizması”**na atıfta bulunmuştur. Önerilen yöntem tekrarlama ve evrişimleri ortadan kaldırmıştır. Dikkat mekanizmaları, modellerin girdiyi işlerken, girdinin belirli bölümlerine özellikle dikkat etmesini isteyerek yetersiz bağlam işleme sorununu çözmüştür. Tüm bağlamı bir adımda yakalamak yerine, belirli bir görevle ilgili en önemli lokmalara odaklandırılmıştır.

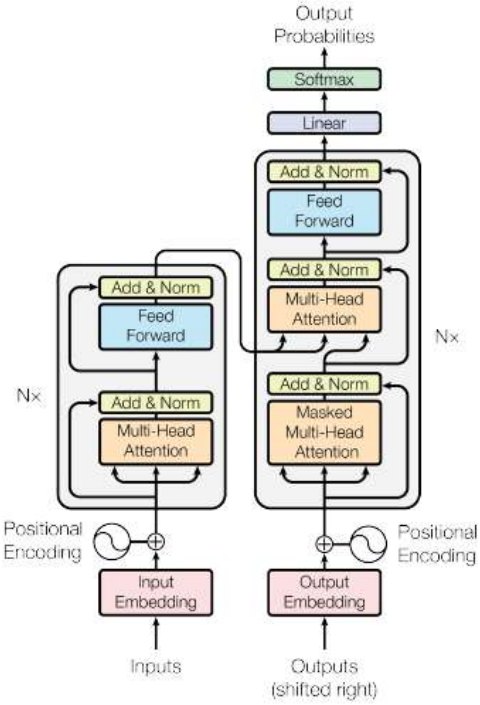
Dikkat mekanizmalarına dayanan öz-dikkat katmanları daha önce girilen her sözcük için dikkat ağırlıklarını hesaplar. Dikkat ağırlıkları her bir lokmanın önemini temsil eder, bu nedenle bir lokmaya ne kadar fazla ağırlık atanırsa o kadar önemli kabul edilir. Örneğin,

Eren’in bir kedisi var.

Tümcesinde dikkat mekanizması "*kedi*" lokmasına "*bir*" lokmasından daha fazla ağırlık verebilir. Bunun nedeni “*kedi*” lokmasının sonuca etkisinin “*bir*” lokmasının etkisinden büyük olmasıdır. Her bir lokmanın ne kadar önemli olduğunu belirlemek için, öz-dikkat katmanları bir metindeki bir lokmanın diğer lokmalarla ilişkilerini değerlendirir. Bir lokmanın diğer lokmalarla çok sayıda ilişkisi varsa, o lokma önemli kabul edilir.

LSTM'ler ve RNN'ler girdiyi ardışık olarak işledikleri için girişe karşılık çıkışın üretilmesi uzun zaman alır. Buna karşın dönüştürücüler paralel çalışırlar, dolayısıyla işlem süresi çok kısadır. Ancak, LSTM'ler ve RNN'ler yavaş olmalarına karşın, lokmaları ardışık işlemeleri nedeniyle hangi lokmanın önce ya da sonra geldiğini, yani lokmaların ya da sözcüklerin sırasını bilirler. Dönüştürücülerin böyle bir özellikleri yoktur. Bu eksikliğin üstesinden gelmek için dönüştürücüler, lokmaları konum bilgilerine yardımcı olan konumsal kodlamalar kullanır. Konumsal kodlamalar, her bir lokmayla ilişkilendirilen ek girdiler veya vektörlerdir.

Şekil-5’te dönüştürücü mimarisi gösterilmiştir. Şekil “Dikkat İhtiyacınız Olan Tek Şey” bildirisinden alınmıştır.



Şekil-5: Dönüştürücü mimarisi

Araştırmacılar ve uygulamacılar dönüştürücüleri uygulamaya başlamışlardır. Uygulamalardan biri Google tarafından 2018 yılında BERT' adıyla duyurulmuştur.

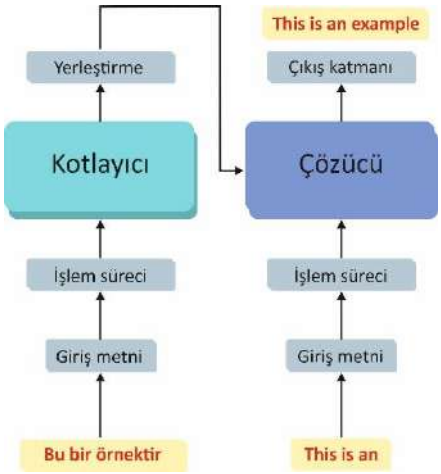
Dil çevirisi için derin öğrenme modeli olan orijinal dönüştürücü mimarisinin basitleştirilmiş bir çizimi Şekil-6’da görülmektedir. Şekildeki dönüştürücü iki bölümden oluşmaktadır: Birinci bölüm giriş metnini işleyen ve metnin bir yerleştirme gösterimini (farklı boyutlarda birçok farklı etkenin sayısal gösterimi) üreten bir kodlayıcı, ikinci bölüm çevrilen metni kod çözücünün her seferinde bir kelime üretmek için kullanabileceği bölümdür. Bu şekil, kod çözücünün yalnızca son sözcüğü üretmesi gereken çeviri sürecinin son aşamasını gösterir.

5.4.1 BERT

BERT (Bidirectional Encoder Representations from Transformers - Dönüştürücülerden Çift Yönlü Kodlayıcı Gösterimleri) ilk olarak tanıtılan dönüştürücü tabanlı dil modellerinden biridir [8]. Maskeli dil modelidir (MLM), yani maskeli lokmalar içeren tümcelerle eğitilmiştir. Modelin, maskeli lokmayı, çevresindeki bağlamı dikkate alarak tahmin etmesi amaçlanmıştır. Yapıyı açıklamak için aşağıdaki örnek tümcenin giriş olarak verildiğini düşünelim:

Askıda bir ***şapka*** var.

Bu örnek tümcede “*şapka*” sözcüğünün maskelendiğini ve dönüştürücünün bulmasını isteyelim. Şapka sözcüğünün öncesinde “*bir*” ve sonrasında “*var*” lokmaları bulunuyor. Dönüştürücü, bu tümceden önceki tümcelere ve bağlamı inceleyerek maskelenen sözcüğün “*şapka*” olduğunu bulur. Tek yönlü çalışan modellerde maskelenmiş lokmayı bulmak daha zordur. BERT, çift yönlü çalışan ilk dönüştürücü modelidir.



Şekil-6: Basitleştirilmiş dönüştürücü mimarisi

6. Diğer Büyük Dil Modelleri

Dönüştürücüler kullanılmaya başlamasıyla farklı özellikte BDM’leri oluşturmaya başlanmıştır. Çizelge-2’de bazı önemli BDM’leri verilmiştir.

Çizelge-2: Bazı Önemli BDM’leri

Model	Üretici Firma	Yıl	Kısa açıklama
GPT-2	OpenAı	2019	İnsan benzeri metinler üretmek ve dil çevirisi, özetleme ve soru yanıtlama gibi DDİ işlevlerini gerçekleştirmek için tasarlanmıştır.
RoBERTa	Facebook	2019	BERT temelli bir BDM’dir. Daha büyük bir veri kümesi üzerinde eğitilmiş, daha verimli bir eğitim yöntemi kullanılmıştır. Bu nedenle başarımı yükseltilmiştir.

DeBERTa	Microsoft	2020	BERT temelli olan bu uygulamada kodlayıcı ve kod çözücü birbirinden ayırarak DDİ işlevlerinin başarımı yükseltilmiştir.
GPT-3	OpenAı	2020	Daha büyük bir veri kümesi üzerinde eğitilen ve daha başarımlı çıktılar üretebilen GPT-2'nin bir üst modelidir.

6.1 GPT ve BERT Karşılaştırması

GPT (Generative Pre-trained Transformer - Üretken Ön-eğitilmiş Dönüştürücü) ve BERT yaygın olarak kullanılan ancak mimarileri farklı BDM’lerdir. GPT, bir tümce içindeki bir sözcük verildiğinde bir sonraki sözcüğü tahmin etmek üzere eğitilen bir modeldir. Bu ön eğitim sayesinde tutarlı ve akıcı tümceler üretebilmektedir. Buna karşın, BERT, tümceleri veya lokmaları duygu analizi, adlandırılmış varlık tanıma ve metin sınıflandırması gibi farklı ulamlara ayırmak üzere eğitilmiş bir modeldir. Bir sözcüğün anlamını çıkarmak için tümcenin önceki ve sonraki bağlamlarını dikkate alan iki yönlü bir modeldir. Duygu analizi ve soru yanıtlama gibi uygulamalarda başarılıdır.

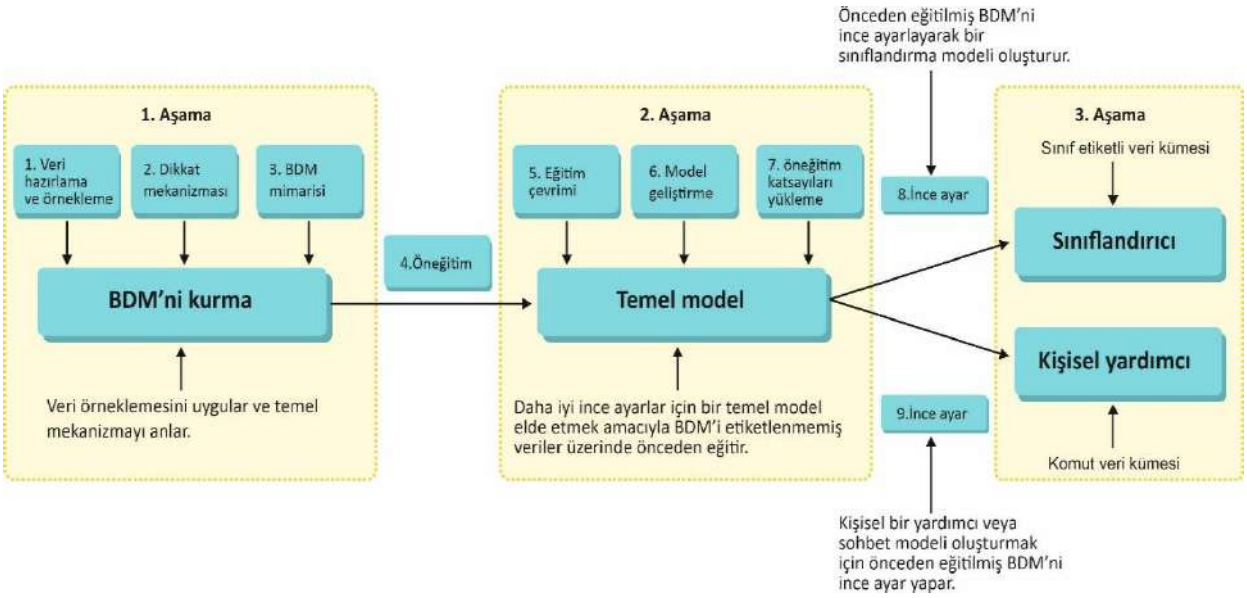
GPT’nin dönüştürücü mimarisi tek yönlü olmasına karşın BERT’inki çift yönlüdür. Bunun sonucu olarak bir sözcüğü tahmin ederken yalnızca önceki sözcüğe bakarken, BERT önceki ve sonraki sözcüklere bakarak tahminde bulunur.

6.2 BDM Oluşturma

Sıfırdan büyük dil modelleri oluşturmak, özellikle temel işi yapay zeka veya DDİ olmayan işletmeler için mantıklı değildir. BDM oluşturma süreci zaman alıcı ve emek gerektireceğinden işletmelerin mevcut BDM’leri gereksinimlerine göre özelleştirmeleri salık verilir. Mevcut BDM’leri özelleştirmek için aşağıda açıklanan üç adımın izlenmesi yerinde olur [9,10].

1. **İşletmenin gereksinimlerini karşılayabilecek modeli araştırmak**: Bu araştırma sırasında modelin boyutu, veri kümesi ve sağlayıcısını araştırmak gerekir.
2. **Modelin ince ayarını yapmak**: Temel BDM’lerinin işletmenin gereksinimlerine ve kullanım biçimine göre ince ayarlarının yapılması gerekir. Örneğin, metin sınıflandırmaya, duygu analizine veya yasal mevzuata uyumu sağlamak üzere ince ayar yapılır.
3. **Modeli en iyi biçime getirmek**: Genel amaçlı olarak hazırlanmış olan modellerde işletme için gereksiz olabilecek veri kümeleri veya yetenekler bulunabilir. Bunlar modelden çıkarılarak bilgisayarın iş yükü azaltılabilir.

Bir BDM’nin oluşturulması Şekil-7’da ayrıntılı biçimde gösterilmiştir.

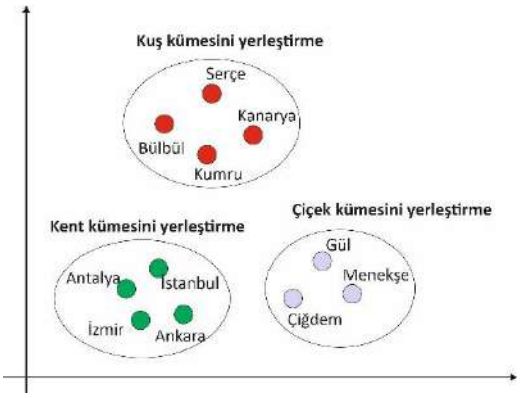


Şekil-7: Bir BDM'nin geliştirilme aşamaları

Şekil-7’de görülen işlevler aşağıda teker teker açıklanmıştır.

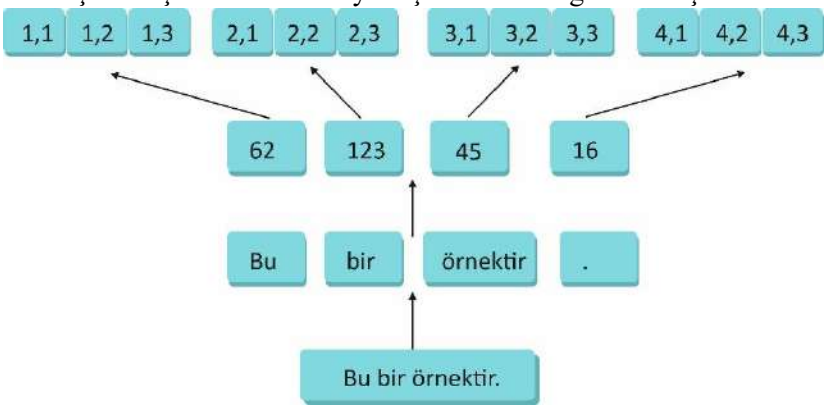
Veri Hazırlama ve Örnekleme

Bu ilk aşamanın görevi, girilen metnin lokmalara ayrılması, lokmaların sayılar ile temsili ve yerleştirilmesidir. Lokmaların vektör biçiminde temsili yerleştirme (embedding) olarak adlandırılır. Yerleştirme için farklı algoritmalar kullanılmaktadır. Bu algoritmalardan biri Word2Vec olarak anılır. Word2Vec'in temel fikri benzer bağlamlarda görünen sözcüklerin benzer anlamlara sahip olma eğiliminde olmasıdır. Şekil-8’de bu fikir tanıtılmıştır.



Şekil-8: İki boyutta yerleşme örneği

Yerleştirme öncesi metin lokmalarına ayrılır. Lokma, metin içinde görünen sözcükler ve noktalama işaretleridir. Tüm metinde geçen sözcükler alfabetik olarak sıralanır; tekrarlananların sırası kaydırılır. Bu aşamadan sonra her sözcüğe sıra numarası atanır. Bir sonraki aşamada her bir lokma için vektör ataması (yerleştirme) yapılır. Şekil-9’da basitleştirilmiş lokmalama ve yerleştirme süreci gösterilmiştir.



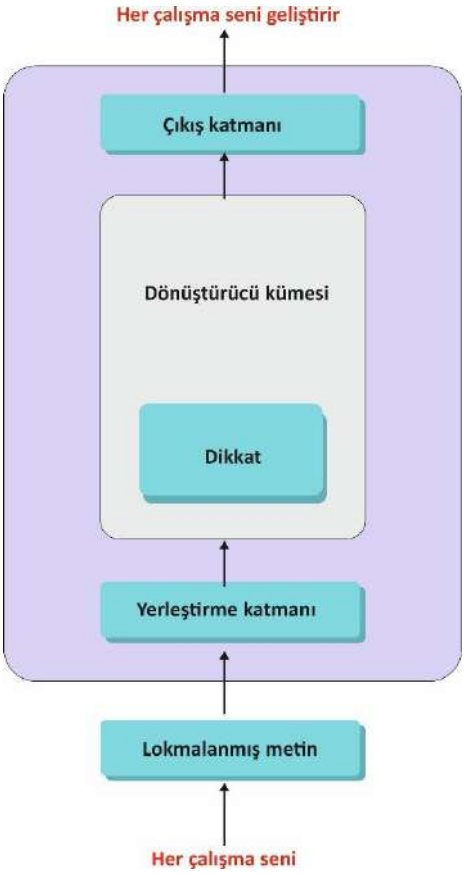
Dikkat Mekanizması

Bir metin içinde bulunan her bir sözcüğün metin içindeki önemi ya da ağırlığını belirtir. Dikkat mekanizmasının türleri şunlardır:

- Basit öz dikkat
- Öz dikkat
- Nedensel dikkat
- Çok başlı dikkat

BDM Mimarisi

GPT gibi BDM’leri, her seferinde bir sözcük (veya lokma) olmak üzere yeni metin üretmek üzere tasarlanmış büyük derin sinir ağı mimarileridir. Ancak, büyük boyutlarına karşın, model mimarisi çok karmaşık değildir. Çünkü bileşenlerinin çoğu tekrarlanır. Şekil-10’da GPT benzeri BDM’nin mimarisi çizilmiştir.



Şekil-10: BDM Mimarisi

Derin öğrenme ve GPT gibi BDM’lerinde, "*parametreler*" terimi modelin eğitilebilir ağırlıklarını ifade eder. Bu ağırlıklar aslında, belirli bir kayıp fonksiyonunu en aza indirmek için eğitim süreci sırasında ayarlanan ve en iyi biçime getirilen modelin iç değişkenleridir.

Ön Eğitim, Eğitim Çevrimi, Model Geliştirme ve Ön Eğitim Katsayılarını Yükleme

Şeki-7’deki 4, 5, 6, 7. Adımlar BDM’nin eğitilmesi ile ilgilidir. Bu aşamalarda, lokmalar arasındaki ağırlık katsayıları hesaplanır ve sonuçlar tekrar güncellenir. Bu işlemlerin sonunda, DDİ’ye yakın tümce oluşturulur.

İnce Ayar

BDM’nin araştırmanın ya da işletmenin gereksinimlerine göre ayıklanması işlemine ince ayar denilmektedir.

Kaynakça

[1] K. S. Jones, Natural Language Processing: A Historical Review, Computational Linguistics, 1994

[2] E. Adalı, Türkçe Doğal Dil İşleme, Akçağ Yayınevi , 2020, ISBN 97860553425519

[3] F.M.Kom, Natural Language Processing Introduction, theory and application, Universitas Gadjah Mada

[4] A. Chockalingam, A. Patel, S. Verma, T. Yeung, A beginner’s Guide to Large Language Models, NVIDIA Corporation. 2023

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, 31st Conference on Neural Information Processing Systems (NIPS 2017)

[6] S.J.D. Prince, Understanding Deep Learning, The MIT Press, March 26, 2025

[7] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, A. Mian, A Comprehensive Overview of Large Language Models, arXiv:2307.06435v10 [cs.CL] 17 Oct 2024

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805v2 [cs.CL] 24 May 2019

[9] J. Kunz, Understanding Large Language Models, Linköping Studies in Science and Technology Dissertation No. 2364

[10] S. Raschka, Build a Large Language Model From Scratch, Manning Publications Co. 2025, ISBN: 9781633437166

O‘ZBEK TILINING ELEKTRON RESURSLARINI YARATISH – ZAMON TALABI

Assalomu alaykum, hurmatli anjuman ishtirokchilari!

O‘zbekiston Respublikasi Prezidentining “Mamlakatimizda o‘zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to‘g‘risida”gi 6084-sonli farmonida Internetda o‘zbek tilidagi kontentni ko‘paytirish, o‘zbek tilini raqamlashtirish, o‘zbek tiliga zamonaviy axborot texnologiyalarini joriy etish kabi vazifalar belgilangan. Buning amaliy natijasi sifatida jamg‘arma tomonidan o‘zbek tili milliy korpusini yaratish bo‘yicha amaliy loyihalar e'lon qilindi, o‘zbek tilining o‘quv elektron lug‘atlarini yaratish bo‘yicha qator loyihalar amalga oshirildi. Ammo bu loyihalar bilan barcha vazifalarni qamrab olish mumkin emas. Shu sababli o‘zbek tili korpuslari, o‘zbek tiliga avtomatik ishlov berish vositalari, o‘zbek tili imlo tekshiruv va tahrir dasturi, o‘zbek tili sinxron tarjimonlari, o‘zbek tilida turli sohalarga oid savol-javob tizimlari, shuningdek, turli elektron lug‘atlarni ishlab chiqish hali ham dolzarbligicha qolmoqda.

Toshkent davlat o‘zbek tili va adabiyoti universiteti olimlaridan iborat ishchi guruh tomonidan keng miqyosda milliy til korpusini ishlab chiqish ishlari olib borilmoqda. “O‘zbek tilining ta'limiy korpusini yaratish”, “Turkiy tillarning lingvodidaktik platformasi”, “O‘zbek tilining morfologik analizatori” amaliy loyihasi doirasida o‘zbek tilining korpusi, o‘zbek tilining morfologik analizatori, turkiy tillarni o‘rgatishga mo‘ljallangan platforma ishlab chiqildi. Bunday loyihalarning yana davom etishida bizning jamoamizda etarlicha salohiyat mavjud. Shu asosda aytish mumkinki, kelgusida bizning jamoamiz professor-o‘qituvchilari, yosh olimlar, doktorant va magistrantlar ishtirokida mana shunday elektron resurslarni ishlab chiqish bo‘yicha ilmiy izlanish va amaliy loyihalarning amalga oshirilishiga ishonamiz va konferensiya ishiga omad tilaymiz!

Sultonbek Ermamatovich Normamatov,
filologiya fanlari doktori, professor

O‘ZBEK TILINING FONEMATIK-FONEMIK SISTEMASI XARITASINI YARATISHNING NAZARIY ASOSLARI

Mengliyev Baxtiyor Rajabovich,
Filologiya fanlari doktori, professor
bakhtiyormingliyev62@gmail.com
TAFU, O‘zbekiston

Annotatsiya. Maqolada o‘zbek tilining fonematik-fonemik sistemasi xaritasini nazariy hamda amaliy jihatdan puxta tahlil qilish, uning zamonaviy lingvistika, xususan, kompyuter lingvistikasidagi o‘rnini aniqlash maqsad qilingan. Tilning fonema va allofonlar qurilishi, singarmonizm va uning bahsli jihatlari, fonologik qoidalar, dialektlar orasidagi fonetik tafovutlar va bu masalalarni hozirgacha o‘rganishda yuzaga kelgan metodologik muammolar (rus tilshunoslik an’analari, eroniy tillar ta’siriga oid iddaolar, eksperimental-empirik tadqiqotlar yetishmasligi, fonetik atlas va transkripsion alifbolarning ommalashmagani) hamda ularni raqamli muhitga tatbiq qilish imkoniyatlari haqida fikr yuritiladi. Shuningdek, turli lingvistik nazariyalar (generativ fonologiya, strukturalizm va boshqa) asosida o‘zbek tilining boy fonetik xususiyatlarini formal modelda aks ettirish masalalari muhokama qilinadi. Tadqiqot yakunida tilni raqamli qayta ishlash (ASR, TTS, mashina tarjimasi) hamda lingvistik tadqiqotlarda foydalanish mumkin bo‘lgan fonemik “xarita”ni yaratish bo‘yicha nazariy tavsiyalar taklif etiladi.

Abstract. This research aims to provide a comprehensive theoretical and practical analysis of the phonematic-phonemic system map of the Uzbek language, highlighting its role in modern linguistics and particularly in computational linguistics. The paper examines the construction of phonemes and allophones, the controversial aspects of vowel harmony (singarmonism), phonological rules, dialect-based phonetic divergences, and methodological issues that have arisen due to (1) the continuation of non-scientific methods inherited from older Russian linguistics, (2) the dominance of claims about Iranian language influence, (3) the prioritization of views held by Iranian or “Iranicized” scholars, (4) insufficient experimental-empirical research using specialized phonetic apparatus, (5) the absence of an Uzbek phonetic atlas, and (6) the limited adoption of Uzbek transcription alphabets and dictionaries. Drawing on various linguistic theories (generative phonology, structuralism, etc.), the paper explores how the rich phonetic features of Uzbek can be rendered in a formal model. Concluding remarks offer theoretical recommendations for developing a phonemic “map” that could benefit both linguistic research and digital language processing systems such as ASR, TTS, and machine translation.

Аннотация. В данной работе проводится всесторонний теоретический и практический анализ карты фонематико-фонемической системы узбекского языка, а также её роли в современной лингвистике и в частности в компьютерной лингвистике. Рассматриваются вопросы, связанные с построением системы фонем и аллофонов, спорные аспекты сингармонизма (гармонии гласных), фонологические правила, диалектные фонетические различия, а также методологические проблемы, возникшие вследствие (1) сохранения несистемных методов, унаследованных от русской лингвистической традиции, (2) доминирования мнения о сильном влиянии иранских языков, (3) приоритета взглядов иранизированных учёных, (4) недостаточной экспериментально-эмпирической базы исследований с использованием специального фонетического оборудования, (5) отсутствия специального фонетического атласа узбекского языка, и (6) ограниченной распространённости транскрипционных алфавитов и словарей. Опираясь на различные лингвистические теории (генеративная фонология, структурализм и др.), автор демонстрирует, как богатые фонетические особенности узбекского языка могут быть формализованы. В заключении даются рекомендации по созданию «карты» фонемной системы для применения в лингвистических исследованиях и цифровых технологиях (ASR, TTS, машинный перевод).

Kalit so‘zlar: *fonema, allofon, fonematik sistema, fonologik qoidalar, singarmonizm, kompyuter lingvistikasi, strukturalizm, generativ fonologiya, eksperimental fonetika, metodologik muammolar, xarita*.

Kirish. O‘zbek tili turkiy tillar oilasiga mansub – boy fonetik va fonologik xususiyatlari bilan ajralib turadi. Tarixiy taraqqiyoti, eroniy tillar bilan chegaradosh hududlarda shakllangan dialektlari va uzoq vaqt davomida rus tilshunosligi an’analari ta’siri ostida o‘rganib kelingani natijasida turli metodologik qarashlar shakllangan. Shuningdek, tilni eksperimental-empirik jihatdan tadqiq etish ishlari yetarli darajada tizimli olib borilmagani ham ba’zi ilmiy xulosalarning bahsli bo‘lib qolishiga sabab bo‘lgan.

O‘zbek tilidagi fonema va allofonlarni aniqlash, singarmonizm masalasini kuzatish, dialektlar fonetikasi, rus va eroniy tillar ta’siriga oid tezislarni tanqidiy tahlil qilish hamda kompyuter lingvistikasida amalga oshirish mumkin bo‘lgan yechimlarni ko‘rib chiqadi. Bunda diqqat markazida turgan muammodan biri – singarmonizm qisman yoki butunlay yo‘qolgan degan iddaolarning qay darajada ilmiy asosga ega ekanligini ko‘rib chiqish. Bu ko‘pincha rus tili va rus tili bo‘yicha mutaxassislar tomonidan yaratilgan ishonchsiz metodologiya yoki eroniy tillar ta’siriga bog‘lab tushuntirilib, o‘zbek tili dialektal kengligidagi real fonetik manzaraga yetarlicha e’tibor berilmagan holatlarga tayanadi.

1. “Fonematik va fonemik tizim” tushunchasi. Fonema – tilning eng kichik ma’no farqlovchi lisoniy birligi. Tilshunoslik tarixida bu tushuncha strukturalizm, Amerika deskriptiv lingvistikasi va generativ fonologiya singari qator yirik yo‘nalishlar tomonidan turlicha ifodalangan:

– strukturalizm (Praga maktabi)da fonemalar zidlanishlar orqali tahlil qilingan;

– Amerika deskriptiv lingvistikasida fonemaning distributsiyasiga asoslanish, minimal juftliklar tahlili amalga oshirilgan;

– generativ fonologiya fonemalarning “tub” va “yuza” shaklini fonologik qoidalar bilan bog‘laydi;

– optimallik nazariyasi (optimality theory)da talaffuz variantlari cheklovlar iyerarxiyasi orqali “optimal” shaklga keltirilgan;

– o‘zbek tilshunosligida fonema masalalari XX asrda qisman tadqiq etilgan. Ammo ko‘p hollarda ushbu tadqiqotlar an’anaviy usullar bilan cheklangan yoki rus lingvistik paradigmasi doirasida bo‘lgan.

Ba’zi manbada “fonemik” istilohi fonemalar tizimiga xos mavhum qatlam, “fonemik” real talaffuzdagi allofonik ko‘rinishlar sifatida ajratiladi. O‘zbek tilining fonematik-fonemik xaritasi deganda esa nafaqat fonemalar ro‘yxati, balki ularning kontekstual o‘zgarishlari, dialektal tafovut va eksperimental ishlar to‘liq bayon qilingan tizim tushuniladi.

2. O‘zbek tilidagi fonemik inventar. O‘zbek tilida 6 ta unli fonema mavjudligi rasman e’tirof etiladi: /a, e, i, o, u, o‘/ (va dialektlarda /ɯ/, /æ/ kabi unlilarning kuzatilishi mumkin):

– kirill yozuvida: <а, э, и, о, у, ў>

– lotin yozuvida: <a, e, i, o, u, o‘>

Bu unlilar turli kontekstda [ɑ], [ɛ], [æ], [ɪ], [ɔ], [ʊ] kabi allofonlarga ham ega bo‘lishi mumkin. Ayniqsa, dialektlarda unli inventarining nozik farqlari mavjud, bu eksperimental fonetika yordamida aniqroq qayd etilishi lozim. Ammo hozirgacha maxsus fonetik apparatlar (spektrograf, formant analizator va h.) bilan tizimli tadqiqotlar yetarlicha olib borilmagani hamon dolzarb muammoligicha qolayotir. O‘zbek tilida quyidagi undoshlar asosan ajratiladi:

– jarangsiz portlovchilar: /p, t, k, q/

– jarangli portlovchilar: /b, d, g, g‘, r, l/

– jarangsiz sirg‘aluvchilar: /s, ʃ, x, h/

– jarangli sirg‘aluvchilar: /z, ʒ, ɣ/ (yoki ba’zi shevalarda /ʁ/)

– burun tovushlar: /m, n, ŋ/

– yarim unli: /j/ (yoki /y/)

Dialektlarda /v/–/w/ farqlanishi, /h/ yoki /x/ talaffuzidagi o‘ziga xosliklar, yumshoq /kʲ/ kabi variantlar kuzatiladi. Lekin bu farqlanishlar ko‘pincha eroniy tillar ta’siri yoki “o‘zbek tilining singarmonizmdan uzoqlashganligi” bilan bog‘lab tushuntirilgani holda yetarli eksperimental dalil keltirilmaganligi sababli

munozaralar davom etmoqda. Bir fonema turli kontekstda har xil akustik xususiyatlar bilan ifodalanishi mumkin. Masalan: /k/ → [k], [kʲ], /g‘/ → [ɣ], [ʁ], /o/ → [ɔ], [o], [ʊ].

Allofonlarni aniq qoidalar bilan tavsiflash kompyuter lingvistikasi uchun ham, adabiy me’yorni targ‘ib qilish uchun ham muhim.

3. Fonologik nazariya va o‘zbek tili. Roman Yakobson va Morris Halle boshchiligidagi strukturalizm fonemalarni minimal xususiyatlar (distinctive features) orqali izohlashni taklif qiladi [1]. O‘zbek tiliga nisbatan qo‘llaganda bu jaranglilik, lablanish, til orqa/old pozitsiyasi, portlovchi/sirg‘aluvchi kabi xossalar asosida soddalashtirilgan tasnifni yaratish imkonini beradi. Xomskiy, Halle [1968] fonologik qoidalar orqali fonemalarning “tub” va “yuza” shakllarini bog‘lash metodi bilan mashhur [2]. O‘zbek tilida bu yondashuv singarmonizm va undosh uyg‘unlashuvi kabi jarayonlarni qoidalash orqali tushuntirishga imkon beradi.

Maqbullik nazariyasi (optimality theory) [3] talaffuz variantlarini “cheklovlar” bilan baholab, eng maqbul (“optimal”) shaklni tanlaydi. Agar o‘zbek tilidagi singarmonizm cheklov darajasida ustun bo‘lsa, unda unlilar uyg‘unligi saqlanadi; aks holda so‘z o‘zgartirilgan shaklda talaffuz qilinadi. Ammo bu nazariyani qo‘llash uchun ham korpus va eksperimental tadqiqotlar lozim.

4. Singarmonizm va undosh uyg‘unlashuvi: bahslar va asl holat. Turkiy tillarda, odatda, unlilar uyg‘unligi (vowel harmony) kuchli saqlangan [7]. Ko‘pincha “o‘zbek tili bu uyg‘unlikni yo‘qotgan yoki juda zaiflashtirgan” degan iddao oldinga suriladi. Biroq bu xulosa quyidagi omillar jihatidan bir qadar bahsli:

– rus tili va rus tili bo‘yicha mutaxassislar qarashlari. O‘zbek tilini, ayniqsa, XX asrda rus tili metodologiyasiga suyangan holda o‘rganish ortidan “singarmonizm deyarli yo‘q” degan qarash keng tarqalgan. Vaholanki, turli shevalar tahlili shuni ko‘rsatadiki, unlilar uyg‘unligi mavjud;

– eroniy tillar ta’siriga oid tezis. Ba’zi tadqiqotchilar o‘zbek tilidagi singarmonizm zaiflashuvini eroniy tillar ta’siri bilan bog‘laydi. Lekin bu ta’sirning real darajasini kuzatish uchun maxsus eksperimental-empirik tadqiqotlar deyarli o‘tkazilmagan.

O‘zbek tilida singarmonizm. Adabiy til nuqtayi nazaridan singarmonizm deyarli sezilmaydi degan qarashlar ishonchsiz va asossiz. Shuningdek, Farg‘ona vodiysi yoki ayrim janubiy hududlarda unlilar uyg‘unlashuvi aralash shakllarda saqlangan. Masalan, “bor + di” = [bordi], “kel + di” = [keldi] kabi oddiy yondashuv bilan cheklanmay, unlining o‘zgarish doirasi kengroq qaralsa, masalan, [kældi], [kʲaldi] kabilar kuzatilishi mumkin. Bular, albatta, muntazam fonologik qoida darajasidami yoki shunchaki fonetik variantlarmi – bu masala hamon chuqur laborator tahlilga muhtoj.

O‘zbek tilida /k/, /g‘/ tovushlarining yumshashi, qattiqlik-yumshoqlik ziddiyatiga tobe bo‘lishi, jarangsiz va jarangli undoshlarning uyg‘unlashuvi kabi jarayonlar kuzatiladi. Bular ko‘pincha:

– so‘z boshi – so‘z o‘rtasi – so‘z oxiri;
– old unli – orqa unli;
– tutash undoshlar orasidagi assimilyatsiya bilan bog‘liq. Ammo bu jarayonlarning real ko‘lami hamma shevada bir xil emas. Shuning uchun bir tekisda mazkur hodisani “yo‘qolgan” deb bo‘lmaydi. Shu o‘rinda quyidagi omillarni ham alohida ta’kidlash kerak:
– XX asr boshidan rus tili modeli asosida shakllangan fonetik-fonologik tahlil usullari ba’zan turkiy tillarning o‘ziga xos, jumladan, singarmonizm kabi parametrlarga to‘liq mos kelmaydi. Shu bois “singarmonizm yo‘qoldi”, - degan xulosa tez-tez uchraydi [6];
– O‘zbek tili XX asr tadqiqotlarida, asosan, Markaziy Osiyoda tarqalgan fors-tojik tillarining kuchli ta’siri ostida fonetik innovatsiyalarni qabul qilgan degan taxmin ustun turdi. Lekin haqiqiy ta’sir darajasini sinchiklab eksperimental tadqiq etish, ayniqsa, chuqur sheva tahlili bilan isbotlash yetarlicha amalga oshirilmagan;
– ayrim tadqiqotchilar (o‘zlari forsiy til fonetikasi yoki eroniy filologiyaga ixtisoslashgan) tabiiy ravishda turkiy fonetik xususiyatlarni ikkinchi darajaga tushirib, eroniy xususiyatlarni kattaroq rol o‘ynaydi deb hisoblashga moyil bo‘ladi;
– o‘zbek shevalari fonetikasini spektrograf, formant analiz, intonatsion tahlil usullari bilan tizimli ravishda o‘rganish ishlari juda kam. Empirik ma’lumotlar bo‘lmasa, har qanday xulosa subyektiv bo‘lib qoladi [5];
– agar butun respublika shevalarini qamrab olgan fonetik atlas bo‘lganida singarmonizmning saqlangan yoki yo‘qolgan nuqtalari aniq xaritada ko‘rinar edi;
– keng jamoatchilik ham, ilmiy doiralar ham fonetik transkripsiyaga oid maxsus resurslardan kam foydalanadi. Bu esa aniq fonologik tahlilga xalal beradi.

Bu omillar o‘zbek tilidagi singarmonizm va, umuman, fonologik jarayonlar haqida bir tomonlama yoki yetarlicha asoslanmagan xulosalarning davom etishiga zamin yaratgan.

5. Kompyuter lingvistikasi doirasida fonematik-fonemik xarita. O‘zbek tilining fonematik-fonemik sistemasi xaritasini raqamli muhitda qo‘llash kreak. Buning uchun:
– bir ma’noli kodlash – har bir fonema (va zarur bo‘lsa, allofon) aniq belgiga ega bo‘lishi zarur;
– kontekstual qoida – singarmonizm, undoshlar uyg‘unlashuvi, allofonik o‘zgarishlar maxsus qoidalarga asoslanib, formal tarzda ifodalanishi lozim;
– moslashuvchanlik – dialektal farqlar va yangi eksperimental ma’lumotlarni qo‘shish imkoniyati hisobga olinadi;
– ilmiy asosli ma’lumotlar – metodologik xatolarni takrorlamaslik uchun eksperimental fonetika ko‘rsatkichlari kiritilishi zarur;
– standartga moslilik – o‘zbek adabiy talaffuz me’yoriga ham zid bo‘lmagan shaklda qurish lozim.

Muammo integrativ tabiatga ega. Masalan:

– nutqni tanish (ASR) da ovozdan so‘zlarni aniqlashda fonema va allofonlar tizimi aniq belgilansa, aniqlik oshadi;

– matndan nutq hosil qilish (TTS) da so‘zlarni fonetik transkripsiyalash va tabiiy talaffuzga yaqin ovoz sintez qilish uchun zarur;

– mashina tarjimasida fonologik xususiyatlar inobatga olingan morfologik tahlil moduli natijani yaxshilaydi;

– orfoepiya va orfografiyada so‘zlarni to‘g‘ri talaffuz qilish, imlo yo‘riqnomalariga muvofiq yozish kabi dasturiy yordamchilar uchun zarur.

6. Fonematik-fonemik xarita loyihasini yaratish bosqichlari sifatida quyidagilarini sanash mumkin:

– respublikadagi barcha hududlardan audio yozuv, tarixiy matnlar va avvalgi tadqiqotlar to‘planadi;

– spektrografik, formant, intensiya, tembr tahlillari yordamida real talaffuz xususiyatlari aniqlanadi;

– dialektlarda uchraydigan fonemalar, adabiy me’yordan chetlanish va allofonlar xaritasi yaratiladi;

– singarmonizm, undosh uyg‘unlashuvi, eliziyalar (tovush tushishi), epentezlar (tovush qo‘shilishi) kabi jarayonlar formal bayon etiladi;

– X-SAMPA yoki IPA asosida transkripsiya jadvali, qoidalarning XML/JSON shakldagi ro‘yxati shakllantiriladi;

– ASR, TTS yoki mashina tarjimasi modullarida qo‘llab, aniqlik ko‘rsatkichlari baholanadi;

– aniqlangan kamchiliklar ustida ish olib boriladi va dialektal ma’lumotlar kengaytiriladi.

Xulosa. Ma’ruza o‘zbek tilining fonematik-fonemik sistemasini xaritaga solish masalasida an’anaviy nazariyalar (strukturalizm, generativ fonologiya, qulaylashtirish nazariyasi va zamonaviy eksperimental fonetika) ning qanchalik zarurligini ko‘rsatib berdi. Ayniqsa, singarmonizm bo‘yicha “saqlangan” yoki “yo‘qolgan” degan qarashlarning birmuncha bahsli ekani va ushbu masalada: 1) rus tilshunoslik metodologiyasidan qolgan yondashuvlar; 2) eroniy tillar ta’siriga oid iddaolarning ustuvorligi; 3) “eroniylashgan” ba’zi olimlar qarashlarining keng yoyilgani; 4) eksperimental-empirik izlanishlar yetishmasligi; 5) fonetik atlas yaratilmaganligi; 6) ommaviy transkripsion resurslar (alifbo, lug‘at) ommalashmaganligi kabi omillar katta rol o‘ynayotganiga urg‘u berildi.

Ilmiy-amaliy yechimlar sifatida quyidagilar ta’kidlanadi:

– metodologiyani yangilash (rus yoki boshqa tashqi metodologiyalarni ko‘r-ko‘rona qo‘llashdan voz kechish, turkiy tillar xususiyatiga xos yondashuvlarga e’tibor qaratish);

– eksperimental tadqiqotlar kengaytirilishi (fonetika laboratoriya, portativ spektrograf va audio ma’lumotlar bazasidan foydalanish, olingan natijalarni ilmiy jamoatchilikka muntazam taqdim etish);

– fonetik atlas yaratish (mintaqa bo‘yicha shevalarni tizimli xaritalash, singarmonizm saqlangan yoki zaiflashgan nuqtalarni aniq belgilash);

– transkripsion resurslarni ommalashtirish (o‘zbek tilining transkripsion lug‘atlarini yaratish va onlayn platformalarda foydalanish imkoniyatlarini kengaytirish);

– kompyuter lingvistikasi bilan integratsiya (ASR, TTS, mashina tarjimasi modullariga singarmonizm va undosh uyg‘unlashuvi qoidalarini integratsiya qilish, shu orqali texnologik infratuzilmani takomillashtirish).

Shunday qilib, o‘zbek tilining fonematik-fonemik xaritasi nafaqat filologik ilm, balki amaliy (dasturiy) ahamiyatga ham ega. Bu bilan tilning raqamli qayta ishlash imkoniyatlari kengayib, nutqni tanish, sintez qilish, tarjima kabi ko‘plab sohada sifat va samaradorlikni oshiradi. Shu bilan birga, milliy lingvistik merosni eksperimental yondashuv orqali chuqur o‘rganish orqaligi ba’zi bahsli masalalarga (xususan, singarmonizmning saqlanganmi yoki yo‘qolganmi degan) to‘liq aniqlik kiritish mumkin bo‘ladi.

Foydalanilgan adabiyotlar:

1. Jakobson R., Fant G., Halle, M. Preliminaries to Speech Analysis. – Cambridge, MA: MIT Press, 1952.
2. Chomsky N. Halle M. The Sound Pattern of English. – New York: Harper & Row, 1968.
3. Prince A., Smolensky, P. Optimality Theory: Constraint Interaction in Generative Grammar. Rutgers University Center for Cognitive Science, 1993.
4. Trubetzkoy N.S. Grundzüge der Phonologie. – Prague: Travaux du Cercle Linguistique de Prague, 1939.
5. Mirtojiyev M. O‘zbek tili fonetikasi. – Toshkent: Mumtoz so‘z, 2015.
6. Xudoyberdiyev J. O‘zbek adabiy tilidan singarmonizmning chiqarilishi. Toshkent: Akademnashr, 2024.
7. Коджаоглу Т. К проблеме небной гармонии гласныхв узбекском языке // Туркология. – Туркистан, 2017.

INTERNET ASOSIDA TIL O‘RGATISHDA KOMPYUTER TEXNOLOGIYALARI ORQALI PRAGMATIK KOMPETENSIYANI RIVOJLANTIRISH

Toirova Guli Ibragimovna,
filologiya fanlari doktori(DSc), professor
tugulijon@mail.ru
BuxDU

Annotasiya: Mazkur maqolada internet asosida til o‘rgatish jarayonida kompyuter texnologiyalari vositasida talabalarning pragmatik kompetensiyasini shakllantirish va rivojlantirish masalalari tahlil qilingan. Shuningdek, onlayn ta’limda yuzaga keladigan pragmatik to‘siqlar va ularni bartaraf etishda multimedia resurslari, sun’iy intellekt, interaktiv platformalardan samarali foydalanish yo‘llari ko‘rsatib berilgan.

Annotation: This article analyzes the development of students’ pragmatic competence through computer technologies in language teaching based on the internet. It also presents the ways of overcoming pragmatic barriers in online education using multimedia resources, artificial intelligence, and interactive platforms.

Аннотация: В данной статье анализируются вопросы формирования и развития прагматической компетенции студентов с помощью компьютерных технологий в процессе обучения языку на основе интернета. Также представлены пути преодоления прагматических барьеров в онлайн-обучении за счёт эффективного использования мультимедийных ресурсов, искусственного интеллекта и интерактивных платформ.

Kalit so‘zlar. *pragmatik kompetensiya, kompyuter lingvodidaktikasi, internet ta’lim, madaniyatlararo kommunikasiya, elektron resurslar, sun’iy intellekt.*

KIRISh. Internet texnologiyalari rivojlanishi oxirgi o‘n yilliklarda ta’lim tizimida, ayniqsa, chet tillarini o‘rgatish sohasida jiddiy o‘zgarishlar keltirib chiqardi. Jahon miqyosida raqamli ta’lim platformalarining kengayishi, interaktiv onlayn kurslar va sun’iy intellektga asoslangan o‘quv ilovalarining yaratilishi natijasida til o‘rgatish nafaqat qulay, balki global darajada ommaviy tus oldi. 2023-yilda UNESCO tomonidan e’lon qilingan ma’lumotlarga ko‘ra, dunyo bo‘yicha 1 milliarddan ortiq kishi internet orqali turli tillarni o‘rganayotgani qayd etilgan. Shuningdek, Coursera, Duolingo, BBC Learning English kabi platformalarning foydalanuvchilar soni yil sayin o‘sib bormoqda.

Pragmatik kompetensiyani shakllantirish, ya’ni tilni muayyan ijtimoiy va madaniy kontekstda to‘g‘ri qo‘llash qobiliyati, bugungi global dunyoda yanada

muhim ahamiyat kasb etmoqda. Xususan, xalqaro ishchi muhitda, ta’lim tizimida va madaniy almashinuv loyihalarida noto‘g‘ri nutqiy munosabatlar ko‘pincha madaniyatlararo tushunmovchiliklarga olib keladi. Shu munosabat bilan, til o‘rgatish jarayonida kompyuter texnologiyalaridan foydalangan holda talabalarni real muloqot vaziyatlariga tayyorlash bugungi kun talabidir.

O‘zbekiston miqyosida ham internet orqali til o‘rgatish sohasi keng quloch yoymoqda. Respublikada 2022-yilda Prezident tashabbusi bilan "Zamonaviy kasb-hunar va xorijiy tillarni o‘rgatish markazlari" tashkil etildi va ularda zamonaviy raqamli texnologiyalar asosida o‘qitish amaliyoti yo‘lga qo‘yildi. Bundan tashqari, Oliy ta’lim, fan va innovasiyalar vazirligining 2023-yil hisobotida qayd etilishicha, onlayn til kurslarida qatnashgan talabalar soni 300 ming nafardan oshib ketgan. Ammo bu jarayonda pragmatik kompetensiyani shakllantirishga etarlicha e’tibor qaratilayotgani yo‘q [19].

Adabiyotlar tahlili. Internet texnologiyalarining til o‘rgatishdagi axamiyati va pragmatik kompetensiyani rivojlantirishga ta’siri haqidagi tadqiqotlar soni kun sayin ortib bormoqda. Modern ta’lim vositalari, xususan, sun’iy intellekt (AI) va kompyuter lingvodidaktikasi imkoniyatlarining til o‘rgatish jarayoniga kiritilishi bu boradagi nazariy va amaliyot tadqiqotlarini yangi bosqichga olib kelgan. Pragmatik kompetensiya, til o‘rgatishda faqat grammatik va leksik bilimlar bilan cheklanmay, balki tilni madaniy va ijtimoiy kontekstda to‘g‘ri qo‘llash qobiliyatini ham o‘z ichiga oladi. Bu borada Smit va Braun o‘z ilmiy tadqiqotlarida pragmatik kompetensiyaning til o‘rgatishdagi ahamiyatini va internet texnologiyalarining ushbu jarayonga qanday ta’sir ko‘rsatishini tahlil qilganlar [15]. Ular internet platforma va multimedia resurslarining pragmatik kompetensiyani rivojlantirishdagi rolini quyidagicha ta’riflaganlar: “Tilni interaktiv usullar va real vaqtdagi muloqotlar orqali o‘rganish talabalarga tilning amaliyotdagi foydalanishini yaxshi tushunishga yordam beradi”[15].

Kompyuter lingvodidaktikasi, shu jumladan, virtual muhitlar va avtomatik pragmatik tahlil platformalarining o‘zlashtirish jarayonidagi ahamiyatiga alohida e’tibor berilgan. Kelli o‘z tadqiqotida kompyuter yordamida pragmatik tahlilning til o‘rganuvchilarga qanaqa yangicha echimlar taqdim etadiganligini ko‘rsatgan. Uning fikricha, “sonlar va ma’lumotlar orqali tilning pragmatik qiyosini aniqlash va o‘rganish talabalarga ko‘p jihatli faoliyatlarni amalga oshirishga yordam beradi” [10]. Kompyuter lingvodidaktikasi orqali tilni o‘rganish, fikrlash, muloqot qilish va rivojlantirish imkoniyatlarini takomillashtirishga imkon yaratadi.

Multimedia resurslar va interaktiv o‘qitish metodlari til o‘rgatishdagi axamiyatini ko‘rsatish uchun bir qancha tadqiqotlar olib borilgan. Chang va Li o‘z tadqiqotlarida multimedia va virtual platformalardagi amaliy o‘qitish usullarining, shu jumladan, geymifikasiya va video-dramalarning, pragmatik kompetensiyani rivojlantirishda qanday ta’sir ko‘rsatishini ko‘rsatganlar. Ularning fikricha, “interaktiv resurslar tilni nafaqat grammatik, balki ma’naviy va ijtimoiy kontekstda

ham to‘g‘ri qo‘llashga imkon beradi” [4]. Internet asosidagi til o‘rgatish jarayonida pragmatik to‘siqlar ham ro‘y berishi mumkin. Bu borada Li olib borgan tadqiqotda, “internet ta’limdagi pragmatik to‘siqlar – bu talaba va o‘qituvchi orasidagi kommunikasiya to‘siklari, ayniqsa, madaniy va ijtimoiy farqlar orqali ro‘y beradi” deb ta’riflangan[12]. Bu tadqiqot pragmatik to‘siqlarni bartaraf etishda kompyuter texnologiyalarining ahamiyatini ko‘rsatadi. Til o‘rgatishdagi yangi usullar, shu jumladan, sun’iy intellekt asosidagi avtomatik pragmatik tahlil va interaktiv resurslarni ishlab chiqishdagi innovasion yondashuvlar, o‘zbekistonida ham joriy etilmoqda. Shunday qilib, pragmatik kompetensiyani rivojlantirish uchun zamonaviy metod va vositalarning ilmiy tahlili, ularning amaliyotga tadbiq qilinishi, til o‘rganish jarayonidagi aniq maqsadlarga erishishda muhim rol o‘ynaydi

Pragmatik kompetensiya – bu tilni muayyan ijtimoiy va madaniy vaziyatda maqsadga muvofiq, ma’noli va ta’sirchan tarzda qo‘llash qobiliyatidir. U leksik va grammatik bilimlardan tashqari, nutqning funksional maqsadi, kontekstga mos niyat va madaniy hassoslikni o‘z ichiga oladi. Masalan, ingliz tilida “Could you open the window, please?” degan iltimos jumlasi grammatik jihatdan to‘g‘ri bo‘lishi mumkin, ammo talabaning madaniy muhitni inobatga olmagan holda bevosita va qattiqroq formada “Open the window!” deyishi kommunikasiyada noqulayliklar keltirib chiqarishi mumkin.

Jahon miqyosida til o‘rgatish metodikasida CEFR (Common European Framework of Reference for Languages) standarti pragmatik kompetensiyani alohida komponent sifatida belgilagan. CEFRga ko‘ra, yuqori darajada til egallagan shaxs faqat grammatik to‘g‘rilik emas, balki muloqot sharoitiga mos nutq strategiyalarini ham samarali qo‘llash qobiliyatiga ega bo‘lishi kerak [6;17].

O‘zbekistonda ham chet tillarini o‘rgatish jarayonida pragmatik bilimlarga bo‘lgan ehtiyoj ortib bormoqda. Buni joriy etilayotgan yangi darsliklar va multimediyali onlayn o‘quv platformalari orqali kuzatish mumkin. Ammo amalda aksariyat kurslar grammatik va leksik ko‘nikmalarni shakllantirishga yo‘naltirilgan bo‘lib, madaniy-ijtimoiy komponentlar etarli darajada qamrab olinmagan. Shuning uchun internet asosidagi kompyuter texnologiyalarini pragmatik o‘qitish maqsadida to‘g‘ri yo‘naltirish muhim vazifalardan biri bo‘lib qolmoqda.

Yangi avlod ta’lim platformalari – Moodle, Google Classroom, Duolingo, ChatGPT, Edmodo va ushbu turkumga kiruvchi boshqa interaktiv vositalar – til o‘rgatishda pragmatik kompetensiyani shakllantirishga xizmat qiluvchi muhim vositalardan biri hisoblanadi. Ushbu platformalarning afzalligi shundaki, ular orqali talabalar real muloqotga yaqin bo‘lgan simulyasiyalangan vaziyatlar, rolli o‘yinlar, muloqot ssenariylari va madaniy kontekstga mos muloqotlarni boshdan kechirish imkoniyatiga ega bo‘ladilar. Masalan, Duolingo ilovasida foydalanuvchi real hayotdagi muloqotlarga asoslangan dialoglarda ishtirok etishi mumkin – bu esa talabaning tilni faqat grammatik jihatdan emas, balki pragmatik jihatdan ham to‘g‘ri qo‘llashini rag‘batlantiradi[13].

ChatGPT kabi sun’iy intellekt asosidagi vositalar orqali talabalar turli madaniy kontekstda turli muloqot strategiyalarini mashq qilishi mumkin. Masalan, bir xil so‘z aytilishi mumkin bo‘lgan turli vaziyatlarni (rasmiy/norasmiy, yosh katta/yosh kichik bilan muloqot) modellashtirish va turli pragmatik yondashuvlarni solishtirish mumkin. Bu esa pragmatik farqlarni inobatga olish qobiliyatini shakllantiradi[16]. Taguchi ta’kidlaganidek, pragmatik kompetensiyani tarbiyalashda simulyativ muloqotlar va sun’iy muhitdagi mashqlar yuqori samara beradi[16].Shu bilan birga, Moodle va Google Classroom kabi platformada o‘qituvchi tomonidan maxsus tayyorlangan multimediyali materiallar, audio va video materiallar, janrlarga oid tahlil topshiriqlari orqali tilning turli uslubiy va pragmatik xususiyatlarini o‘rgatish mumkin. Masalan, rasmiy maktub yozish, blog yozish, elektron xat orqali muloqot kabi turli janrlarda talabalarni tayyorlash orqali ularda ma’noli va maqsadga yo‘naltirilgan nutq strategiyalari shakllanadi.

Elektron platformalarning yana bir afzalligi – talabalar faoliyatini avtomatik baholash, muloqotdagi xatolarni aniqlash va pragmatik masalalarda to‘g‘ri yo‘l ko‘rsatish imkoniyatidir. Jahon bankining 2022-yil ta’lim bo‘yicha hisobotida ta’kidlanganidek, raqamli vositalar orqali tashkil etilgan ta’lim jarayonida talabalarning nutqiy faolligi 37% ga ortadi va bu pragmatik ko‘nikmalarni rivojlantirishda muhim omil hisoblanadi [18] .

Onlayn muloqotda madaniyatlararo omillar til o‘rgatish jarayonida asosiy rol o‘ynaydi, chunki til faqat grammatik va leksik elementlardan iborat emas, balki u shaxslar va jamiyatlarning turli madaniy qiymatlarini, ishoralarini va qarashlarini ham o‘z ichiga oladi. Til va madaniyat bir-biri bilan yaqin bog‘liq bo‘lgan, shuningdek, har bir madaniyatning o‘zgacha til ishlatish shakllari va xususiyatlari mavjud. Internet platformada amalga oshiriladigan muloqotlar bu farqlarni tushunish va qabul qilish imkonini beradi.

Madaniyatlararo kompetensiya (Ing. intercultural competence) – bu odamning turli madaniyatlar orasida faol va samarali muloqot o‘rnatish qobiliyatidir. Onlayn muloqotda buni ta’minlash uchun talabalarga turli madaniy kontekstlarda bo‘lib o‘tadigan so‘zlashuvlar, ijtimoiy me’yorlar, bag‘rikenglik va fahmlashni o‘rgatish kerak. Onlayn platformalar orqali turli madaniy ishoralarning to‘g‘ri tushunilishini ta’minlash uchun video-dramalar, rol o‘yinlari va multimedia kontentlar muhim vosita bo‘lib xizmat qiladi. Masalan, onlayn platformadagi videodramalar orqali talabalar turli madaniy kontekstlar bilan tanishishadi. Bunday videomateriallar orasida, odatda, rasmiy va norasmiy muloqotlar, egalik qilish va ijtimoiy xatolarni engishga qaratilgan ssenariilar bo‘lishi mumkin. Videodramalar talabalarga muloqotdagi ijtimoiy vaziyatlarni, mavjud ijtimoiy madaniyatlarni chuqurroq tushunish va ularning ijtimoiy kontekstdagi ahamiyatini aniqlash imkoniyatini beradi.

Rol o‘yinlari onlayn ta’limda interkultural kompetensiyani rivojlantirishda alohida ahamiyatga ega. Ushbu o‘yinlarda talabalar turli madaniy vaziyatlarda

qatnashishadi va ushbu vaziyatlarga mos keladigan ijtimoiy va pragmatik kodifka uyg‘unlashishni o‘rganishadi. Masalan, kimdir biznes yoki rasmiy muloqotlarda qatnashishi mumkin, boshqasi esa fuqarolik xuquqlarini muhokama qilishda yoki kundalik hayotdagi o‘zgarishlarni qabul qilishda qatnashadi. Bunday mashqlar orqali talabalar tadqiqot va muloqot strategiyalarini moslashtirishga o‘rganadilar. Multimedia kontentlar, xususan, videolar, interaktiv o‘zgarishlar va animasiyalar, talabalarga tilning turli madaniy janrlarini o‘rganish imkonini beradi. Multimedia tasvirlarini va videomateriallarni taqdim etish til o‘rgatish jarayonini aniq va dinamik qiladi, shunday qilib talabalarga turli xil madaniy kontekstlarda o‘z bilimini amaliy ravishda qo‘llashga yordam beradi. Bundan tashqari, o‘rganilayotgan tildagi interkultural ogohlik, shundaki, talabalar to‘g‘ri ishlatish va o‘zgarishlarga moslashishni o‘rganadilar. Bu borada bir qator tadqiqotlar amalga oshirildi. Masalan, 2021-yildagi tadqiqotga ko‘ra, onlayni dasturlar va multimedia kontentlari orqali madaniyatlararo kompetensiyani rivojlantirish, nafaqat o‘rganilayotgan tilga oid bilimlarni, balki o‘zgaruvchan ijtimoiy va madaniy muhitda amalga oshiriladigan muloqotlar malakasini ham oshirdi [2,3]. Shuningdek, ushbu tadqiqotlar, video-dramalar va rolli o‘yinlar talabalarning tildagi pragmatik kompetensiyasini shakllantirishda muhim ahamiyatga ega ekanligini ko‘rsatdi.

Pragmatik to‘siq — bu til o‘rganish va muloqot jarayonida paydo bo‘ladigan muammolar, ular insonning tilni ijtimoiy va madaniy kontekstda to‘g‘ri va samarali ishlatish qobiliyatiga to‘sqinlik qiladi. Pragmatik to‘siqlar til o‘rganuvchining so‘zlashuvdagi samaradorligini, aniqlikni va mazmunni to‘g‘ri etkazishini murakkablashtiradi.

Rus tilidagi "Как с гуся вода" va o‘zbek tilidagi "Suvdan quruq chiqmoq" iboralari tilshunoslikda ikki xalqning madaniy va pragmatik kontekstiga mos keladi. Ular o‘xshash ma’no va maqsadga ega bo‘lib, insonning vaziyatlarga nisbatan barqarorligini, stressga e’tibor bermasligini va muammolardan omad bilan chiqib ketishini bildirish uchun ishlatiladi.

Rus tilida "Как с гуся вода" iborasi g‘ozlarning tabiiy xususiyatiga asoslangan bo‘lib, ular suvda suzishsa ham, o‘ziga xos yog‘langan parhalari tufayli hech qachon cho‘kmaydi. Bu ibora odamning muammolardan, tanqidlardan, xavflardan qandaydir ta’sir ko‘rmasdan, omad bilan o‘tishini ko‘rsatadi. Shuningdek, o‘zbek tilidagi "Suvdan quruq chiqmoq" iborasi ham shu ma’noni o‘zida mujassamlashtiradi, ya’ni inson muammolardan yoki xavfdan ta’sirlanmay, o‘ziga xos muqobillik va mas’uliyat bilan va kerakli reaksiya bilan vaziyatdan chiqishni bildiradi.

Shunday qilib, bu iboralar tilshunoslikda frazeologizmlarning amaliy va pragmatik ishlatilishining o‘xshashligini ko‘rsatadi. Buning orqali har bir tilda frazeologizmlar bilan ma’no to‘ldirilgan har bir xulosa, tilni amaliy va ma’naviy jihatdan qanchalik yaxshi izohlash va ularning madaniy kontekstlarini anglatishga yordam beradi.

O‘zbek va rus tillaridagi ayrim frazeologizmlar o‘xshash pragmatik mazmunga ega bo‘lishi mumkin, ular jamiyatdagi muayyan ijtimoiy va madaniy kontekstlarga asoslanib shakllanadi. Masalan, o‘zbek tilidagi "Bir burda nonga zor" va rus tilidagi "сидет на хлебе и воде" iboralari o‘rtasidagi ma’nolar va ularning pragmatik konteksti o‘xshashdir.

1. O‘zbek tilidagi "Bir burda nonga zor" iborasi, asosan, qiyin hayot shartlarida yashovchi yoki ehtiyojga muhtoj bo‘lgan shaxsni anglatadi. Bu ibora, insonning moddiy yoki ma’naviy yordamga ehtiyojini, uning hayotdagi kurashish jarayonidagi og‘irliklarini ta’riflaydi. Bu, o‘z navbatida, ijtimoiy munosabatlarda shaxsning yangicha holatga tushishini ko‘rsatib, uning har kungi kurashishini va ko‘proq ehtiyojni qondirish uchun bor kuchini ishga solishini ifodalaydi.
2. Rus tilidagi "сидет на хлебе и воде" iborasi esa, odamning qiyin vaziyatda, omadsizlik va kerakli manbalardan mahrum bo‘lishini, juda kam yoki engil taomlanishni anglatadi. Bu, asosan, shaxsning kurashishga va muammolarga qaramasdan, qiyin va kambag‘al hayot sharoitida yashashini belgilaydi.

Ikkala ibora ham ijtimoiy va madaniy kontekstiga mos ravishda insonning qiyin vaziyatlardan omad bilan chiqishini yoki ularga qarshi kurashishini ko‘rsatadi. Biroq, har ikki iboraning pragmatik mazmunidagi farqlar, tillar orasidagi nazariy va amaliyotga oid qo‘shimcha jihatlarning muhimligini namoyon etadi.

Shuningdek, ushbu frazeologizmlar tildagi mavjud metaforalarning amaliy va madaniy ko‘rinishlarini ifodalagan holda, tilshunoslik va pragmatikada frazeologizmlarning ro‘yxatini shakllantirishga hissa qo‘shadi. Bu iboralar, nafaqat tilning ko‘p qirraliligini, balki turli madaniy kontekstlarda qanday ma’lumotlar ishlatilishi mumkinligini ham ko‘rsatadi.

Pragmatik to‘siqlarni bartaraf etish uchun quyidagi usullar va metodikalar ishlatish mumkin:

1. Modellashtirish va rol o‘yinlari – onlayn platformalar orqali turli muloqot vaziyatlari simulyasiya qilinadi, bunda talabalar hayotiy vaziyatlarda ijtimoiy-mədaniy faktorlarni hisobga olib, turli situasiyalarda noto‘g‘ri ta’bir va adashishlarga yo‘l qo‘ymasdan muloqot olib borishga harakat qiladilar. Bunday interaktiv mashqlar pragmatik to‘siqlarni yo‘q qilishda samarali bo‘lishi mumkin. Masalan, Moodle va Google Classroom platformalarida talabalarga turli tillardagi muhokamalarni tartibga solishga yordam beruvchi vazifalar berilsa, ularning kommunikativ qobiliyatlari rivojlanadi.
2. Madaniy kontekstni tushunish – til o‘rganuvchiga, uzbek va rus tillaridagi frazeologizmlar, ifoda va shuningdek, so‘zlar va o‘rganilgan madaniyatning mazmunini taqdim etish muhimdir. Madaniy kontekstga oid materiallarni ishlatish (masalan, video-dramalar va multimedia kontentlar) talabalarga ikki tilda ham to‘g‘ri nutq va ifodaga ega bo‘lishni amalga oshirishga yordam beradi. Shunday usullar orqali pragmatik to‘siqlarni bartaraf etish uchun talabalarga har bir vaziyatdagi madaniy va ijtimoiy xususiyatlar tushuntiriladi.

3. Avtomatik mulohaza va sun’iy intellekt orqali yordam – Qurilgan kompyuter modellari, shu jumladan, sun’iy intellektga asoslangan avtomatik mulohaza berish tizimlari pragmatik xatolarni aniqlash va to‘g‘ri baho berish imkonini beradi. Masalan, ChatGPT yoki boshqa NLP (Tabiiy tilni qayta ishlash) tizimlari talabalarga muloqotdagi xatolarni aniqlab, ularni tahlil qilib, aniq tavsiyalar beradi.

4. Situativ modellashtirish – Pragmatik kompetensiyani rivojlantirish uchun talabalarga turli amaliy vaziyatlar modellashtiriladi, shuningdek, ularning muammolarni hal qilish qobiliyatini kuchaytirish maqsadida interaktiv va realistik ssenariylar ishlab chiqiladi. Bu yordam beradi, chunki til o‘rganuvchi har xil tildagi vaziyatlar uchun to‘g‘ri ishlatishni o‘rganadi, shuningdek, ularning fahmlovchi va kommunikativ ko‘nikmalarini mustahkamlaydi.

Xulosa qilib aytganda, internet asosidagi ta’limda pragmatik kompetensiyani rivojlantirish zamonaviy til o‘rgatishdagi muhim va ustuvor vazifalardan biri hisoblanadi. Bu jarayonda, kompyuter lingvodidaktikasi vositalarini, jumladan, multimedia texnologiyalarini, virtual muloqot muhitlarini va sun’iy intellektga asoslangan mashqlarni faol qo‘llash muhim ahamiyatga ega. Ushbu vositalar nafaqat tilning grammatik va leksik komponentlarini, balki uning pragmatik, madaniy va emosional jihatlarini ham chuqur o‘rganish imkonini yaratadi. Bu esa talabalarga tilni hayotda va turli muloqot vaziyatlarida to‘g‘ri va samarali qo‘llashga yordam beradi. Tavsiyalar: Til o‘rgatishda madaniy kontekstni qamrab olgan interaktiv resurslar ishlab chiqish, bu orqali talabalarning nafaqat tilni, balki uning madaniy va ijtimoiy kontekstlarini ham o‘zlashtirishiga erishish mumkin; Sun’iy intellekt yordamida avtomatik pragmatik tahlil platformalarini yaratish, bu talabalarning tilni pragmatik nuqtai nazardan analiz qilish va uning turli kontekstlarda qo‘llanishini yaxshilashga yordam beradi; Talabalarning nutqiy faoliyatini baholashda kompyuter texnologiyalarini joriy etish, bu hamdagi mustaqil ishlashga imkon yaratadi va ularning to‘g‘ri va samarali muloqot olib borish qobiliyatlarini oshiradi. Shu bilan birga, pragmatik kompetensiyaning rivojlanishi mukammal til o‘rgatish tizimida muhim o‘rin tutadi. Zamonaviy ta’lim vositalari orqali bu maqsadlarga erishish uchun yanada ko‘proq tadqiqotlar va innovasion yo‘nalishlar talab etiladi.

Foydalanilgan adabiyotlar:

1. Bardovi-Harlig, K. (2013). Developing L2 Pragmatics. Language Learning, 63(S1), 68–86.
2. Baxmann, L. F. (1990). Fundamental Considerations in Language Testing. Oxford University Press.
3. Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics.

4. Chang, R., & Li, H. (2019). Multimedia resources and interactive teaching methods in foreign language learning. International Journal of Educational Technology, 29(2), 99–110.
5. Chapelle, C. A. (2001). Computer applications in second language acquisition. Cambridge University Press.
6. Council of Europe. (2020). CEFR Companion Volume with New Descriptors. Strasbourg: Council of Europe Publishing.
7. Coursera Impact Report. (2022). Retrieved from: https://about.coursera.org/impact
8. Crystal, D. (2006). Language and the Internet. Cambridge University Press.
9. Duolingo Language Report. (2023). Retrieved from: https://blog.duolingo.com
10. Kelly, P. (2021). Automated pragmatic analysis in language education: A new horizon. Journal of Educational Technology, 58(1), 120–134.
11. Kramsch, C. (1998). Language and Culture. Oxford University Press.
12. Li, T. (2022). Pragmatic barriers in online language learning: A systematic review. Journal of Language and Communication Studies, 35(3), 220–238.
13. Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic skills. Foreign Language Annals, 53(2), 209–233.
14. President qarori: Yoshlar va aholini zamonaviy kasb-hunarlarga o‘qitish dasturi. (2022).
15. Smith, J. & Brown, A. (2020). Pragmatic competence and technology-enhanced language learning. Language Teaching Research, 24(4), 437-455.
16. Taguchi, N. (2015). Instructed Pragmatics at a Glance: Where Instruction Has Been and Where It Should Go. Language Teaching, 48(1), 1–50.
17. UNESCO Institute for Statistics. (2023). Global Education Monitoring Report.
18. World Bank. (2022). Remote Learning During COVID-19: Lessons from Today, Principles for Tomorrow. Retrieved from: https://www.worldbank.org
19. Олий таълим, фан ва инновациялар вазирлиги ҳисоботи. (2023). Retrieved from: https://edu.uz

I SHO‘BA. O‘ZBEK TILINING FORMAL GRAMMATIKASI YARATISH

OT SO‘Z TURKUMIGA OID BIRLIKLARNING LEKSIK-GRAMMATIK XUSUSIYATI

Elov Botir Boltayevich,
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Bekmurzayeva Maftuna Baxtiyor qizi,
Kompyuter lingvistikasi mutaxasisligi magistranti
baratovamaftuna09@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tilidagi ot so‘z turkumi, uning leksik-grammatik xususiyatlari, morfologik shakllanishi va sintaktik funksiyalari tahlil qilinadi. Maqolada otlarning turdosh va atoqli bo‘lishi, ularning semantik guruhlari, shuningdek, otlarni klassifikatsiya qilish usullari haqida keng ma’lumotlar keltirilgan. Kompyuter lingvistikasi nuqtayi nazaridan, otlarni avtomatik teglash va ularning semantik va morfologik tahlili ham ko‘rib chiqiladi. Otlarning tilshunoslikdagi o‘rni va ularning zamonaviy tadqiqotlarda qanday yangi yondashuvlar bilan o‘rganilishi, jumladan, affiksatsiya, kompozitsiya va abbreviatsiya usullari orqali yasalishi tahlil qilinadi.Tabiiy tilni qayta ishlashda (NLP) ot so‘z turkumini morfologik teglash va ular uchun lingvistik baza yaratish muhim masala hisoblanadi. Otlarning semantik tasnifi, grammatik kategoriyasi, morfologik xususiyatlari morfologik teg uchun muhim hisoblanadi.

Abstract. This article analyzes the noun part of speech in the Uzbek language, its lexical-grammatical features, morphological formation, and syntactic functions. The article provides detailed information about the distinction between animate and inanimate nouns, their semantic groups, as well as methods for classifying nouns. From the perspective of computational linguistics, the automatic tagging of nouns and their semantic and morphological analysis are discussed. The article also analyzes the role of nouns in linguistics and the new approaches in their study in contemporary research, including the formation of nouns through affixation, composition, and abbreviation. In natural language processing (NLP), an important issue is the morphological tagging of nouns and the creation of a linguistic database for them. The semantic classification of nouns, their grammatical category, and morphological characteristics are crucial for the morphological tag.

Аннотация. В данной статье анализируется существительное как часть речи в узбекском языке, его лексико-грамматические особенности, морфологическое формирование и синтаксические функции. В статье приведены подробные данные о различиях между одушевленными и неодушевленными существительными, их семантические группы, а также методы классификации существительных. С точки зрения компьютерной лингвистики рассматриваются автоматическая тегировка существительных и их семантический и морфологический анализ. Также анализируется роль существительных в языкознании и новые подходы в их изучении в современных исследованиях, включая образование существительных с помощью аффиксации, композиции и аббревиации. В обработке естественного языка (NLP) важным вопросом является морфологическая тегировка существительных и создание лингвистической базы данных для них. Семантическая классификация существительных, их грамматическая категория и морфологические характеристики являются важными для морфологического тега.

Kalit so‘zlar: *ot so‘z turkumi, semantik tasnif, grammatik kategoriya, morfologik xususiyat, teg.*

Kirish

Ot so‘z turkumi tilning eng asosiy leksik-grammatik birliklaridan biri bo‘lib, har qanday tabiiy tilda muhim sintaktik va semantik ahamiyat kasb etadi. O‘zbek tilida otlar shaxs, narsa, hodisa ifodalovchi so‘zlar guruhini tashkil etib, ularning morfologik shakllanishi va sintaktik rolini belgilovchi turli grammatik omillar mavjud.

Adabiyotlar tahlili

Tilshunoslikda ot so‘z turkumining leksik-grammatik xususiyatlari yuzasidan turli tadqiqotlar amalga oshirilgan. A. M. Mahmudov (2000), Sh. R. Rahmatullayev (1982), N. G‘. Mallaev (1976) kabi olimlar tomonidan otlarning grammatik kategoriyalari, sintaktik vazifalari va semantik tasnifi chuqur o‘rganilgan. Shu bilan birga, zamonaviy lingvistika doirasida otlarning grammatik strukturasi, ularning morfologik va sintaktik xususiyatlari bo‘yicha ham tadqiqotlar davom etmoqda.

Kompyuter lingvistikasi nuqtayi nazaridan otlarning avtomatik teglanishi, ularning semantik va morfologik jihatdan tahlil qilinishi sohasida olib borilgan tadqiqotlar [1] ushbu sohadagi yangi yondashuvlarni ochib beradi. Ushbu maqolada an’anaviy va zamonaviy tadqiqotlar asosida otlarning lingvistik tavsifi keng tahlil qilinadi.

Asosiy qism

Predmetlik ma’nosini anglatadigan so‘z turkumi *ot* deyiladi. Ot bu ma’noni son, egalik va kelishik grammatik kategoriyalari ifodalaydigan ma’no bilan birgalikda anglatadi. [2]

Ot borliqdagi bir turdagi predmetdan birining yoki shu turdagi predmetning umumiy nomini bildirishiga kora *atoqli va turdosh* otga bo ‘linadi.

Atoqli otlar atab qo‘yilgan yakka shaxsning nomi, borliqdagi yakka narsa, voqea-hodisalarning nomi hisoblanadi va ***[NP]*** (***Proper noun)*** shaklida teglanadi. Atoqli ot quyidagi ma’noviy guruhlarga ega:

Atoqli otlarning ma’noviy guruhlari: *[Qo‘lda teglanadi]*

1	Shaxs nomi	(Otabek)	**[[NPTitl]]**
2	Geografik nom	(O‘zbekiston)	**[[NPGeo]]**
3	Tashkilot, muassasa, korxona nomi	(Xalq ta’limi vazirligi)	**[[NPOrg]]**
4	Samoviy yoritgichlar nomi	(Zuhro, Yupiter)	**[[NPCos]]**
5	Tarixiy sana yoki bayram nomi	(Mustaqillik kuni)	**[[NPDt]]**
6	Hayvon nomi	(To‘rtko‘z, Arslon)	**[[NPAnm]]**
7	Maxsulot nomi	(Siltana – ovqat)	**[[NPPrd]]**
8	Ilohiy tushuncha nomi	(Iso, Jannat)	**[[NPDvn]]**

Turdosh otlar ot so‘z turkumidagi so‘zlarning asosiy qismini tashkil qiladi. Turdosh otlar bir jinsdagi shaxs, narsa, hayvonlar, voqea-hodisalarning nomlari bo‘lib ular haqidagi umumiy predmetlik tushunchasini anglatadi va ***[[NC]]*** (***Noun Cognate***) kabi teglanadi.

Turdosh otlarning leksik-semantik ma’no turlari *[Qo‘lda teglanadi]*

1	modda nomi	(oltin, kumush)	**[[NCM]]**
2	shaxs tavsifi nomi	(chol, yigit)	**[[NC]]**
3	narsa/buyum nomi (qalam, kitob)	(qalam, kitob)	**[[NCPrd]]**
4	o‘simlik va uning qismlari nomi	(daraxt, barg)	**[[NCPlt]]**
5	miqdor nomi	(misqol, tonna)	**[[NCNU]]**
6	tashkilot va muassasa nomi	(maktab, idora)	**[[NCOrg]]**
7	o‘rin-joy nomi	(qishloq,shahar)	**[[NCOrg]]**
8	vaqt-payt nomi	(kun, fasl, tong)	**[[NCWhl]]**
9	inson tana a’zolari nomlari	(soch, yurak)	**[[NCBdyPrt]]**
10	hayvon va uning a’zolari nomi	(kuchuk, ot, tuyoq)	**[[NCAn]]**
11	oziq-ovqat nomi	(non, suv)	**[[NCFd]]**
12	diniy-ilohiy tushuncha nomi	(nomoz, rukn)	**[[NCDvn]]**
13	faoliyat/jarayon nomi	(uchrashuv, tortishuv)	**[[NCEv]]**

Bevosita kuzatishga ko‘ra otlar ikkiga bo‘linadi:

1	aniq otlar	**[[NCnc]]**
2	mavhum otlar	**[[NAbs]]**

Aniq otlar konkret shaxs yoki narsani bildiradi. Aniq otlar bevosita ko‘rish va sanash mumkin bo‘lgan narsalarni anglatgani uchun sanoq son bilan ifodalangan sifatlovchi oladi, shuningdek, birlik va ko‘plik shakllarida ham qo‘llanadi: *ruchka – uchta ruchka – ruchkalar, qiz – beshta qiz – qizlar, shahar – yigirmata shahar – shaharlar* kabi. Aniq otlarni teglash jarayoni avval qo‘lda amalga oshiriladi so‘ng avtomatlashtiriladi.

Mavhum otlar mavhum tushunchalarni, belgi, xususiyat, harakat va holatni narsa sifatida bildiradi. Bunday otlar (Mavhum otlar) ko‘plikda kelmaydi: *sevgi, g‘am, baxt, alam, muhabbat* kabi. Mavhum otlarni teglash jarayoni avval qo‘lda amalga oshiriladi so‘ng avtomatlashtiriladi. [2]

Tuzilishiga ko‘ra*[Qo‘lda teglanadi]*

1. Sodda ot [[ND]]

Sodda otlar faqat bir asosdan tarkib topgan otlardir. Sodda otlar tub va yasama bo‘ladi *(daftar, ishlarimiz, o‘roq).*

2. Qo‘shma ot [[NCmp]]

Qo‘shma otlar ikki yoki undan ortiq asosdan tarkib topgan otlardir *(kungaboqar, beshbarmoq, qolqop).* Qo‘shma otlar qoshib yoki ajratib yoziladi. Ikkinchi qismi turdosh ot bo‘lgan atoqli otlar qo‘shib yoziladi *(Oltariq, Yangibozor).* Ikkinchi qismi atoqli ot bo‘lgan qo‘shma otlat ajratib yoziladi *(O‘rta Osiyo, O‘rta Chirchiq).*

3. Juft ot [[NCpl]]

Juft otlar ikkita otning juftlashuvidan hosil bo‘ladi va chiziqcha bilan yoziladi *(kitob-daftar, ota-ona, orzu-armon, gap-so‘z).*

4. Takroriy ot [[NRep]]

Takroriy otlar otni aynan yoki tovush o‘zgarishi bilan takrorlash orqali hosil qilinadi. Takroriy otlar chiziqcha bilan ajrtib yoziladi *(savat-savat, dasta-dast, uy-puy).*

Otning funksional shakllari mavjud bo‘lib, ular ma’lum bir qo‘shimcha grammatik ma’no ifodalash, otlarning kategorial shakllariga xos bo‘lmagan biror sintaktik vazifasiga moslashish uchun xizmat qiladi. Otga xos funksional shaklning yasalishga ko‘ra 3 tipi mavjud:

1) sintetik shakl;

2) analitik shakl;
3) juft va takror shakl.

Otlarning yasalishi natijasida yasama otlar paydo bo‘ladi. Yasama otlar affiksatsiya, kompozitsiya, abbreviatsiya yo‘llari bilan yasaladi.

Yasalishiga ko‘ra

4.1. Tub [[ND0]]
4.2. Yasama [[ND1]]
2.1. affiksatsiya [[ND1Af]]
So ‘z yasovchi qo‘shimcha qo‘shilishi orqali yasaladi *(buyruq, o‘roq)*.
2.2. kompozitsiya [[ND1Cmp]]
Ikkita mustaqil so‘z qo ‘shilib yangi ma’noli so‘z hosil bo‘ladi, kompazitsiya usuli orqali qo‘shma va juft ot hosil qilinadi *(tuyaqush, oybolta)*.
2.3. abbreviatsiya [[ND1abv]]
Abbreviatsiya usuli bilan ot yasalganda, qisqartma otlar hosil qilinadi (BMT, ToshDO‘TAU, SamDU, O‘zMU).

Predmetning o‘zini yoki to‘dasini bildirishiga ko‘ra*[Qo‘lda teglanadi]*

3.1. Jamlovchi ot **[[NN]]**
Jamlovchi otlar bir jinsdagi narsalarning yig‘indisini bir butun holatda anglatadigan turdosh ot. *(Xalq, armiya, meva)*

3.2. Yakka ot **[[NN1]]**
Yakka otlar to ‘da sostaviga kirgan alohida predmetlarni bildiradi. Bunday otlarning birlik shaklida ham ko‘plik shaklida ham narsalar yakka-yakka holatda tasavvur qilinadi. *(dehqon-dehqonlar, kishi-kishilar, odam-odamlar)*

Xulosa. O‘zbek tilida ot so‘z turkumi tilning eng asosiy leksik-grammatik birliklaridan biri sifatida, uning semantik va grammatik tahlili tilshunoslikda katta ahamiyatga ega. Otlar, nafaqat leksik ma’nolarni ifodalashda, balki sintaktik tuzilmalarda ham muhim rol o‘ynaydi. O‘zbek tilida otlarning tasnifi, ularning grammatik kategoriyalari va semantik guruhlari bo‘yicha ilgari surilgan tadqiqotlar o‘zbek tili grammatikasining murakkabligini va boyligini ko‘rsatadi. Bundan tashqari, zamonaviy kompyuter lingvistikasi tadqiqotlari otlarning avtomatik tahlili, ularning morfologik va semantik jihatlarini aniqlashda yangi imkoniyatlarni taqdim etadi. Bu esa nafaqat tilshunoslarga, balki tilni ishlatish va o‘rganish jarayonlarini avtomatlashtirishga intilayotgan texnologik sohalarga ham katta foyda keltiradi.

Otlarning morfologik shakllanishi, sintaktik roli va semantik xususiyatlari o‘zbek tilining tilshunoslikda o‘rganilishi kerak bo‘lgan muhim sohalaridan biridir. Kelajakda bu sohada olib borilgan tadqiqotlar va yangi metodlar tilni yanada mukammal va chuqur tahlil qilish imkonini yaratadi. O‘zbek tilida ot so‘z turkumi leksik-grammatik jihatdan keng ko‘lamli tadqiqot talab qiluvchi murakkab tizimni tashkil etadi. Semantik va grammatik xususiyatlar nuqtayi nazaridan otlarning tasnifi, ularning morfologik shakllanishi va sintaktik funksiyalari tilshunoslikda muhim o‘rin tutadi. Zamonaviy lingvistika va kompyuter lingvistikasi kontekstida otlarning avtomatik tahlil qilinishi ularning yanada chuqur o‘rganilishiga zamin yaratadi.

Foydalanilgan adabiyotlar:

1. Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. Stanford University Press.
2. Tursunov U., Muxtorov A., Rahmatullayev Sh. Hozirgi o‘zbek adabiy tili. Toshkent “O‘zbekiston” 1992
3. Hojiyev, A. (2019). O‘zbek tili morfologiyasi. Toshkent: Fan.
4. Hamroyev, M. (2020). O‘zbek tilida so‘z turkumlari tasnifi. O‘zbek tili va adabiyoti, 3, 45-52.
5. Mengliyev, B. (2021). O‘zbek tili morfologik analizatori. Kompyuter lingvistikasi va sun’iy intellekt, 2(1), 12-24.
6. Elov B., Hamroyeva Sh., Abdullayeva O., Uzakova M. O‘zbek tilida POS tegging masalasi: muammo va takliflar / Uzbekistan: Language and Culture. Applied philology. 2022/2(5). – 51-68-b.
7. Elov B., Hamroyeva Sh., Xudayberganov N., Yodgorov U., Yuldashev A. Pos tagging of Uzbek texts using hidden Markov models (HMM) and Viterbi algorithm. O‘zMU xabarlari. Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti ilmiy jurnali. 2023 yil Maxsus son
8. Sokirko, A. V. (2021). Morphological analysis algorithms for Turkic languages. Computational Linguistics and Intellectual Technologies, 15(2), 156-169.
9. https://cyberleninka.ru https://github.com/Step2AI/Uzbek-tilidagi-Suniy-Intellektmanbalari.
10. https://github.com/Step2AI/Uzbek-tilidagi-Suniy-Intellektmanbalari.

UDK: 81.33

KAM RESURSLI TILLARDA G2P MODEL TAYYORLASHDA FONETIK KORPUSLARNING AHAMIYATI: O‘ZBEK TILI MISOLIDA

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Maxmudjonova Gulshaxnoz Ulug‘bek qizi,
tayanch doktorant
gulshaxnozmahmudjonova@gmail.com
ToshDO‘TAU

Annotatsiya: Ushbu maqolada kam resursli tillar, xususan o‘zbek tili uchun Grafema-fonema (G2P) modelini yaratishda fonetik korpuslarning o‘rni va ahamiyati yoritilgan. G2P konversiyasi – matnni fonemalarga aylantirish jarayoni bo‘lib, zamonaviy nutq sintezi (TTS) tizimlarining asosiy komponentlaridan biridir. O‘zbek tili fonetik tizimining murakkabligi, undagi tarixiy, fonologik va morfonologik xususiyatlar, shuningdek, dialektal tafovutlar fonetik korpuslar asosida formallashtirilmaguncha, modelning sifatli ishlab chiqilishi mushkul bo‘ladi.

Abstract: This article explores the significance and role of phonetic corpora in developing Grapheme-to-Phoneme (G2P) models for low-resource languages, with a particular focus on the Uzbek language. G2P conversion – the process of mapping text to phonemes – is a core component of modern Text-to-Speech (TTS) systems. Due to the complexity of Uzbek phonetics, including its historical, phonological, and morphonological features, as well as dialectal variation, a high-quality G2P model cannot be effectively developed without the formalization of phonetic corpora.

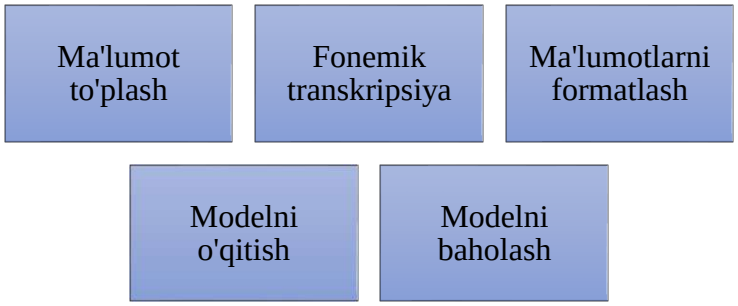
Аннотация: В данной статье рассматриваются значение и роль фонетических корпусов при создании моделей преобразования графем в фонемы (G2P) для языков с ограниченными ресурсами, в частности, узбекского языка. Преобразование G2P – это процесс трансформации текста в фонемы, который является ключевым компонентом современных систем синтеза речи (TTS). Сложность фонетической системы узбекского языка, включая её исторические, фонологические и морфофонологические особенности, а также диалектные различия, делает невозможным создание качественной G2P-модели без формализации фонетического корпуса.

Kalit so‘zlar: *fonetik korpus, grafema, transkripsiya, alifbo, lug‘at, ma’lumotlar bazasi*

G2P (grafemadan fonemaga o‘tkazish) modeli matndagi so‘zlarni ularning talaffuziga avtomatik konvertatsiya qilish vazifasini bajaradi. Bu texnologiya nutqni sintezlash (Text-to-Speech, TTS) va nutqni avtomatik tanish (Automatic Speech Recognition, ASR) tizimlarida muhim rol o‘ynaydi[1]. Masalan, TTS jarayonida tovushlarning to‘g‘ri chiqarilishini ta’minlash uchun matnni fonetik transkripsiyaga aylantirish zarur bo‘ladi – ayniqsa yangi yoki lug‘atda yo‘q so‘zlarda (Out-of-vocabulary, OOV)[2]. Yetakchi xalqaro tillar (ingliz, fransuz, rus) uchun bunday modellar mavjud bo‘lib, ularni tayyorlashda keng qamrovli talaffuz lug‘atlari va fonetik korpuslardan foydalaniladi[3]. Biroq, kam resursli tillarda – jumladan, o‘zbek tilida – bunday fonetik ma’lumotlar bazasi yo’qligi G2P modelini yaratishni qiyinlashtiradi.

O‘zbek tili dunyoda o‘n million insonlar tomonidan so‘zlashiladi (turli baholashlarga ko‘ra 30-40 million atrofida[4] va turkiy tillar orasida ikkinchi eng ko‘p so‘zlovchiga ega tildir. Shunga qaramay, hozirgi paytda o‘zbek tilida talaffuz lug‘atlari yoki fonetik belgilangan matn korpuslari deyarli mavjud emas. Natijada, G2P modelining aniqligi uchun zarur bo‘lgan orfografiya – tovush mosliklarini mashinaga o‘rgatish murakkablashadi. Masalan, bir necha tillar uchun yaratilgan ko‘p tilli G2P tizimlari[5] o‘zbek tilini ham qamrab olgan bo‘lsa-da, ularning resurslari to‘liq ishonchli emas: CharsiuG2P loyihasi o‘zbek tili bo‘yicha ishlatilgan talaffuz lug‘atida xatoliklar borligini va uni yangilash zarurligini qayd etgan[6]. Bu holat kam resursli tilga xos bo‘lgan muammo – *fonetik ma’lumotlarning sifati va standartlashmagani* – modelga bevosita ta’sir qilishini ko‘rsatadi.

Fonetik korpus tuzish bosqichlari (1-rasmga qarang). O‘zbek tili uchun G2P modelini yaratishda birinchi navbatda fonetik korpus – ya’ni so‘zlar va ularning to‘g‘ri fonemik transkripsiyasidan iborat ma’lumotlar to‘plami – shakllantiriladi. Korpus tuzish jarayoni quyidagi bosqichlarni o‘z ichiga oladi:



1-rasm. Fonetik korpus tuzish bosqichlari

Ma’lumot to‘plash: avvalo, o‘zbek tilida reprezentativ matnlar to‘plami yig‘iladi. Bu jarayonda adabiy va og‘zaki nutqni qamrab oluvchi turli janr va uslubdagi matnlar (rasmiy hujjatlar, yangiliklar, badiiy asarlar, og‘zaki nutq namunalari) tanlanadi. Maqsad – korpusda o‘zbek tilidagi barcha asosiy tovush

birikmalari va holatlarini qamrab olishdir. Korpusga kiritiladigan so‘zlar soni imkon qadar ko‘proq va xilma-xil bo‘lishi lozim (masalan, 50 – 100 ming atrofida so‘z shakllari). Yig‘ilgan xom matnlar tozalash va normalizatsiya qilinadi: keraksiz belgi va shum (masalan, raqamlar, qisqartmalar) chiqarib tashlanadi, yoki standart ko‘rinishga keltiriladi. So‘zlar alohida leksik birliklar sifatida ajratiladi va korpusning leksik ro‘yxati shakllantiriladi.

Fonemik transkripsiya va annotatsiya: keyingi bosqichda korpusdagi har bir so‘zning fonemik transkripsiyasi tayyorlanadi. Buning uchun transkripsiya standarti tanlanishi zarur. Ilmiy amaliyotda xalqaro IPA (International Phonetic Alphabet) keng qo‘llaniladi, chunki u barcha tovushlarni unifikatsiyalangan belgilarda ifodalash imkonini beradi. O‘zbek fonetikasi uchun moslashtirilgan alifbo ham qo‘llaniladi. Transkripsiyani lingvist-mutaxassislar qo‘lda bajarishi yoki maxsus yarim-avtomatik vositalar yordamida amalga oshirishi mumkin. Hozirgi kunda mavjud ba’zi dasturlar (Epitran, CharsiuG2P, MFA) boshqa tillardagi modellarga asoslanib o‘zbekcha taxminiy transkripsiya berishi mumkin, lekin ularni ehtiyotkorlik bilan qo‘llash lozim – yuqorida ta’kidlanganidek, ko‘p tilli model bergan transkripsiya xatolarini inson tomonidan tuzatish talab etiladi. Demak, annotatsiya jarayoni sifat jihatdan nazoratlangan bo‘lishi kerak: har bir so‘zning transkripsiyasi kamida ikki mutaxassis tomonidan tekshirilishi, bir xil transkripsiya me’yoriga rioya etilishi zarur. Annotatsiyada fonemalar izchil belgilanishi (bir xil tovush har doim bir xil belgi bilan ifodalanishi shart), urg‘u yoki intonatsiya kabi qo‘shimcha prosodik belgilar zarur bo‘lsa, alohida belgi bilan ko‘rsatilishi mumkin.

Ma’lumotlarni formatlash va bazaga kiritish: transkripsiyalangan korpus ma’lumotlari lug‘at shaklida tartiblanadi, har bir qatorda so‘z va uning fonetik transkripsiyasi yozilgan format (masalan, JSON ko‘rinishda) saqlanadi. Bu lug‘at G2P modelini o‘qitish uchun tayyor ma’lumotlar bazasi vazifasini bajaradi. Korpus tuzishda, shuningdek, morfologik belgilarni ham saqlab qo‘yish foydali bo‘ladi, chunki o‘zbek tili agglyutinativ til bo‘lib, morfemalar chegarasida tovush o‘zgarishlari kuzatilishi mumkin. Korpusning har bir yozuvi ko‘rinishi jihatdan: *“so‘z – [fonemik transkripsiya]”* tarzida bo‘ladi. Misol: kitob — [k i t ɒ b], quyosh — [q u j ɒ ʃ] kabi.

Modelni o‘qitish (tayyorlash): tayyor fonetik korpus asosida G2P modeli o‘qitiladi. G2P modellarini ishlab chiqishda ikki asosiy yondashuv mavjud: lug‘atga asoslangan va qoidaga asoslangan usullar. Lug‘atga asoslangan usulda katta hajmdagi tayyor talaffuz lug‘ati modelga yuklanadi va undan statistik yoki neyron tarmoq usullari bilan yangi so‘zlar uchun talaffuz generatsiya qilinadi. Qoidaga asoslangan yondashuvda esa tilshunoslikdagi fonologik qoidalar dasturlashtiriladi va faqat istisno so‘zlar uchun alohida lug‘at yuritiladi. Hozirgi zamon G2P modellari ushbu yondashuvlarning kombinatsiyasidan foydalanadi. Masalan, Phonetisaurus[7] kabi vositalar WFST asosida lug‘atni o‘rgatib, keyin yangi so‘zlarni statistik

qoidalar yordamida transkripsiya qiladi[8]. NVIDIA NeMo yoki CharsiuG2P kabi platformalar esa neyron tarmoqqa asoslangan yondashuvdan foydalanadi (masalan, Transformer arxitekturasi yoki ByT5 modeli) va bir nechta til uchun birgalikda o‘qitilib, keyin ma’lum bir tilga moslashtiriladi. Bizning holatda, korpus nisbatan kichik va tilimiz xususiyatlari o‘ziga xos bo‘lgani uchun, qoida-lug‘atli model konsepsiyasini qo‘llash maqsadga muvofiq: ya’ni, o‘zbek tilining fonetik qonuniyatlarini inobatga olgan holda, kichik lug‘at va qoidalar integratsiyasi bilan ishlaydigan modelni tanlash. Bu jarayonda Phonetisaurus yoki MFA[9] singari vositalar yordamida avval grafema-fonema moslashtirish (alignment) amalga oshirilib, so‘ng n-gram model yoki neyron tarmoqni o‘qitish orqali grafema ketma-ketliklaridan fonema ketma-ketliklariga xarita hosil qilinadi. Modelni tayyorlashda korpusning 80-90% qismi o‘qitish uchun, qolgan qismi test va tasdiqlash (validation) uchun ajratiladi. Natijada, model yangi berilgan so‘zning imlosidan uning talaffuzini bashorat qilishni o‘rganadi.

Modelni baholash: G2P model sifatini o‘lchash uchun fonemik xato darajasi (Phoneme Error Rate, PER) yoki so‘z bo‘yicha aniqlik kabi mezonlar qo‘llanadi. Tayyor model korpusdan ajratib olingan test to‘plamidagi so‘zlar uchun transkripsiya yaratadi va ular mutaxassis tayyorlagan etalon transkripsiya bilan taqqoslanadi. Kam resursli til sharoitida modelning dastlabki natijalari kutilganidek pastroq bo‘lishi mumkin; bu holda modelni takomillashtirish uchun korpusni kengaytirish (yangi so‘zlar qo‘shish), yoki model arxitekturasini murakkablashtirish kabi choralar ko‘riladi. Shu tariqa, metodologiya bosqichlari takomillashtirib borilishi ham mumkin – masalan, model chiqishlaridagi tizimli xatolar tahlil qilinib, ularni tuzatish uchun korpus annotatsiyasiga qo‘shimcha qoidalar yoki misollar kiritiladi.

Fonemik transkripsiya standarti va annotatsiya prinsiplari. Fonetik korpusni annotatsiya qilishda fonemik transkripsiya standartining to‘g‘ri tanlanishi va izchil qo‘llanishi hal qiluvchi ahamiyatga ega. O‘zbek tilining tovush tizimi uchun alifbo IPA standartiga moslashtirilishi kerak. Transkripsiyada unli fonemalar uchun IPA belgilarini qo‘llaganda, ularning allofonik variantlari ham e’tiborga olinadi. O‘zbek tilida ba’zi kombinator hodisalar yuz beradi, ular transkripsiya me’yorlarida aks ettirilishi lozim. Masalan, jarangli/jarangsizlanish assimilatsiyasi: so‘z ichida yoki qo‘shimcha qo‘shilganda, ba’zan yonma-yon undoshlar bir-biriga ta’sir qiladi. Transkripsiya jarayonida barcha qoidalar ishlab chiqilib, bir xil uslubda bajarilishi kerak. Bu degani, biror tovushni ifodalash uchun qanday belgi tanlansa, hamma joyda shuni qo‘llash lozim; birikma yoki qisqartmalarning talaffuzi bo‘yicha formallashtirilib, qat’iy rioya qilinadi. Masalan, “SMS” so‘zi harflab o‘qiladi (es-em-es) yoki “YUNESKO” qisqartmasi [Yunesko] deb o‘qiladi – buni oldindan hal qilib, korpusga to‘g‘ri kiritish kerak. Annotatsiya prinsiplari bo‘yicha qo‘llanma tuzilib, ularda turli murakkab holatlar uchun yechimlar ko‘rsatiladi. Bu prinsipial

jihatlar model sifatiga katta ta’sir qiladi – agar bir xil so‘z turlicha transkripsiya qilinsa, model buni shovqin sifatida qabul qilib, noto‘g‘ri o‘rganadi.

Fonetik korpus sifati va model natijalari. G2P modelining muvaffaqiyati ko‘p jihatdan tayyorlangan fonetik korpusning sifatiga bog‘liq. Agar korpusda xatolar yoki nomuvofiqliklar bo‘lsa, model noto‘g‘ri tarzda o‘rganadi va bu talaffuz xatolariga olib keladi. Shu bois, korpus annotatsiyasida qo‘yilgan qoidalarga qat’iy amal qilish va uni imkon qadar to‘liq va xatosiz qilish muhim. Misol uchun, o‘zbek tilida “o” va “o‘” fonemalarini adashtirish modelni chalg‘itishi va natijada “ol” va “o‘l” so‘zlarini farqlay olmaslikka olib kelishi mumkin. Shunday holatlarda model chiqishini tekshirib, korpusdagi potentsial xatoni tuzatish lozim.

Annotatsiya muammolarining yana biri – inson omili. Fonetik transkripsiya qo‘lda qilinganida, ayniqsa katta hajmda, ba’zan subyektiv farqlar paydo bo‘lishi mumkin. Turli lingvistlar bir tovushni turlicha talqin qilishlari yoki diakritik belgilar qo‘llashdagi uslubi farq qilishi mumkin.

Fonetik korpus hajmi ham ahamiyatli. Dastlabki bosqichda korpus kichik bo‘lsa, model yetarlicha qoidalarni o‘rganmasligi mumkin. Tadqiqotlar shuni ko‘rsatadiki, Transformer kabi zamonaviy modellar katta ma’lumot talab qiladi va bitta tilning ma’lumotlari kam bo‘lsa, ko‘p tilli o‘qitish orqali samaradorlikni oshirish mumkin. Bizning kontekstimizda ham agar o‘zbek tilida 50 ming so‘zli korpus yetarli natija bermasa, uni turkiy tillar (masalan, qozoq, turk, uyg‘ur) ma’lumotlari bilan birga o‘qitib, keyin o‘zbekchaga moslashtirish (fine-tuning) strategiyasini qo‘llash mumkin. SIGMORPHON kabi tanlovlarda ham kam resursli tillar uchun transfer learning yondashuvi samarali ekani ko‘rsatilgan[10]. Biroq, bunday ko‘p tilli modelda har bir tilning o‘ziga xos fonetikasi “yo‘q bo‘lib ketmasligi” uchun ehtiyot bo‘lish zarur – masalan, turk tilidagi unlilar uyg‘unlashuvi qoidalari modelga singib, o‘zbek tiliga noto‘g‘ri tatbiq bo‘lmasligi kerak. Shu sababdan biz taklif qilayotgan yondashuv fonetik jihatdan asoslanadi: ya’ni, modelga boshidan o‘zbek tilining fonologik xususiyatlari (masalan, unlilar uyg‘unlashuvi yo‘qligi, qator va tor unlilar farqi, g‘/q tovushlari mavjudligi kabi) haqida ma’lumot beriladi yoki qoidalar shaklida kiritiladi. Bu bilimlar modelning ichki qatlamlarida yoki qoidalovchi modulida hisobga olinsa, u holda korpus kichikroq bo‘lsa ham, model ma’lum qoidalarga tayangan holda to‘g‘riroq ishlaydi. Masalan, yapon tilida kam resurs sharoitida mutaxassislar aynan qoidabazali G2P tizimini taklif qilishgan – lingvistik adabiyotga asoslanib, maxsus qoidalar to‘plami tuzilgan va u ko‘p holatlarda neyron modelga teng natija bergan[11]. O‘zbek tilida ham xuddi shunday, tilshunoslikka asoslangan yondashuvni joriy etish mumkin: bu modelga fonetik korpusdan tashqari fonologik qoidalarni qo‘lda kiritishni anglatadi.

Fonetik korpusning roli. Kam resursli til – xususan, o‘zbek tili – uchun G2P modelini muvaffaqiyatli yaratishda fonetik korpus hal qiluvchi ahamiyatga ega.

Talaffuzi aniq belgilangan so‘zlar jamlanmasi modelni to‘g‘ri o‘qitish va tekshirish imkonini beradi. Korpus hajmi va sifati qanchalik yuqori bo‘lsa, model natijalari shunchalik ishonchli chiqadi. Korpus yaratishda barcha asosiy tovush birikmalari va fonologik hodisalarni qamrab olish lozim, aks holda model notanish holatlarda xatoga yo‘l qo‘yishi mumkin[12].

Yakunda, o‘zbek tili misolida ko‘rsatildiki, kam resursli til uchun G2P model yaratish tilshunos va sun’iy intellekt mutaxassislarining yaqin hamkorligini talab etadi. Fonetik korpus shakllantirish va uni tahlil qilish orqali biz modelning birinchi talqinini yaratishimiz mumkin. Keyingi amaliy ishlar bosqichida mazkur nazariy ishlab chiqilgan yondashuv asosida G2P modelining dastlabki amaliy tadbiqi amalga oshiriladi va uning natijalari baholanadi. Ushbu konseptual ish kelgusida nafaqat o‘zbek tili, balki unga o‘xshash ahvoldagi boshqa kam resursli tillar (masalan, tojik, qirg‘iz, turkman) uchun ham foydali bo‘lishi mumkin – zero, til texnologiyalari sohasida resurslarni yaratish va ulardan samarali foydalanish globallashgan dunyoda kichik tillarning rivoji uchun muhim ahamiyat kasb etadi.

Foydalanilgan adabiyotlar:

1. Irie E., Samson J., Saee Suhaila. A Review on Grapheme-to-Phoneme Modelling Techniques to Transcribe Pronunciation Variants for Under-Resourced Language. Pertanika Journal of Science and Technology. 2023.
2. NVIDIA. NeMo Toolkit Documentation. – URL: https://docs.nvidia.com/nemo-framework/user-guide/24.09/nemotoolkit/tts/g2p.html
3. Kim B. C., Lee G. B., Lee J. H. Hybrid Grapheme-to-Phoneme Conversion for Unlimited Vocabulary. Department of Computer Science & Engineering, Pohang University of Science & Technology, Pohang, Korea. 1999
4. en.wikipedia.org
5. CharsiuG2P – URL: https://github.com/lingjzhu/CharsiuG2P
6. El-Hadi, C., Mhania, G. Phonetisaurus-based letter-to-sound transcription for Standard Arabic. In 2017 5th International Conference on Electrical Engineering - Boumerdes. 2017
7. Phonetisaurus – URL: https://github.com/AdolfVonKleist/Phonetisaurus
8. github.com
9. Montreal Forced Aligner – URL: https://github.com/MontrealCorpusTools/mfa-models
10. Salatas, J. Phonetisaurus: A WFST-driven Phoneticizer – Framework Review. Retrieved from https://www.johnsalatas.com/phonetisaurus-wfst-review, 2012
11. Jiampojamarn S. Grapheme-to-Phoneme Conversion and Its Application to Transliteration. – University of Alberta, 2009.
12. Cheng S., Zhu P., Liu J., Wang Z. A Survey of Grapheme-to-Phoneme Conversion Methods // *Applied Sciences*, 2024.

O‘ZBEK TILIDA MOSLASHUV YO‘LI BILAN BIRIKKAN SO‘Z BIRIKMALARINI MODELLASHTIRISH MASALASI

Abdullayeva Oqila Xolmo‘minovna,
Filologiya fanlari falsafa doktori (PhD)
abdullayevaoqila@gmail.com
ToshDO‘TAU doktoranti DSc

Nizomova Fotima Bo‘rixo‘ja qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
nizomovaf44@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur maqolada o‘zbek tilida moslashuv yo‘li bilan birikkan so‘z birikmalarining nazariy asoslari tahlil qilindi. Moslashuvli so‘z birikmalari ikki yoki undan ortiq so‘zning o‘zaro grammatik va semantik jihatdan bog‘lanishidan iborat bo‘lib, ular tilning morfologik va sintaktik qoidalari asosida hosil qilinishi ta’riflandi. So‘z birikmalarida so‘zlar o‘rtasidagi kelishik, son va egalik kabi grammatik munosabatlar aniq ifodalanishi qayd etildi. Shuningdek, maqolada bunday so‘z birikmalarini modellashtirishning qoidalar asosidagi yondashuvi va grammatik tizimlar orqali amalga oshirilishi tavsiflandi. Moslashuvli so‘z birikmasini modellashtirishning kompyuter lingvistikasi sohasidagi ahamiyati va zarurati haqida fikr bildirildi.

Abstract. In this article, the theoretical foundations of word combinations formed through adaptation in the Uzbek language were analyzed. Adapted word combinations consist of two or more words connected grammatically and semantically, and they are formed based on the morphological and syntactic rules of the language. It was noted that grammatical relationships such as agreement, number, and possession between words in word combinations are clearly expressed. Furthermore, the article describes the rule-based approach to modeling such word combinations and their implementation through grammatical systems. The importance and necessity of modeling adapted word combinations in the field of computational linguistics were also discussed.

Аннотация. В данной статье были проанализированы теоретические основы словосочетаний, образующихся через адаптацию в узбекском языке. Адаптированные словосочетания состоят из двух или более слов, связанных грамматически и семантически, и они образуются на основе морфологических и синтаксических правил языка. Было отмечено, что грамматические отношения, такие как согласование, число и принадлежность, между словами в словосочетаниях выражаются явно. Также в статье описан подход моделирования таких словосочетаний на основе правил и их реализация через

грамматические системы. Также была высказана важность и необходимость моделирования адаптированных словосочетаний в области компьютерной лингвистики.

Kalit so‘zlar: *So‘z birikmasi, moslashuvli so‘z birikmasi, teglash, modellashtirish, sintaktik annotatsiyalash.*

Kirish. Tilshunoslikda so‘z birikmasi tushunchasining mazmuni va ahamiyati ko‘plab olimlar tomonidan turli xil yondashuvlar asosida talqin qilingan. So‘z birikmasi nutqdagi so‘zlarning grammatik va semantik jihatdan o‘zaro bog‘lanib, yaxlit bir tushunchani ifodalashiga asoslanadi. Fortunatov, Peterson, Peshkovskiy kabi jahonda mashhur olimlar so‘z birikmalarining sintaksisdagi asosiy o‘rni va ularning shakllanishiga e’tibor qaratgan bo‘lsalar, Shaxmatov va Vinogradov so‘z birikmasining nominativ funksiyasini va uning gapdan tashqari mavjudlik xususiyatini ta’kidlaydilar.

Nutqda, gapda so‘zlar ma’lum grammatik qoidalar asosida bir-biri bilan bog‘langan bo‘ladi. Masalan, uning qizi (gapirdi), yosh bola (yugurdi), (maqsadim-shu) kitobni o‘qimoq, sekin gapirmoq. Misollardagi har bir juft grammatik butunlikni hosil qilgan bo‘lib, yaxlit, biroq, o‘z ichida qismlarga bo‘linadigan tushunchani, bir butun tushunchani bildiradi. Bu kompleksdagi-qo‘shilmadagi so‘zlarning a’loqasi grammatik jihatdan shakllangandir. Ikki yoki undan ortiq mustaqil so‘zdan tuzilib, bir grammatik butunlik, semantik yaxlitlik tashkil qilgan har bir kompleksi so‘z birikmasi sanaladi[1].

So‘z birikmasi narsa, belgi yoki harakatni yakka holda emas, balki ikki va undan ortiq so‘zlarning o‘zaro aloqaga kirishuvi orqali ifodalaydi, deya ta’rif berish bilangina chegaralanadi [2].

Moslashuv yo‘li bilan birikkan so‘z birikmalari ta’rifi va tavsifi

Moslashuv (tilshunoslikda) - tobe (bog‘liq) aloqaning bir turi bo‘lib, bunda birikmaning tobe qismi o‘z shaklini hokim qismning shakliga moslaydi, shu bilan hokim qismga bog‘lanadi. Hokim qismning shakli o‘zgarishi bilan tobe qism ham ma’no jihatdan shu shaklga mos bo‘lgan shaklni oladi. O‘zbek tilida moslashuv aloqasi, asosan, qaratqich-qaralmish va ega-kesim aloqasida ko‘rinadi va bu moslik shaxs va son jihatidan bo‘ladi, ya’ni qaratqich yoki ega qaysi shaxs-sonni bildirsa, qaralmish yoki kesim xuddi shu shaxs-sonni bildiruvchi shaklda bo‘ladi: *men yozyapman, sen yozyapsan, u yozyapti, biz yozyapmiz, siz yozyapsiz, ular yozishyapti. Mening ukam, sening ukang, uning ukasi.*

Hokim qism (qaratqich yoki ega) 3-shaxsni bildiruvchi otlardan ifodalanganda, hokim va tobe qism sonda moslashishi ham, moslashmasligi ham, ya’ni hokim qism ko‘plik son, tobe qism birlik son shaklida bo‘lishi mumkin: *sozandalar keldilar (kelishdi) yoki sozandalar keldi.* Ega 3-shaxsdagi shaxs bildirmaydigan so‘zlardan bo‘lganida, ega ko‘plik, kesim esa birlik shaklida qo‘llanadi, ya’ni son jihatdan moslashish bo‘lmaydi: gullar ochildi. So‘z birikmalari

tarkibidagi so‘zlarning grammatik bog‘lanishida moslashuvning o‘rni haqida ham olimlar tamonidan alohida fikr mulohazalar bildirilgan[3].

Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo‘linadi: semantik teg, sintaktik teg va grammatik teg. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur.

POS teglash – bu berilgan gapdagi har bir so‘zshaklga uning turkum (ot, fe’l, sifat, son, ravish yoki olmosh)ga mansubligini belgilash (teglash) vazifasidir. POS teglash tabiiy tilni qayta ishlash (Natural Language Processing, NLP)ning asosiy vazifalaridan biri bo‘lib, pipeline konveyerining muhim bosqichi hisoblanadi [4].

Modellashtirish – Kompyuter lingvistikasida tabiiy tilni modellashtirish inson faoliyati davomida axborot almashinuvida foydaniladigan va ushbu faoliyatga asosan doimiy o‘zgarib boradigan ko‘p bosqichli belgilar tizimi bilan bog‘liq jarayondir. Ma’lumotlarni tahlil qilish uchun an’anaviy vositalar, metodologiyalar va ko‘nikmalardan foydalanish uchun berilgan matnni strukturlangan formatga o‘tkazish lozim. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) platformasi moslashuvchan, mustahkam va samarali komponentlar to‘plamini taqdim etadi [5].

1-jadval. Universal va o‘zbek tilidagi POS teglar ro‘yxati

T/r	Teg	O‘zbek tilidagi teg	Ma’nosi o‘zbekcha misollar
1	N	N	Ot *(N)* olma *(apple)*
2	V	VB	Fe’l *(V)* yugurmoq *(run)*
3	ADJ	JJ	SIFAT*(Adjective)* qizil/katta *(red/large)*
4	NUM	NUM	SON *(Numeral)* besh *(five)*
5	ADV	RR	RAVISH *(AdV)* tez *(fast)*
6	PRON	P	OLMOSH *(ProN)* bu *(this)*
7	AUX	II	KO‘MAKCHI *(Auxiliary)* bilan *(with)*
8	CONJ	C	BOG‘LOVCHI *(Conjunction)* va *(and)*
9	PART	Prt	YUKLAMA *(Particle)* faqat *(only)*
10	MOD	MD	MODAL *(Modal)* darhaqiqat *(actually)*
11	IMIT	UH	TAQLID *(Imitation)* kuk-kuk *(imitation of hen)*
12	INTJ	UV	UNDOV *(Interjection) Urraaa*! *(when you win)*

J. Ibragimov o‘zining “O‘zbek tili matnlarida so‘zlarning bevosita (kontakt) va bilvosita (distant) birikuvchanligini dastlabki dasturiy modellashtirish” maqolasida o‘zbek tilidagi matnlar tarkibidagi birikmalarni modellashtirishda umume’tirof etilgan ramzlardan foydalanganligini aytib, quyidagicha misollar keltiradi. *Ustodning muborak ko‘zoynaklarini taqib, xalqimning o‘tmishiga qaradim* gapida 1) *o‘tmishiga qaradim* - N + SFN + V + SFv. Bunda N - OT, SFN - otning sintaktik shakllari, V – fe’l, SFv – fe’lning sintaktik shakllari. *O‘tmishiga qaradim* kabi birikmalar tipik modelining submodelini quyidagicha berish mumkin bo‘ladi:

N + SG + PS_aff + Acc_CS_aff + V + TS aff + Shs aff. Bunda Not, SG - birlik, PS _aff - egalik shakli, Acc_CS_aff - jo‘nalish kelishigi, V – fe’l, TS aff (zamon shaklini inglizcha “Tense” so‘zi asosida shu ramz orqali ifodalashga qaror qildik), ShS aff- shaxs - son shakli (shaxs - son shaklini shunday ifodalashga qaror qildik). 2) *taqib qaramoq* - V + V + SFv. Bunda V – fe’l, V- fe’l, SFv – fe’lning sintaktik shakllari. 3) *xalqimning o‘tmishiga qaradim* - N^d + SFN + Nk + SFN + V + SFv 4) *ustodning ko‘zoynaklari* - N + SFN + N + SFN Bunda N + OT, SFN - otning sintaktik shakllari, N + ot, SFN - otning sintaktik shakllari. 5) *muborak ko‘zoynaklari* - ADJ + N + PS_aff . Bunda ADJ + sifat, N + ot, SFN - otning sintaktik shakllari [6].

Bundan tashqari S.Nazarova o‘zining “Birikmalarda so‘zlarning erkin bog‘lanish omillari” nomli tadqiqotida ism+ism qolipli so‘z birikmalarning modelini ishlab chiqqan.

1) [ot $^{\text{qaratqich kelishigi}}$ → ot $^{\text{egalik qo‘shimchasi}}$ = ot ^{qaratuvchi} + ot ^{qaralmish}]: kitobning varag‘i;

2) [ot $^{\text{qaratqich kelishigi}}$ → sifat $^{\text{egalik qo‘shimchasi}}$ = ot ^{qaratuvchi} + sifat ^{qaralmish}]: daraxtning mo‘rti;

3) [sifat $^{\text{qaratqich kelishigi}}$ → ot $^{\text{egalik qo‘shimchasi}}$ = sifat ^{qaratuvchi} + ot ^{qaralmish}]: gulning/qizilining hidi;

4) [sifat $^{\text{qaratqich kelishigi}}$ → sifat $^{\text{egalik qo‘shimchasi}}$ = sifat ^{qaratuvchi} + sifat ^{qaralmish}]: olmaning/kattasining chuchugi;

5) [ot $_{\text{turdosh}}$ $^{\text{qaratqich kelishigi}}$ → sifat $^{\text{egalik qo‘shimchasi}}$ = ot ^{qaratuvchi} + son ^{qaralmish}]: gulning bittasi. [7].

O‘zbek tilidagi so‘z birikmalarini teglash va modellashtirishda moslashuvli so‘z birikmasini namuna qilib olaylik, har bir qoidaga mos bo‘lgan quyidagi teg va madellarni yana bir namuna sifatida keltirib o‘tsak.

Nazariyalarda so‘z birikmasining ikki qismini tobe va hokim qism deb nomlashgan. Moslashuvli so‘z birikmasida esa bu ikki qismga qaratuvchi va qaralmish deb ham nomlashadi. Har ikki qismni nazariyasi va turkumiga ko‘ra otli va fe’lli birikmalariga bo‘lib o‘rganish an’anaviy tilshunoslikda juda ommalashgan.

Biz namunaviy teg va modellarimizni qaratuvchi va qaralmish terminlaridagi qoidalardan foydalangan holda ishlab chiqishni tanladik.

1. **Qoida:** Qaratuvchi kishilik olmoshi bilan ifodalanganida qaralmishga shu olmoshga mutanosib nisbatlovchi qo‘shiladi.

2-jadval. 1-qoida bo‘yicha modellashtirish

T/r	Namuna gaplar	O‘zbek tilidagi POS tegging modeli	Ingliz tilidagi POS tegging modeli	Moslashuvli so‘z birikmasining umumiy belgilanishi

1.	Mening uyim	P_ning+N_im	P+N	SB_{mos}
2.	Mening uylarim	P_ning+N_lar_im	P+N	SB_{mos}
3.	Sening uying	P_ning+N_ing	P+N	SB_{mos}
4.	Sening uylaring	P_ning+N_lar_im	P+N	SB_{mos}
5.	Uning uyi	P_ning+N_i	P+N	SB_{mos}
6.	Uning uylari	P_ning+N_lar_im	P+N	SB_{mos}
7.	Bizning uyimiz	P_ning+N_im	P+N	SB_{mos}
8.	Bizning uylarimiz	P_ning+N_im	P+N	SB_{mos}

2. **Qoida:** Qaratuvchi ot leksema bilan, boshqa turkum leksemalarining ot tabiatli turi, shuningdek otlashib kelgan holati bilan ifodalanganida qaralmishga III shaxs nisbatlovchisi qo‘shiladi:

3-jadval. ot leksema bilan ifodalanishi:

T/r	**Namuna gaplar**	**O‘zbek tilidagi POS tegging modeli**	**Ingliz tilidagi POS tegging modeli**	**Moslashuvli so‘z birikmasining umumiy belgilanishi**
1.	Uyning devori	N_ning+N_i	N+N	SB_{mos}
2.	Uylarning devori	N_lar_ning+N_i	N+N	SB_{mos}
3.	Uyning devorlari	N_ning+N_lar_i	N+N	SB_{mos}
4.	Uylarning devorlari	N_lar_ning+N_lar_i	N+N	SB_{mos}

4-jadval. olmoshning ot turlari bilan ifodalanishi:

T/r	**Namuna gaplar**	**O‘zbek tilidagi POS tegging modeli**	**Ingliz tilidagi POS tegging modeli**	**Moslashuvli so‘z birikmasining umumiy belgilanishi**
1.	Kimning kitobi	P_ning+N_i	P+N	SB_{mos}
2.	Kimlarning kitoblari	P_lar_ning+N_lar_i	PRON+PRON	SB_{mos}
3.	Allanarsaning sharpalari	P_ning+N_lar_i	PRON+N	SB_{mos}
4.	Allanarsalarning narsalari	P_lar_ning+N_lar_i	PRON+N	SB_{mos}

5-jadval. ikki va undan yuqori miqdorni anglatadigan miqdor son, shuningdek, hamma kabi leksema bilan ifodalanishi:

T/r	Namuna gaplar	O‘zbek tilidagi POS tegging modeli	Ingliz tilidagi POS tegging modeli	Moslashuvli so‘z birikmasining umumiy belgilanishi
1.	O‘nning yarm + i	NUM_ning+Num_i	NUM+Num	SB_{mos}
2.	Hammaning fikr + i	P_ning+N_i	PRON+N	SB_{mos}

Xulosa. Moslashuvli so‘z birikmalarini umuman olganda so‘z birikmasini (moslashuv, boshqaruv, bitishuv) teglash va modellashtirishning kompyuter lingvistikasi sohasidagi ahamiyati tilni avtomatik qayta ishlashda, balki uni aniq va to‘g‘ri tushunishda ham katta rol o‘ynaydi. Har bir grammatik qoidani va tuzilmani modellashtirish, kompyuter tizimlariga tilni yanada samarali o‘rganish va foydalanish imkonini beradi. Shu bois, so‘z birikmalarini teglash va modellashtirish masalalari kompyuter lingvistikasi sohasida dolzarb bo‘lib, NLP dasturlarini ishlab chiqish va rivojlantirishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Фортунатов Ф. Ф. О преподавании грамматики русского языша в; средней школе. Избраннме трудьг Т. II. М., 1957, с. 451.
2. Anorbekova A., Mirzayeva Sh. Hozirgi o‘zbek adabiy tili. – Toshkent, 2011. – B.171
3. Sobirjonova M. Moslashuvli so‘z birikmasi haqida ba’zi mulohazalar. - Toshkent, 2022.
4. Elov B.B. Xudayberganov N.U. O‘zbek tili korpusi matnlarini pos teglash usullari: file:///C:/Users/User/Downloads/elov-b.b.-xudayberganov-n.u.-ozbek-tili-korpusi-matnlarini-pos-teglash.pdf
5. Большакова Е.И., Воронцов К.В., Ефремова Н.Э., Клышинский Э.С., Лукашевич Н.В., Сапин А.С. Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие — М.: Изд-во НИУ ВШЭ, 2017. - 269 с.
6. Ibragimov J. O‘zbek tili matnlarida so‘zlarning bevosita (kontakt) va bilvosita (distant) birikuvchanligini dastlabki dasturiy modellashtirish. *// Zamonaviy dunyoda innovatsion tadqiqotlar: Nazariya va amaliyot nomli ilmiy, masofaviy, onlayn konferensiya materiallari.* https://doi.org/10.5281/zenodo.7515431
7. Nazarova S. Birikmalarda so‘zlarning erkin bog‘lanish omillari: filol. fan. nomz…. diss. avtoreferat. – Toshkent: 1997. – B.26.

UDK: 81.33

O‘ZBEK TILI FONETIKASINI TTS TIZIMLARI UCHUN FORMALLASHTIRISH

Maxmudjonova Gulshaxnoz Ulug‘bek qizi,
tayanch doktorant
gulshaxnozmahmudjonova@gmail.com
ToshDO‘TAU

Annotatsiya: Matndan ovozga aylantirish tizimlari (Text-To-Speech, TTS) o‘zbek tilida sifatli ishlashi uchun ushbu tilning fonetik va fonologik jihatlarini aniq formallashtirish zarur. Ya’ni, o‘zbek yozuvining harflarini fonemalarga birma-bir moslab, tovush o‘zgarishlari, urg‘u va intonatsiya qonuniyatlarini ham aniq belgilash kerak. Tilshunoslikda o‘zbek tilining tovush tizimi yaxshi o‘rganilgan bo‘lib, unli va undosh fonemalar, ularning xususiyatlari va o‘zaro munosabatlari ko‘plab ilmiy manbalarda yoritilgan. Ushbu maqolada o‘zbek tili fonetik-fonologik tizimi TTS uchun qanday formal ifodalash mumkinligi tahlil qilinadi.

Abstract: To ensure the high-quality performance of Text-To-Speech (TTS) systems in the Uzbek language, it is essential to formally define its phonetic and phonological properties. This includes establishing a clear mapping between written letters and phonemes, as well as systematically accounting for sound alternations, stress, and intonation patterns. While the sound system of Uzbek has been extensively studied in linguistics, with detailed analyses of its vowels, consonants, and their features, this article explores how these phonetic-phonological aspects can be formally represented for effective TTS development.

Аннотация: Для обеспечения качественной работы систем синтеза речи (Text-To-Speech, TTS) на узбекском языке необходимо чётко формализовать его фонетические и фонологические особенности. Это включает в себя установление точного соответствия между буквами письма и фонемами, а также описание звуковых изменений, ударений и интонационных закономерностей. Несмотря на то, что звуковая система узбекского языка хорошо изучена в лингвистике и подробно описана в научных источниках, в данной статье анализируется, как эти фонетико-фонологические аспекты могут быть формализованы для целей построения TTS-систем.

Kalit so‘zlar: *fonema, fonologiya, singarmonizim, ohang, urg‘u, formalizatsiya, TTS, konversiya*

O‘zbek tilining fonetik-fonologik xususiyatlari. O‘zbek tili turkiy tillar oilasiga mansub bo‘lib, fonetik tizimi o‘ziga xos tarixiy rivojlanish yo‘lini bosib o‘tgan. Fonetik tizimning kelib chiqishini o‘rganish, til tarixidagi tovush o‘zgarishlarini tushunish orqali ularni zamonaviy texnologik tizimlarga to‘g‘ri integratsiya qilish mumkin. Xususan, turkologiyada mashhur bo‘lgan A.N. Samoylovich va N.A.Baskakov [1] singari olimlarning tasniflari o‘zbek tilining fonetik shakllanish bosqichlarini chuqur tahlil qilish imkonini beradi. A.N. Samoylovich turkiy tillarni fonetik belgilar asosida bir nechta guruhlarga ajratgan. Unga ko‘ra, o‘zbek tili *“Тағлиқ”* guruhiga mansub bo‘lib, ushbu guruhga xos asosiy fonetik belgilar — *тоғ, тоққиз, айақ, бол, сариғ, қалған* kabi qadimiy fonetik shakllarning saqlanib qolganidir. Bu guruhdagi tillarda, ayniqsa, *тоғ* so‘zining *тағ* shaklida talaffuz qilinishi, qadimgi til qatlamining saqlanganligidan dalolat beradi. Samoylovichning ta’kidlashicha, ushbu guruhga qadimgi o‘zbek tili, hozirgi o‘zbek va uyg‘ur tillari kiradi. Turkiy tillarning fonetik jihatdan shakllanishini yanada chuqurroq tahlil qilgan N.A. Baskakov esa o‘zbek tilini *Qarluq guruhi* tarkibida ko‘rsatadi. Bu guruhga xos bo‘lgan fonetik belgilar orasida quyidagilar ajralib turadi:

unlilar sonining 6 tadan 9 tagacha o‘zgaruvchanligi;
q va k fonemalarining farqlanishi;
so‘z oxirida j, g, g‘ tovushlarining saqlanishi;
p, t, k, q jarangsiz fonemalarining ikki unli orasida ham jarangsizligicha qolishi;
qaratqich kelishigining -ning shakli bilan to‘liq qo‘llanilishi

Bu belgilar hozirgi adabiy o‘zbek tili fonetik sistemasining tarkibiy qismlarini tashkil etadi va ularni zamonaviy G2P (grafema-fonema konversiyasi) yoki TTS modellarida qat’iy formal qoidalarga solish lozimligini ko‘rsatadi. Tarixiy fonetik tavsiflardan foydalanish TTS modellarida fonetik birliklar to‘plami yanada chuqurroq aniqlashga yordam beradi. Masalan, eski o‘zbek til qatlamlarida mavjud bo‘lgan **[ü], [ö], [ɨ]** fonemalari hozirgi o‘zbek tilda rasmiylashtirilmagan bo‘lsa-da, ayrim dialektlar yoki tarixiy matnlarni ovozli sintezlashda bu fonemalar kerak bo‘lishi mumkin. Shu sababli tarixiy fonetik tahlil zamonaviy texnologik model uchun kengaytirilgan fonema tizimi yaratishga asos bo‘lib xizmat qiladi. Umuman olganda, o‘zbek tilining fonetik formallashtirilishi tarixiy taraqqiyot bosqichlarini inobatga olgan holda amalga oshirilsa, semantik va morfologik jihatdan aniq, barqaror fonemik modelga ega bo‘lish mumkin bo‘ladi. Fonemalar tizimini ilmiy tavsiflashda Mirtojiyev [2], Abduazizov [3] kabi olimlar unli va undoshlarning farqlovchi belgi va variantlarini chuqur tahlil qilgan. Jumladan, unlilarning cho‘ziqlik-qisqaligi, lablanish darajasi, undoshlarning portlovchi yoki sirg‘aluvchan turi kabi xususiyatlar fonologik tahlillarda e’tiborga olingan.

Hozirgi o‘zbek tili tovush tizimida unli fonemalar soni an’anaviy ravishda 6 ta deb qabul qilingan: *a, e, i, o, u, o‘*. Ushbu unlilar til pozitsiyasi (old-orqa) va lablarning ishtirokiga ko‘ra farqlanadi. Masalan, i – tor, old qator, lablanmagan unli; u – tor, orqa qator, lablangan unli; a – keng, old qator (yoki markaziy) lablanmagan unli; o – orta-keng, orqa qator, biroz lablangan unli tovush. O‘zbek tilining undosh fonemalarini Irisqulov, Tursunov va Rahmatullayev 25 ta deb keltiradi [4], Jamolxonov va Sayfullayeva esa kitoblarida 24 ta ekani ko‘rsatiladi [5]. Undoshlar ovoz va shovqinning ishtirokiga ko‘ra: jaranglilik (jarangli va jarangsiz), hosil bo‘lish o‘rni va usuli kabi belgilarga ko‘ra tasniflanadi.

O‘zbek tili fonologiyasining muhim xususiyati – singarmonizm (unli fonemalarning uyg‘unlashuvi)nin mavjudligidir. Qadimgi turkiy til va eski o‘zbek tilida qat’iy fonetik uyg‘unlashuv bo‘lib, unlilar **old** va **orqa qator** opozitsiyasini hosil qilgan (masalan, *i/ы, ü/u, ö/o, ä/a* qarama-qarshi juftliklari). O‘zbek tilida ham ba’zan bir bo‘g‘inli so‘zlarning struktura elementlarini — unli va undosh tovushlar o‘rtasida tovush qatorlari bo‘yicha o‘zaro xoslanish hollari kuzatiladi. Xususan, GC, CGC strukturali so‘zlarda tovushlar palatal garmoniyaga muvofiq birikadi: *қил, хил, бил, тил, сиз, жил* va bosh. Ba’zan old va orqa qator variantlari ma’no farqlash funksiyasini ham bajarishi mumkin. Masalan, *tıq — tik*. Bir qator so‘z yasovchi va forma yasovchi affikslar variativlikka ega. O‘zak fonetik strukturasiga muvofiq affikslarning variantlaridan biri qo‘llaniladi. Affikslar variantlari shu fonetik qurshovda biri o‘rnida, ikkinchisi qo‘llanolmaydi. Masalan: — ық ё — ик (қылық — кичик), — қы (— гы) — ги (чопқы — чолғы — сузги), — қа — га (тоққа — бизга) va boshqalar. Bu faktorlar turkiy tillar uchun xarakterli bo‘lgan singarmonizm qonuniyatining o‘zbek tilida ham ayrim o‘rinlarda mavjudligini ko‘rsatadi[6]. Yuqoridagi umumiy tavsifdan ko‘rinadiki, o‘zbek tilining fonemalar majmuasi va ularning xususiyatlarini aniqlash TTS tizimi uchun poydevor vazifasini o‘taydi. Quyida avvalo o‘zbek tilidagi unlilarni formal tizimga solish va harflardan fonemalarga mapping (mos tushirish) masalasi ko‘rib chiqiladi.

Tovush hodisalarini formalizatsiya qilish: pozitsion o‘zgarishlar va assimilyatsiya. O‘zbek nutqida tovushlar kombinatsiyasiga ko‘ra turli fonetik hodisalar sodir bo‘ladi. Eng ko‘p uchraydigani – assimilyatsiya, ya’ni bir xil kategoriyadagi tovushlarning (undosh bilan undoshning) bir-biriga moslashuvi [7]. Dissimilyatsiya – bir tovushning boshqa tovushga ko‘chishi) Bundan tashqari, fonetik hodislarning boshqa turlariga eliziya – unli tovush bilan tugagan so‘zga unli bilan boshlangan so‘zning qo‘shilishi natijasida bir unli tovushning tushishi (ayta oladi > aytoladi) ham kuzatiladi. Nutqda ortiqcha tovush aytilmaydi. Masalan, *“shanba”* so‘zini [shamba] deb talaffuz qilish odatiy. Bunday hodisalarni TTS matnni sintezlaganda odatda muhim hisoblamaydi (chunki u yozuvga to‘liq amal qiladi). Ammo jonliroq intonatsiya uchun ba’zi holatlarda qo‘llanishi mumkin. Formal jihatdan, eliziyaga oid qoidalar kiritilishi murakkab bo‘lib har doim ham kerak emas, chunki orfografik matnni imkon qadar to‘liq o‘qish TTSning vazifasi.

TTS tizimi matnni ovozga aylantirishda bu hodisalarni hisobga olib, sun’iy nutqni tabiiy nutqqa yaqinlashtirishi lozim. Buning uchun fonetik qoidalar bazasi yoki formal tovush o‘zgartirish modullari tuziladi. **Pozitsion o‘zgarishlar** – tovushning so‘z boshida, o‘rtasida yoki oxirida kelishiga bog‘liq holda o‘zgarishi. O‘zbek tilida ba’zi jarangli undoshlar so‘z oxirida kuchsiz pozitsiyada qisman jarangsizlanishi mumkin. Masalan, **g‘** harfi so‘z oxirida [q] yoki [x]ga yaqin eshitiladi: *“yorug‘”* [yoruq] kabi. Yoki aksincha, **q** so‘z oxirida ba’zan [g‘] tarzida jarangli eshitilishi mumkin (Toshkent shevasida: *“yo‘q”* → [yog‘]). Bu holatlar ko‘proq shevada namoyon bo‘ladi. TTSda bunday o‘zgarishlarni aks ettirishda odatda adabiy me’yorga yaqinroq talaffuz tanlanadi. Bunday hodisalarni qoida sifatida kiritish-kiritmaslik sintezatorni qanday nutq modeliga yo‘naltirilganiga bog‘liq – standart shaklda kiritmaslik ham mumkin. Yuqoridagi misollar shuni ko‘rsatadiki, o‘zbek tilining fonetik hodisalarini formalizatsiya qilish – ya’ni, qoidalar ko‘rinishida ifodalash – TTS sifatini oshirishga yordam beradi. Maxsus fonologik qoidalar to‘plami tuzilib, grafemadan fonemaga o‘tkazishda qo‘llanadi. Bu jarayon tabiiy tilni qayta ishlashda muhim ahamiyatga ega bo‘lib, matndan nutqqa va avtomatik nutq tanish tizimlarida ham qo‘llanadi. Grafema-fonema konversiyasi va unga oid qoidalarni ishlab chiqish o‘zbek tili uchun juda dolzarb [8] – bu borada maxsus modellarning nazariy asoslari va texnik yechimlari ishlab chiqilmoqda. Demak, TTS tizimi uchun fonetik hodisalarni bir mexanizmga solish – sintezator “qoidalar bazasi”ni yaratish deganidir. Reallashgan misol (qoidalar qo‘llanilishi): masalan, TTS tizimi *“Yaxshi ko‘rishingizni bilaman”* jumlasini ovozga aylantiradi. Avvalo grafemalar fonemaga o‘tkaziladi:

yaxshi	jˈæxʃy
ko‘rishingizni	kˌoryʃyŋˈyzny
bilaman bylˈæmæn	bylˈæmæn

Keyin qoidalar qo‘llanadi: /jˈæxʃy/ shu tarzda qoladi: /kˌoryʃyŋˈyzny/ ichida “ng” → [ŋ] qoidasiga binoan, ya’ni ng harfi oldidan unli kelsa, n va g talaffuz qoidasi kiritiladi va /kˌoryʃyngˈyzny/ bo‘ladi; /bilæmæn/ ichida esa biror o‘zgarish talab qilinmaydi. So‘ngra urg‘u va ohang belgilari qo‘shilib, tovush sintezi amalga oshiriladi. Natijada iloji boricha tabiiy jarangga ega nutq hosil bo‘ladi.

Urg‘u va ohangning TTSda modellashtirilishi. Ovoz sintezida faqat fonemalarni to‘g‘ri aniqlash yetarli emas – prozodiya, ya’ni urg‘u va ohang (intonatsiya)ni ham to‘g‘ri berish kerak. O‘zbek tilida urg‘u so‘z tarkibidagi bo‘g‘inlardan biriga tushadigan urg‘udir. Intonatsiya esa gap turiga ko‘ra shakllanadi (masalan, so‘roq gap ohangi ko‘tariladi, hayrat ohangi maxsus intonatsiya bilan ifodalanadi). TTS tizimida buni modellashtirish uchun bir necha yondashuv mavjud: matnga maxsus belgi qo‘yish, qoida asosida ohang yaratish yoki

neyron tarmoqlar orqali prozodiya sintezi. Bizning vazifamiz – formallashtirish nuqtai nazaridan buni qanday ifodalash mumkinligini ko‘rsatish.

Urg‘uni formal belgilash: o‘zbek tilida urg‘u qoidasi sodda – so‘zning ohirgi bo‘g‘ini urg‘uli. Demak, TTS bu qoidani default holat sifatida qabul qilishi mumkin. Biroq ba’zi so‘zlar (yoki qo‘shimchalar) urg‘uga ega emasligi ma’lum: masalan, -mi so‘roq yuklamasi, -dir tasdiq yuklamasi urg‘u olmaydi, urg‘u so‘zning oxirgi asosiy bo‘g‘inida qoladi (*“keldimmi”* so‘zida **mi** urg‘usiz). Shuning uchun TTS uchun quyidagi qoidalarni formallashtirish kerak:

Urg‘u qoidasining formal ifodasi: “har bir so‘zda oxirgi bo‘g‘in urg‘uli, agar so‘z -mi, -chi, -dir kabi urg‘usiz yuklamalar bilan tugasa, urg‘u o‘sha yuklamadan oldingi bo‘g‘inga ko‘chadi”. Bu qoida asosida sintezator fonemik chiqishda urg‘uli bo‘g‘in oldiga [ˈ] belgisini qo‘yishi mumkin (IPA transkripsiyada urg‘u belgisi). Masalan: *“yaxshi ko‘rishingizni bilaman”* misolidagi *“bilaman”* so‘zi bo‘lsa, urg‘u oxirgi bo‘g‘inda, yoki *“keldingizmi”* (urg‘u **-ingiz** bo‘g‘inida, chunki **mi** urg‘usiz) deb belgilash. Buni belgilash uchun ko‘p tizimlarda matnga diakritik belgi qo‘yiladi yoki maxsus markup ishlatiladi. Masalan, SSML (Speech Synthesis Markup Language) da <emphasis level="strong"> kabi teglar bilan urg‘uni ko‘rsatish mumkin. Bizning holatda, soddalik uchun, fonetik notatsiyaga [ˈ] belgisini qo‘yish kifoya.

Ohang (intonatsiya) modellashtirish: Intonatsiyani formallashtirish urg‘uga nisbatan qiyinroq, chunki ohang uzluksiz o‘zgaruvchi kattalik (ton) bo‘lib, gapning turiga va ichki tuzilishiga bog‘liq. Ammo TTSda asosan quyidagi qoidalar qo‘llanadi:

Gap tugagach, ohangning ko‘tarilish-pasayishi gapning maqsad bog‘lanishiga (kommunikativ funksiyasiga) bog‘liq. Darak gap ohangi odatda pasayish bilan yakunlanadi (kuchli nuqta intonatsiyasi), so‘roq gap ohangi ko‘tarilib tugaydi (so‘roq intonatsiyasi), buyruq yoki undov gaplarda esa ohang mazmuniga ko‘ra maxsus egri chiziq hosil qiladi (masalan, buyruqda qat’iy pasayish).

Bu qoidalarni formal ifodalash uchun TTS kirish matnida belgilar yoki teglar ishlatiladi. Masalan, so‘roq gapni belgilovchi ? belgisi tizimga ohangni ko‘tarib tugatishni buyuradi. Agar oddiy matn bo‘lsa, TTS punktuatsiya belgilariga qarab intonatsiya tanlaydi.

Formal misol: “Siz keldingiz.” va “Siz keldingizmi?” jumlalarini olaylik. Birinchisi darak gap, ikkinchisi so‘roq. TTS uchun qoidalar: “agar gap oxirida [?] bo‘lsa, ohang = so‘roq ohangi (intonatsiya ko‘tariladi)”, “agar [.] bo‘lsa, ohang = tugash ohangi (pasayadi)”. Bu qoida asosida sintezator ohangni belgilaydi.

Intonatsion pauza va urg‘u ham muhim: murakkab gaplarda pauzalar kerak bo‘ladi. Masalan, “Men, albatta, ertaga boraman” gapida verguldan keyin qisqa pauza bor, ohang biroz tushadi va keyin davom etadi. TTS buning uchun vergulni pauza sifatida qabul qilib, intonatsiyani moslashtiradi. Bu ham qoidaga keltiriladi: “, belgisi = qisqa pauza, ohang biroz ko‘tarilib davom ettirilsin” kabi.

TTS uchun urg‘u va intonatsiya modellashtirishni formal qoidalar orqali hal etish mumkin, lekin bugungi kunda ko‘proq neyron tarmoqli sintezatorlar bu jarayonni o‘z-o‘zidan o‘rganadi.

Fonetik formallashtirishning zarurligi. TTS tizimlari uchun fonetik formallashtirish – ya’ni, tilning tovush tizimini va talaffuz qoidalarini rasmiy qoidalarga solish – nihoyatda muhim ahamiyatga ega. Buning bir necha asosli sabablari bor:

To‘g‘ri va tabiiy talaffuz: Agar o‘zbek tilining harf-fonema mosligi va tovush hodisalari aniq formalashmasa, TTS noto‘g‘ri yoki notabiiy talaffuz chiqarishi mumkin.

Yozuv va talaffuzdagi farqlarni aniqlash: O‘zbek tilida ayrim so‘zlarda yozuv va talaffuz farqlanishi mumkin (transliteratsiya, qadimiy ismlarda, xorijiy nomlarda). Qaysi so‘zlarda ng harfi, qaysi so‘zlarda n va g harfi talaffuz qilinishini qoidaga kiritish kerak. Masalan: *Tungi, keyingi, oldingi* kabi so‘zlarda n va g harflari ajratilsa, ko‘ngil, singil,, tong so‘zlarida ajratilmaydi. Formal qoidalar esa har bir mumkin bo‘lgan grafik kombinatsiyani to‘g‘ri talaffuzga xaritalaydi.

Standartlashtirish: Fonetik formallashtirish tilni standart me’yor asosida “kompyuter tushuna olish” shakliga keltiradi. Bu til me’yorini saqlash va targ‘ib qilishga ham xizmat qiladi. Masalan, yuqorida aytilganidek, ayrim kishilar *“yo‘q”* so‘zini [yok] deb talaffuz qilishi mumkin, boshqalar [yoq] deb, yoki [yog‘] shaklida. TTS esa to‘g‘ri standart [yo‘q] tarzida aytishi lozim. Formal model buni ta’minlaydi. TTS shu me’yorni ushlab qoladi va butun auditoriyaga yagona, aniq talaffuzni yetkazadi.

Yuqoridagi omillar tufayli fonetik formallashtirish TTS jarayonining muhim qismi sanaladi. O‘zbek tilining fonetik formallashtirilgan modeli TTS tizimining “miya”sidir. Aynan shu model matnning harflar ketma-ketligini jonli tovushlarga aylantirish yo‘lini belgilab beradi. Bu model qanchalik to‘liq va ilmiy asosli tuzilgan bo‘lsa, TTS ovozi shunchalik tabiiy va ravon chiqadi.

Maqolada “O‘zbek tili fonemasini TTS tizimlari uchun formallashtirish” masalasi har tomonlama ko‘rib chiqildi. Avvalo, o‘zbek tilining fonetik va fonologik tizimi – unli va undosh fonemalar majmuasi – ilmiy manbalar asosida tahlil qilindi. O‘zbek tilida 6 ta unli va ~24 ta undosh fonema adabiy me’yorda mavjud bo‘lsa-da, ayrim tadqiqotchilar fonologik oppozitsiyalarni e’tiborga olib unlilar sonini 8–9 tagacha deb hisoblashadi. O‘zbek tili uchun harflardan fonemalarga o‘tkazish qoidalari aniq ifodalash TTSning grafema-fonema konversiya modulini shakllantiradi. Urg‘u va intonatsiya uchun TTSga qoida berish usullari muhokama qilinib, so‘z urg‘usini belgilash va gap ohangini tinish belgilari orqali boshqarish yondashuvlari keltirildi.

Fonetik formallashtirishning zarurligi, xususan, TTS sifatini oshirish, to‘g‘ri talaffuzni ta’minlash va integratsiyalashgan tizimlar uchun muhim qoidalar asoslandi. Ilmiy manbalarga tayanib aytish mumkinki, har bir yangi til uchun TTS

yaratishda birinchi qadam – shu tilning fonema inventari va orfoepik qoidalarini to‘liq ishlab chiqishdir [9]. O‘zbek tili ham bundan mustasno emas. Ta’kidlash joiz, o‘zbek tilida ilgari bunday formal model va qoidalar to‘plami yaratilgan ishlar kam (yoki ichki loyihalar doirasida qolgan).

O‘zbek tilining TTS tizimi uchun fonetik formallashtirilishi – bu tilning ilmiy asosda “raqamlashtirilishi” demakdir. Ushbu maqolada keltirilgan tavsif va takliflar real TTS tizim qurishda dastlabki qo‘llanma vazifasini o‘tashi mumkin. Kelgusida bu sohada yanada chuqurroq izlanishlar olib borilib, o‘zbek tilining intonatsion modellari, kontekstga oid ko‘p variantli talaffuz (masalan, omonim so‘zlarni farqlash) kabi masalalar ham hal etilsa, ona tilimiz zamonaviy nutq texnologiyalarida to‘laqonli qo‘llanilishiga erishiladi.

Foydalanilgan adabiyotlar:

1. Mahmudov Q. O‘zbek tilining tarixiy fonetikasi. — Toshkent: Fan, 2006. — 168 b.
2. Mirtojiyev M.M. O‘zbek tili fonetikasi. — Toshkent: Fan nashriyoti, 2013.
3. Abduazizov A.A. O‘zbek tilining fonologiyasi va morfonologiyasi. — Toshkent: O‘qituvchi, 1992.
4. Irisqulov M.T. Tilshunoslikka kirish. 2-nashr. — Toshkent: Yangi asr avlodi, 2009.
5. Sayfullayeva R.R., Mengliyev B.R., Qurbonova M.M., Raupova L.R., Abuzalova M.Q., Yo‘ldosheva D.N. Hozirgi o‘zbek tili. — Buxoro: Durdona nashriyoti, 2021.
6. Nurmonov A. O‘zbek tilining fonologiyasi va morfonologiyasi. – Toshkent: O‘qituvchi, 1990. – 128 b.
7. Jamolxonov H. Hozirgi o‘zbek adabiy tili. — Toshkent: Talqin, 2005.
8. Maxmudjonova G.U. Grafema-fonema (G2P) modeli: nazariy asoslari va o‘zbek tili uchun ahamiyati. Qo‘qon DPI ilmiy xabarlari, 2025.
9. Nath C., Sarma B. A Grapheme to Phoneme Based Text to Speech Conversion Technique in Unicode Language // Data and Metadata, 2023.

O‘ZBEK TILINING BAZIS SO‘ZLARIGA ASOSLANGAN KONTEKST-ERKLI GRAMMATIKA ISHLAB CHIQISH VA UNING SINTEZI

Nilufar Soliyeva Qahramon qizi,
magistrant
nilufarsoliyeva732@gmail.com
UrDU

Annotatsiya. Mazkur tezisda kontekst-erkli grammatika tushunchasining nazariy asoslari, uning tabiiy tilni qayta ishlashdagi ahamiyati va o‘zbek tiliga moslashuv imkoniyatlari tahlil qilinadi. Ayniqsa, o‘zbek tilining morfologik xususiyatlari, sintaktik qurilishi va bazis (asosiy) so‘zlar tarkibini aniqlash mezonlari asosida kontekst-erkli model uchun zarur bo‘lgan grammatik tuzilmalarni shakllantirish imkoniyatlari ko‘rib chiqiladi. Shuningdek, mavjud grammatik yondashuvlar bilan solishtirma tahlil o‘tkazilib, ushbu yondashuvning afzalliklari va qo‘llanish muammolari haqida xulosa qilinadi. Ushbu tahliliy yondashuv o‘zbek tilini avtomatik qayta ishlash, xususan, mashina tarjimasi, avtomatik tahlil va sintez jarayonlarida yangi ilmiy asos bo‘lishi mumkin.

Abstract. This thesis examines the theoretical foundations of context-free grammar, its significance in natural language processing, and its adaptability to the Uzbek language. Special attention is given to the morphological features and syntactic structure of Uzbek, along with the criteria for identifying basic (core) vocabulary, in order to construct the necessary grammatical structures for a context-free model. A comparative analysis is also conducted with existing grammatical approaches, outlining the advantages and limitations of this method. This analytical perspective may serve as a new scientific foundation for automatic processing of the Uzbek language, particularly in machine translation, automatic analysis, and synthesis tasks.

Аннотация. В данной тезисной работе рассматриваются теоретические основы понятия контекстно-свободной грамматики, её значение в обработке естественного языка и возможности адаптации к узбекскому языку. Особое внимание уделяется морфологическим особенностям узбекского языка, его синтаксической структуре, а также критериям определения базисных слов с целью формирования необходимых грамматических конструкций для модели КСГ. Кроме того, проводится сравнительный анализ с существующими грамматическими подходами, выделяются преимущества и проблемы применения данного подхода. Представленный аналитический обзор может служить научной основой для автоматической обработки узбекского языка, в

частности в задачах машинного перевода, автоматического анализа и синтеза.

Kalit so‘zlar: *kontekst-erkli grammatika, o‘zbek tili, sintaksis, morfologiya, bazis so‘zlar, tabiiy tilni qayta ishlash, mashina tarjimasi*.

Zamonaviy sun’iy intellekt va tabiiy tilni qayta ishlash (NLP) texnologiyalarining rivojlanishi natijasida til strukturasi va uning kompyuter tomonidan avtomatik tarzda qayta ishlanishi dolzarb muammoga aylangan. Bu jarayonda tilning sintaktik va morfologik xususiyatlarini formal modellar asosida ifodalash alohida ilmiy yo‘nalishga aylangan. Ayniqsa, kontekst-erkli grammatika (KEG) modeli shu kabi ehtiyojlarga samarali yechim sifatida ko‘plab tillarda qo‘llanilib, muvaffaqiyatli natijalar bermoqda. Ushbu model yordamida grammatik qoidalarga asoslangan holda gaplarni avtomatik sintez qilish, ularni sintaktik jihatdan tahlil qilish va kompyuter dasturlarida real qo‘llash imkoniyati yuzaga keladi. Kontekst-erkli grammatika dastlab Noam Chomsky tomonidan 20-asrning ikkinchi yarmida nazariy jihatdan ilgari surilgan bo‘lsa-da, amaliyotda, ayniqsa so‘nggi yillarda tabiiy tillarni qayta ishlash, mashina tarjimasi, ovozli yordamchilar, chatbotlar va avtomatik matn yaratish tizimlarida keng qo‘llanila boshlandi. Bu modelning afzalligi shundaki, u har bir til birligining tuzilmaviy rolini aniqlab, uni kontekstda joylashuviga ko‘ra ifodalash imkonini beradi. Bunda sintez jarayoni grammatik qoidalarga tayanib, tilning mantiqiy va shakliy tuzilmasini aniq aks ettiradi.

Mazkur tezisda aynan shunday yondashuv ilgari suriladi: o‘zbek tilining eng ko‘p ishlatiladigan, grammatik va semantik jihatdan faol bazis so‘zlari tanlab olinadi, ularning morfologik variantlari va sintaktik holatlari KEG qoidalari asosida modellashtiriladi, so‘ngra bu qoidalar asosida avtomatik sintez mexanizmi shakllantiriladi. Bu jarayon orqali o‘zbek tilidagi grammatik jihatdan to‘g‘ri va kontekstga mos gaplarni yaratish imkoniyati yuzaga keladi. Shuningdek, bu tizim mashina tarjimasi, til o‘rgatuvchi platformalar va boshqa NLP tizimlariga samarali integratsiya qilinishi mumkin.

O‘zbek tili kontekst-erkli grammatika modellarini ishlab chiqish uchun juda qulay va shu bilan birga murakkab til hisoblanadi. Bu, bir tomondan, o‘zbek tilining agglutinativ (yopishuvchi) morfologik tuzilishi, ikkinchi tomondan esa sintaktik erkinlik darajasining nisbatan yuqoriligi bilan bog‘liq. O‘zbek tilida so‘zlar o‘zaro qo‘shilib, grammatik ma’no hosil qiluvchi affikslar yordamida murakkab birliklar tashkil etadi. Shu sababli, har bir birlikning grammatik vazifasi kontekstdan kelib chiqib o‘zgaradi. Masalan, “kelgan” so‘zi turli kontekstlarda sifatdosh, fe’l, yoki hatto gapda kesim vazifasida ishtirok etishi mumkin.

Kontekst-erkli grammatika aynan mana shu o‘zgaruvchanlikni tahlil qilish va ifodalashda muhim rol o‘ynaydi. Bu modelda har bir qoidaning amal qilishi uchun uning chap va o‘ng konteksti aniqlanadi. Bu, o‘z navbatida, o‘zbek tilida fe’l

shakllarining zamon, nisbat, shaxsga qarab moslashuvi, holatlar bilan bog‘lanishi va gapda vazifasining farqlanishini to‘g‘ri tahlil qilish imkonini beradi. Masalan:

Men bor**dim** (fe’l – o‘tgan zamon, 1-shaxs, birlik)

Sen bor**ding** (fe’l – o‘tgan zamon, 2-shaxs, birlik)

U bor**di** (fe’l – o‘tgan zamon, 3-shaxs, birlik)

Ushbu uchta shakl morfologik jihatdan o‘xshash bo‘lsa-da, kontekst-erkli model ularning o‘zaro farqini tahlil qilish va sintetik yo‘l bilan hosil qilishda muhim ahamiyat kasb etadi. O‘zbek tilining bazis (asosiy) so‘zlariga asoslangan model yaratishda dastlab, tilning eng ko‘p ishlatiladigan va grammatik jihatdan eng faol birliklarini aniqlash zarur. Bunday so‘zlar, odatda, o‘zbek tilining ommaviy matnlar korpusi asosida, chastotali tahlil orqali aniqlanadi. Masalan, “bor”, “kel”, “ol”, “qil”, “bilan”, “uchun”, “men”, “sen”, “u”, “shu”, “bu” kabi birliklar bazis korpusining asosini tashkil qilishi mumkin. Bazis so‘zlarga asoslangan kontekst-erkli grammatika yaratishda yana bir muhim bosqich – bu bazis birliklarning sintaktik va morfologik variantlarini tizimlashtirishdir. Har bir so‘z o‘zining imkonli kombinatorik holatlari, ya’ni qanday affikslar bilan qanday kontekstlarda bog‘lanishi mumkinligi bo‘yicha tahlil qilinadi. Bu jarayonda qoidaga asoslangan (rule-based) va statistik usullar qo‘llanishi mumkin. Masalan, “ol-” fe’li:

ol**di** – o‘tgan zamon, 3-shaxs

ol**dim** – o‘tgan zamon, 1-shaxs

ol**gan** – sifatdosh

ol**indi** – majhul nisbatda

Shuningdek, bu shakllarning kontekstda qanday ishlatilishini aniqlashda avtomatik tahlil tizimlarining yordami zarur bo‘ladi. Bu yerda mashina o‘rganishi (machine learning) yondashuvlari, ayniqsa neyron tarmoqlar asosida ishlovchi modellar katta ahamiyat kasb etadi. Ammo ularga kontekst-erkli qoida tizimi integratsiya qilinmaguncha, bunday modellar o‘zbek tilining murakkab grammatikasini to‘liq o‘zlashtira olmaydi. Bundan tashqari, sintetik grammatika — ya’ni, til birliklarini yaratish (sintez) yo‘li bilan yangi shakllarni hosil qilish — kontekst-erkli yondashuv asosida ancha ishonchli va moslashuvchan bo‘ladi. Masalan, avtomatik matn yaratish yoki mashina tarjimasi tizimida o‘zbek tilining grammatik shakllari foydalanuvchining kontekstiga mos tarzda sintez qilinadi: ya’ni, model “sen uchun yozgan edim” shaklini foydalanuvchi ifodalayotgan nuqtai nazarga mos shaklda yaratadi. O‘zbek tilida erkin so‘z tartibi mavjudligi (masalan, “men kitobni o‘qidim”, “kitobni men o‘qidim”, “o‘qidim men kitobni”) kontekst-erkli yondashuvni talab qiladi. So‘z tartibining erkinligi formal qoidalarga

asoslangan modelda chalkashliklarga olib keladi, kontekst-erkli model esa bu holatlarni mantiqiy kontekst orqali farqlashga qodir.

Tabiiy tilni avtomatik tarzda qayta ishlash texnologiyalari hozirgi kunda sun’iy intellekt sohasida katta rol o‘ynamoqda. Bu texnologiyalarning muhim yo‘nalishlaridan biri — til sintezi bo‘lib, u grammatik jihatdan to‘g‘ri gaplar va matnlar yaratish imkonini beradi. Sintez jarayonida ishlatiladigan asosiy model – bu formal grammatikalardan biri bo‘lgan kontekst-erkli grammatika (KEG) hisoblanadi. KEG yordamida tilning sintaktik tuzilmasi qat’iy qoidalar asosida tavsiflanadi va bu qoidalar asosida yangi gaplarni yaratish imkoniyati tug‘iladi. KEG modeli to‘rtta asosiy komponentdan tashkil topadi:

N – no-terminal belgilar to‘plami, Σ – terminal belgilar (ya’ni real so‘zlar), P – produktsiyalar (o‘zgartirish qoidalari), S – boshlang‘ich belgi (odatda, gapni anglatadi). KEG qoidalari yordamida til strukturasi daraxtsimon shaklda ifodalanadi va sintaktik bog‘liqliklar aniqlanadi.

Sintez jarayoni KEG modelidagi qoidalarning teskari yo‘nalishda ishlatilishi orqali amalga oshiriladi. Ya’ni, avval boshlang‘ich belgi — odatda **S** belgisi olinadi, so‘ngra unga mos produktsiya qoidasi tanlanadi va u bosqichma-bosqich terminal belgilar to‘plamiga aylantiriladi. Natijada, grammatik jihatdan to‘g‘ri gap yoki ibora hosil qilinadi. Bu jarayon, kompyuter dasturlari yordamida avtomatlashtirilgan holda bajariladi va real til sintezi amaliyotida juda muhim ahamiyat kasb etadi.

Misol tariqasida quyidagi qoidalarni olaylik:

S → NP VP
NP → Det N
VP → V NP
Det → bu | bir
N → bola | o‘quvchi | kitob
V → o‘qiydi | yozadi | oladi

Ushbu qoidalardan foydalangan holda sintez jarayoni quyidagicha kechadi:

1) Boshlang‘ich belgi sifatida S belgilanadi.
2) S → NP VP qoidasi asosida NP va VP hosil qilinadi.
3) NP → Det N qoidasi orqali NP belgisi “bu bola” shaklida sintez qilinadi.
4) VP → V NP esa “o‘qiydi kitob” kabi segmentlarni beradi. Yakuniy natijada, masalan, “Bu bola kitob o‘qiydi” gapini hosil qilish mumkin bo‘ladi. Bu esa til sintezining eng sodda, lekin amaliy modelidir.

Yuqoridagi jarayon davomida hosil bo‘lgan natijani sintaktik daraxt (parse tree) shaklida ham ifodalash mumkin. Bu daraxt tildagi har bir birikmaning qanday

grammatik komponentlardan tashkil topganini ko‘rsatadi. Masalan, S – gap elementi NP (ot birikmasi) va VP (fe’l birikmasi) orqali tuzilganini, NP esa Det (aniqlovchi) va N (ot) orqali ifodalanganini ochiq-oydin ko‘rsatadi. Bu sintaktik daraxt KEG asosidagi sintez mexanizmining vizual va tushunarli ifodasidir.

Shuningdek, KEG asosidagi sintez mexanizmi til strukturasi qat’iy bo‘lgan agglutinativ tillar, jumladan o‘zbek tili uchun ham moslashtirilishi mumkin. O‘zbek tilining morfologik jihatdan boyligi (ya’ni qo‘shimchalar tizimi), SOV tartibi (ega-to‘ldiruvchi-kesim) va kesimning gap oxirida kelishi KEG qoidalari orqali ifodalanib, sintez mexanizmida muvaffaqiyatli qo‘llanishi mumkin. Bunda bazis so‘zlar (eng ko‘p uchraydigan va grammatik jihatdan markaziy so‘zlar) KEGning terminal belgilariga aylanadi. No-terminal belgilar esa tilning asosiy sintaktik birliklarini tashkil qiladi: masalan, NP, VP, PP (predlogli birikma) va hokazo.

Sintez mexanizmining algoritmik bosqichlari quyidagicha ifodalanadi:

1. Boshlang‘ich belgi tanlanadi (S);
2. S ga mos produktsiya qoidalari tanlanadi;
3. Har bir no-terminal belgi mos terminal yoki no-terminal birliklarga aylantiriladi;
4. Terminal belgilar orqali gap yakuniy ko‘rinishda hosil qilinadi.

Ushbu ketma-ketlik asosida KEG sintez mexanizmi hatto murakkab gaplarni hosil qilishga imkon beradi. Masalan, ergash gaplar, bog‘lovchili yoki bog‘lovchisiz gaplar, holli fe’l shakllari, inkor shakllari ham ushbu model doirasida ifodalanishi mumkin. Ammo KEG asosidagi sintez modelida ayrim cheklovlar ham mavjud. Ular orasida semantik nazoratning cheklanganligi, stilistik moslikni saqlab qolishdagi qiyinchiliklar va morfologik variantlarning ko‘pligi tufayli qoidalar sonining ko‘payib ketishi kabi holatlar mavjud. Biroq zamonaviy NLP tizimlarida KEGni semantik tarmoqlar, lemmatizatorlar va tahlilchilar bilan birgalikda ishlatish orqali bu muammolarni kamaytirish mumkin. Kontekst-erkli grammatika (KEG) modeli zamonaviy tabiiy tilni qayta ishlash (NLP) texnologiyalarida muhim o‘rin tutadi. U tilning sintaktik va morfologik tuzilishini qat’iy qoidalarga asoslangan holda modellashtirish, tahlil qilish va sintez qilish imkoniyatini beradi. KEG ayniqsa grammatik jihatdan murakkab va agglutinativ tillar, jumladan o‘zbek tili uchun ham nihoyatda foydali model hisoblanadi. Ushbu tadqiqotda KEG modeli asosida o‘zbek tilining bazis so‘zlariga tayangan holda sintez mexanizmini shakllantirish masalasi atroflicha o‘rganildi.

Tahlillar shuni ko‘rsatdiki, o‘zbek tilining so‘z yasovchanligi, affikslar tizimi va erkin so‘z tartibi KEG modelini qo‘llashda alohida yondashuvni talab etadi. Shu sababli, bazis (eng ko‘p ishlatiladigan) so‘zlarni aniqlash, ularning grammatik variantlarini sistemalashtirish va KEG qoidalari orqali sintez algoritmini ishlab

chiqish zarur bo‘ldi. Tadqiqotda bu bosqichlar izchil tarzda amalga oshirildi: no-terminal va terminal belgilar aniqlandi, produktsiya qoidalari ishlab chiqildi va real gaplar sintez qilish imkoniyati sinovdan o‘tkazildi.

KEG asosidagi sintez mexanizmi orqali o‘zbek tilida grammatik jihatdan to‘g‘ri, kontekstga mos va mantiqiy gaplar yaratish mumkinligi amalda isbotlandi. Bunday yondashuv mashina tarjimasi, avtomatik matn sintezi, chatbotlar va ovozli yordamchilar kabi zamonaviy texnologiyalar uchun poydevor vazifasini bajaradi. Shuningdek, tilni o‘rgatish, grammatik strukturani tushuntirish kabi ta’limiy maqsadlar uchun ham KEG modeli asosidagi yondashuvlar samarali qo‘llanishi mumkin. Shuni alohida ta’kidlash lozimki, KEG modeli o‘zbek tilining formal qoidalarini chuqur o‘rganish, sintaktik va morfologik birliklarni tizimli shaklda tasniflash, ularni algoritmik tarzda tahlil qilish va avtomatik tarzda gap yaratish jarayonida ilmiy asos bo‘lib xizmat qiladi. Kelgusida bu yondashuvni semantik tahlil tizimlari, neyron tarmoqlarga asoslangan modellar, korpus lingvistikasi va leksikografik tadqiqotlar bilan integratsiyalash orqali yanada mukammallashtirish mumkin.

Foydalanilgan adabiyotlar:

1. N. Chomsky, Syntactic Structures, 2nd ed. Berlin: Mouton de Gruyter, 2002.
2. D. Jurafsky and J. H. Martin, Speech and Language Processing, 3rd ed. Draft, Stanford University, 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/
3. A. V. Aho and J. D. Ullman, The Theory of Parsing, Translation, and Compiling: Volume I: Parsing. Englewood Cliffs, NJ: Prentice-Hall, 1972.
4. K. Knight, “Automated Knowledge Acquisition for Machine Translation,” Machine Translation, vol. 5, no. 2, pp. 129–144, 1990.
5. D. Klein and C. D. Manning, “Accurate unlexicalized parsing,” in Proc. 41st Annual Meeting of the Association for Computational Linguistics, Sapporo, Japan, 2003, pp. 423–430.

GAPDAGI BOSHQARUV, MOSLASHUV MUNOSABATLARINING XUSUSIYATLARI (KELISHIKLAR ORQALI IFODALANADIGAN SINTAKTIK MUNOSABATLAR)

Sayfiddinova Ziynat Abduzohidovna,
ziynatsayfiddinova@gmail.com
ToshDO‘TAU 3-kurs talabasi

Annotatsiya: Tilshunoslikda sintaktik munosabatlar gap tuzilishini tashkil etuvchi asosiy birliklar o‘rtasidagi bog‘lanishlarni belgilaydi. Umuman olganda, sintaktik munosabatlarning uch turi mavjud: boshqaruv, moslashuv va bitishuv.

Abstract: In linguistics, syntactic relations define the connections between the basic units that make up the structure of a sentence. In general, there are three types of syntactic relations: control, coordination, and conjunction.

Аннотация: В лингвистике синтаксические отношения определяют связи между основными единицами, составляющими структуру предложения. В целом существует три типа синтаксических отношений: контроль, координация и согласование.

Kalit so‘zlar: *Boshqaruv, yordamchi so‘z, moslashuv, bitishuv, so‘z birikmasida bog‘lanish usuli.*

Kirish

An’anaviy tilshunoslikda ham “sintaktik munosabat” va “sintaktik aloqa” terminlari farqlanadi. Sintaktik munosabat sintaktik birliklar o‘rtasidagi mazmuniy sintagmatik munosabatni, sintaktik aloqa esa shakliy sintagmatik munosabatning ifoda tomonini bildiradi deb ta’kidlaydi A.Nurmonov. Ma’lum sintaktik shaklning (gap bo‘lagi shaklining) boshqa sintaktik shakl bilan ketma-ket munosabati shakliy sintagmatik munosabat sanaladi.

So‘zlarning grammatik munosabatga kirishuvi bilan hosil bo‘lgan sintaktik konstruktsiya qo‘shilma sanaladi. Bu xilma-xil qo‘shilmalarning eng sodda ko‘rinishi ikki mustaqil so‘zdan tuzilib, uning tarkibidagi komponentlarning o‘zaro bog‘lanishi bir (tobe) elementning ikkinchi (hokim) elementga ergashishi formasida bo‘lavermaydi. So‘zlar nok va olma, anjir va anor, o‘qidi va yozdi shaklida – biri ikkinchisiga tobelanmay, teng holda bog‘lanishi ham mumkin. Bu bog‘lanish modellar shaklida ot va ot, fe’l va fe’l shakllarida so‘z tizmalarini hosil qiladi. Bunda komponentlarning huquqi teng darajada bo‘ladi, hokim-tobelik aloqasi ko‘zga tashlanmaydi.

Sintaktik konstruktsiyaning komponentlari orasidagi bog‘lanish bu komponentlar orqali ifodalangan ma’nolarning (predmet, belgi va ular orasidagi munosabatni ifodalovchi ma’nolarning) bog‘lanishini bildiradi. Bu ma’nolar material dunyoning kishi ongida aks etishidan tug‘iladi.

Har qanday shakliy sintagmatik aloqa ma’lum mazmuniy sintagmatik munosabatni ifodalaydi va bu ikki munosabat birligi shakl va mazmun dialektikasini aks ettiradi. Masalan, ega va kesimning shakliy munosabatidan tashkil topgan men keldim sintaktik qurilmasi subyekt va predikatning mazmuniy munosabatini ifoda qiladi. An’anaga ko‘ra, ega bilan kesim gapning struktur asosini tashkil qiladi. Ega va kesim munosabati minimal gap deb ham yuritiladi. qolgan bo‘laklar esa shu bo‘laklardan biriga bog‘lanib, gapni yoyiq tusga kiritadi. Shunga muvofiq sintaktik munosabat ierarxiyasini ikkita sathga ajratish ham mumkin:

1. Predikativ sath (ega va kesim munosabati).

2. Nopredikativ aloqalar sathi.

Moslashuv. Bunda bog‘langan har ikki so‘z bir-biriga muvofiqlashib, tobe so‘z shu aloqaga mos hokim so‘z formada keladi; ulardan birining o‘zgarishi bilan boshqasi ham o‘zgarib, shunga mos formaga o‘tadi – o‘z shakllanishini unga barobarlashtiradi. Masalan, terimchining hikoyasi.

Qaratqich bilan qaralmish moslashuv yo‘li bilan bog‘lanadi; shaxs va sonda moslashadi. Bunday tobelanishning asosiy xususiyati tobelanish bir qismning boshqasini aniqlashga xizmat qildirishidir. Tobe so‘z bilan hokim so‘z bir-biriga moslashadi, harakat, belgi va predmetning qaysi predmetga tegishliligi ko‘rsatiladi. Masalan, mening vatanim, gullarning guli, talabalarning o‘qishi. Gullarning guli - Nargis. Bunda guli (hokim) so‘z talaba bilan gullarning (tobe so‘z) qaratqich kelishik qo‘shimchalarini olib kelgan. Dutorimning torlari qo‘sh bulbuldek sayraydi kabi. Moslashuv hodisasi so‘z birikmasining doirasi bilangina cheklanmaydi. Ega va kesim moslashuv yo‘li bilan bog‘lanadi.

Bunday birikma “qaratuvchi va qaralmish” qolipida tuziladi; birinchi a’zo ot, ot tur yoki otlashib kelgan leksema bilan ifodalanadi, qaratqich kelishigida bo‘ladi. Shunday leksema shakl sintaktik jihatdan ikkinchi a’zoga bog‘lanadi. Ikkinchi a’zoga birinchi a’zo bilan sintaktik bog‘lanish tufayli nisbatlovchi qo‘shiladi; nisbatlovchidan oldin sonlovchi qatnashadi, birlik yoki ko‘plik sonlovchisi obyektiv voqelikka ko‘ra tanlab qo‘shiladi; turlovchi esa birikmaning boshqa bir leksema yoki birikma bilan sintaktik bog‘lanishi natijasida qo‘shiladi, shunga ko‘ra ikkinchi a’zo (birikma yaxlitligicha) asli leksema (birikma) holatida bo‘ladi, leksema shaklga (birikma shaklga) keyingi sintaktik bog‘lanish natijasida aylanadi: hovlining o‘rtasi , cholning nabirasi- kabi. a’zo munosabati jihatidan qaralmishdan anglashiladigan predmet yoki belgi qaratuvchidan anglashiladigan

predmetga (predmet tasavvuridagi belgiga) tegishli ekani (nisbat berilayotgani) ifodalanadi: o‘rta hovliga, nabira cholga tegishli. Qaratuvchili birikma ifodalaydigan ma'nolar deb xoslik, butunning bo‘lagi, to‘dadan ajratilgan qism, o‘ringa, vaqtga munosabat ma'nolari ta'kidlanadi. Asli bu ma’nolar – birikmaning ma’nolari emas, balki qaratuvchi, qaralmish bo‘lib keladigan leksemalardan anglashiladigan ma’nolar. Qaratuvchili birikmada turlovchilar paradigmasining bir vakili-qaratqich kelishigi bilan nisbatlovchilar paradigmasi morfemalari orasidagi bog‘liqlik namoyon bo‘ladi.

Bunda kelishik sintaktik aloqani ta’minlashda qatnashadi va sintaktik vazifa ko‘rsatkichi bo‘ladi; nisbatlovchi esa sintaktik aloqani ta’minlashda qatnashadi, lekin shakl o‘zgartiruvchi ekanidan qat’i nazar, sintaktik vazifaga nisbatan betaraf bo‘ladi. Qaralmishga nisbatlovchi qaratuvchidan anglashiladigan shaxs, son, modal ma’nolarga muqobil holda tanlab qo‘shiladi. Qaralmishga qaysi nisbatlovchi qo‘shilishini qaratuvchining kelishik shakli emas, balki kelishik qo‘shiladigan asos, shu asosdan anglashiladigan shaxs, son, modal ma’no belgilaydi: mutanosiblik asosdan anglashiladigan grammatik ma’nolar bilan morfema (nisbatlovchi) ifodalaydigan ma’nolar orasida voqe bo‘ladi. Qaratuvchi tarkibidagi kelishik shakli va qaralmish tarkibidagi sonlovchi mutanosiblik aloqasida qatnashmaydi.

Mutanosiblik qaratuvchi bilan qaralmishning grammatik shakllarida voqe bo‘lmaganligi sababli bunday birikma a’zolari orasidagi sintaktik aloqa moslashuv emas, balki muvofiqlashuv deyiladi (Rus tilshunosligida ega bilan kesim orasidagi bunday aloqa koordinatsiya deb nomlanadi: lot. co(n) - birgalikda, ordinatio - “tartibli joylashuv”. Bunday sintaktik aloqaga sootvetstviye ham deyiladi). Qaralmishga qaysi nisbatlovchining qo‘shilishi qaratuvchining qanday turkum leksemasi bilan ifodalanishiga, qisman uning grammatik shakliga bog‘liq: 1. Qaratuvchi kishilik olmoshi bilan ifodalanganida qaralmishga shu olmoshga mutanosib nisbatlovchi qo‘shiladi: mening uy + im-, mening uylar +im-, sening uy + ing-, sening uylar + ing-, uning uy + i-, uning uylar + i-, bizning uy + imiz-, bizning uylar + imiz- kabi.

Moslashuv (tilshunoslikda) — tobe (bog‘liq) aloqaning bir turi bo‘lib, bunda birikmaning tobe qismi o‘z shaklini xokim qismning shakliga moslaydi, shu bilan hokim qismga bog‘lanadi. Hokim qismning shakli o‘zgarishi bilan tobe qism ham ma’no jihatdan shu shaklga mos bo‘lgan shaklni oladi. O‘zbek tilida M. aloqasi, asosan, qaratqichqaralmish va ega-kesim aloqasida ko‘rinadi va bu moslik shaxs va son jihatidan bo‘ladi, ya’ni qaratqich yoki ega qaysi shaxssonni bildirsa, qaralmish yoki kesim xuddi shu shaxssonni bildiruvchi shaklda bo‘ladi: men yozyapman, sen yozyapsan, u yozyapti, biz yozyapmiz, siz yozyapsiz, ular yozishyapti. Mening ukam, sening ukang, uning ukasi.

Hokim qism (qaratqich yoki ega) 3-shaxsni bildiruvchi otlardan ifodalanganda, hokim va tobe qism sonda moslashishi ham, moslashmasligi ham, ya’ni hokim qism ko‘plik son, tobe qism birlik son shaklida bo‘lishi mumkin: sozandalar keldilar (kelishdi) yoki sozandalar kel-d i. Ega 3-shaxsdagi shaxs bildirmaydigan so‘zlardan bo‘lganida, ega ko‘plik, kesim esa birlik shaklida qo‘llanadi, ya’ni son jixatdan moslashish bo‘lmaydi: gullar ochildi.

Aloqa-munosabat (sintaktik) shakllari (sintaktik shakl hosil qiluvchi vosita) – kelishik, egalik, kesimlik ko‘rsatkichlari. Aloqa-munosabat shakllari shakliy omil sifatida so‘zlarning erkin bog‘lanishida faol: Men qishloqqa o‘tdim gapida *-ga* shakli qishloq so‘zini o‘tmoq fe’li bilan, *-dim* kesimlik shaklidagi *-m* ko‘rsatkichi men so‘zini o‘tmoq fe’liga tobelash vazifasini o‘tamoqda.

Egalik qo‘shimchalari mustaqil so‘zlarning hokim mavqeda kelishini ifodalab, uning tobe so‘zga munosabatini ko‘rsatadi (sizning uyingiz). Shuningdek, bog‘lash bilan birgalikda, hokim va tobe so‘zning shaxs va sonda moslashishini ham ta’minlaydi.

Kelishik – mustaqil so‘zning barchasi uchun tobelikni ta’minlovchi formal ko‘rsatkich. *Masalan,* bosh kelishik egani, tushum kelishigi to‘ldiruvchini, qaratqich kelishigi aniqlovchini hokim bo‘lakka tobelaydi.

Shaxs-son qo‘shimchalari, kesimlikning boshqa shakllari bilan birgalikda, o‘zi hosil qilayotgan hokim bo‘lakni tobe bo‘lak – egaga bog‘laydi.

Yordamchi so‘z. Sintaktik aloqani ifodalashda qo‘llanadigan yordamchi vositalar (ko‘makchi, bog‘lovchi va bog‘lama) birikuvlarda shakliy omil sifatida namoyon bo‘ladi.

Ko‘makchining bir so‘zni ikkinchi so‘zga tobelovchi vosita ekanligi, ma’no va vazifa jihatdan kelishikka o‘xshashligi (*telefonda gaplashmoq – telefon orqali gaplashmoq*), shuningdek, so‘zni sintaktik aloqaga kiritishda o‘zi bilan birga uning biror kelishik shaklida bo‘lishini ham talab qilishi ma’lum: Marraga qadar bir tekisda yugurib borishdi. (A.G‘ul.).

Bog‘lovchi odatda, so‘z tizmalari, gaplarning a’zolarini bog‘lash uchun xizmat qiladi. Masalan: olma va anor. Men o‘qidim va yozdim. Teng bog‘lovchi birinchi misolda so‘z tizmasi a’zolarini, ikkinchi misolda esa qo‘shma gap a’zolarini bog‘lash uchun xizmat qilgan.

Bog‘lama ot kesimni ega yoki boshqa tobe a’zo bilan bog‘lash uchun xizmat qiladi: *Biz o‘tgan yili talaba bo‘ldik.* Bunda bo‘ldik bog‘lamasi shaxs-son va zamon ma’nolarini ifodalagan hamda kesimni uning shaxs (biz), zamon (o‘tgan yili) kengaytiruvchilari bilan bog‘lagan.

So‘z tartibi. So‘zlar erkin birikuvda grammatik shakl asosida bog‘langanda, so‘z tartibi o‘zgarsa-da, u ma’noviy munosabat yoki bog‘lanishga ta’sir qilmaydi: *Sen kitobni keltirdingmi? Kitobni sen keltirdingmi? Keltirdingmi sen kitobni?* birikuvidagi so‘z o‘rni almashtirilganiga qaramay, keltirdingmi so‘zshakli hokim, sen va kitobni so‘zshakllari tobe a’zo sifatida qolavergan. Chunki ular sintaktik mavqeini ta’minlovchi vosita – morfologik ko‘rsatkichga ega. So‘zning hokim-tobeligi (sintaktik mavqei)ni ko‘rsatuvchi vosita bo‘lmaganda so‘z tartibining qimmati oshadi *(achchiq olma)*. Bunda so‘zning gapdagi funksiyasi, asosan, tartibi bilan belgilanadi, tartib o‘zgarsa, mazmun bilan birga, sintaktik holat ham o‘zgaradi: Hayot go‘zal *(ega+kesim),* go‘zal hayot *(aniqlovchi+aniqlanmish).* Gapda uslubiy maqsadlarda so‘z tartibi o‘zgarishi mumkin: O‘qituvchi intizom haqida gapirdi. Intizom haqida o‘qituvchi gapirdi. Tartibning bunday o‘zgarishi a’zolarning hokim-tobeligiga ta’sir qilmaydi, biroq gapning uslubiy jihatiga, fikrdagi ma’no nozikligiga ta’sir ko‘rsatadi.

Ohang ham, yuqorida aytilganidek, birikuvning tabiatini ko‘rsatuvchi nisbiy vosita. Gap tugallangan ohang bilan aytiladi, uning tarkibidagi ayrim birikuvlar sanash (olma va anor), ayrimlari ko‘tariluvchi ohangga (qizil qalam) ega bo‘ladi.

Ohang (pauza) birikuvning sintaktik bo‘linishini ko‘rsatadi. Masalan, Ko‘p qavatli uy birikmasini ikki xil (ko‘p//qavatli uy va ko‘p qavatli//uy tarzida) aytish mumkin. Demak, sintaktik aloqani, unga bog‘liq ravishda mazmuniy farqlanishni ifodalashda ohangning o‘rni bor. Ya’ni ko‘p//qavatli uy tarzida birinchi so‘zdan keyin pauza qilinishi natijasida *(ko‘p)* so‘zshakli tobe va qavatli uy birikmasi unga hokim mavqeda bo‘lib, bunda qavatli uylarning ko‘pligi haqidagi ma’no anglashiladi. Ikkinchi holatda (ko‘p qavatli//uy) esa ko‘p qavatli birikmasi tobe, uy so‘zshakli esa hokim mavqeda, birikmadan esa uyning qavatlari ko‘pligi ma’nosi namoyon bo‘lgan.

Sanash ohangi bir xil bo‘lakning aloqasini ifodalaydi: Oltin, kumush tosh bo‘lar. (Maq.)

Qo‘shma gapning tarkibiy qismlari orasida ham turli: qarama-qarshilik *(Kunduzi issiq – kechasi sovuq)*, shart *(Sen kel – u boradi)*, sabab *(Kecha borolmadim – mehmonlar kelib qoldi)* – ma’noviy munosabatlar mavjud.

Ko‘rinadiki, ayrim o‘rinda so‘zlar va gaplarni bog‘lashni ohangning o‘zi bajaradi.

So‘z birikmasida bog‘lanish usuli

So‘z birikmasi a’zolari orasidagi bog‘lanish, ya’ni tobelanish uch xil: bitishuv, moslashuv, boshqaruv. Bu aloqa turining o‘ziga xos xususiyati va farqi a’zolarni biriktiruvchi ko‘rsatkich yoki boshqa vositadan kelib chiqadi. Masalan, tez yurmoq, qalamni olmoq, kitobning varag‘i birikmalarida tobe so‘zning qaysi

shaklda kelishi ko‘pincha, uning hokim so‘ziga *(qalamni olmoq)* bog‘liq bo‘lsa, ba’zan har ikkala a’zo bir-birining qanday shaklda bo‘lishini belgilab qo‘yadi *(kitobning varag‘i)*.

Bitishuvda birikkan so‘zlarning o‘zaro aloqasi, asosan, tartib va ohang bilan belgilanadi. Bitishuv aloqasida morfologik ko‘rsatkich bo‘lmasa-da, tartib muhim vosita sifatida namoyon bo‘ladi. Bunda tobe so‘z ushbu mavqeda kelish uchun shakliy jihatdan o‘zgarmaydi. Zero, tobe a’zo sifatida namoyon bo‘layotgan so‘z ma’noviy va vazifaviy jihatdan bunga xoslangan bo‘ladi. Misollar: qizil gul, oq qog‘oz. Ba’zan ot turkumidagi so‘zlar ham hech qanday vositasiz tobe vazifada kela oladi: oltin soat, tosh yo‘l kabi. Bunda ular mohiyatan tobe a’zolik doimiy belgisi bo‘lgan sifat, son, ravish turkumiga yaqinlashadi. Bu hodisa nutqiy bo‘lib, oltin, tosh leksemasining lisoniy mohiyati bilan belgilanmagan.

Quyidagi so‘zlar bitishuv yo‘li bilan bog‘langan:

1) ravish +fe’l: o‘z-o‘zidan gapirmoq;

2) sifat+fe’l: yaxshi so‘zlamoq;

3) son+ot: o‘nta qalam, sakkiz daftar;

4) ot+ot: kumush qoshiq, chang yo‘l;

5) olmosh+ot: hamma odam, qanday kitob.

Ayrim manbalarda tobe so‘z vazifasida ravishdosh (shoshilib gapirdi), sifatdosh (o‘qigan bola) yuzaga chiqqan birikuv ham bitishuvli birikma sifatida qaraladi. Vaholanki, bu erda ravishdosh ko‘rsatkichi *-ib* fe’lni fe’lga (shoshilib gapirdi), sifatdosh shakli *-gan* fe’lni otga (o‘qigan bola) bog‘lash uchun xizmat qiladi. Shuningdek, bularning morfologiyada lug‘aviy-sintaktik shakl sifatida qaralishi ham ular hosil qilgan birikuvni bitishuvli emas, balki boshqaruvli birikuv sifatida baholashni taqozo etadi.

Bitishuvda a’zolarning hokim-tobelik holati, asosan, tartib va joylashuv omili asosida belgilanadi: (tobe a’zo ya’ni tobe so‘z oldin, hokim so‘z keyin keladi hamda a’zolar orasiga boshqa so‘zni kiritib bo‘lmaydi, *(kumush qoshiq birikmasini kumush o‘nta qoshiq deb o‘zgartirib bo‘lmaydi)*. Ba’zan bitishuvda ham tartib ikkinchi o‘ringa tushib qolgandek tuyuladi: Majlis bo‘ladi – Majlis kechqurun bo‘ladi. Lekin bunda majlis so‘zi fe’lga bitishuv asosida emas, balki bosh kelishikda bog‘lanayotganligini esdan chiqarmaslik lozim.

Demak, bitishuv aloqasida tartib va ohang muhim rol o‘ynaydi. Bunda tobelik tartib va joylashuvdan, so‘zning grammatik xususiyati va ma’nosidan anglashilishi ayon bo‘lib turadi.

Kelishik affiksi o‘z ma’nolarini yo‘qotgan tezda kelmoq, o‘z-o‘zidan raqsga tushib ketmoq, birdan gapirmoq birikuvini ham bitishuv aloqasi sifatida qaramoq

lozim. Chunki kelishik qo‘shimchasi bu SBda o‘z tizimidan ajralib, so‘z tarkibida *«qotib qolgan».* Bular – kelishik qo‘shimchasini olgan holda ravishga ko‘chgan so‘z.

Boshqaruv aloqasi bir tomonlama shakllangan SBlar tarkibida amal qiladi. Faqat tobe a’zolari birikish uchun tobelovchi qo‘shimcha olgan SB boshqaruv yo‘li bilan bog‘langan deyiladi. Bunda hokim so‘z tobe so‘zning qanday shaklda bo‘lishini boshqaradi. Masalan, kitobni o‘qimoq birikmasida tobe so‘zning tushum kelishigi shaklida kelishi o‘qimoq fe’lining o‘timliligi bilan belgilangan. Agar bu fe’l o‘timsiz fe’lga aylantirilsa, (o‘qildi) tobe so‘zdagi tushum kelishigi shakli o‘z-o‘zidan g‘oyib bo‘ladi: kitobni o‘qimoq-kitob o‘qildi[1].

Boshqaruv aloqasida tobe so‘zni har xil turkumdagi so‘z boshqarishi mumkin. Misollar: uyga bormoq *(fe’lli boshqaruv)*, mendan katta *(sifat boshqaruvi)*, tovushdan tez *(ravishli boshqaruv)*, o‘qishda birinchi *(son boshqaruvi)*, aytgan bola *(otli boshqaruv)*. ya’ni tobe so‘z oldin, hokim so‘z keyin keladi hamda a’zolar orasiga boshqa so‘zni kiritib bo‘lmaydi, *(kumush qoshiq birikmasini kumush o‘nta qoshiq deb o‘zgartirib bo‘lmaydi)*. Ba’zan bitishuvda ham tartib ikkinchi o‘ringa tushib qolgandek tuyuladi: Majlis bo‘ladi – Majlis kechqurun bo‘ladi. Lekin bunda majlis so‘zi fe’lga bitishuv asosida emas, balki bosh kelishikda bog‘lanayotganligini esdan chiqarmaslik lozim.

Demak, bitishuv aloqasida tartib va ohang muhim rol o‘ynaydi. Bunda tobelik tartib va joylashuvdan, so‘zning grammatik xususiyati va ma’nosidan anglashilishi ayon bo‘lib turadi.

Kelishik affiksi o‘z ma’nolarini yo‘qotgan tezda kelmoq, o‘z-o‘zidan raqsga tushib ketmoq, birdan gapirmoq birikuvini ham bitishuv aloqasi sifatida qaramoq lozim. Chunki kelishik qo‘shimchasi bu SBda o‘z tizimidan ajralib, so‘z tarkibida «qotib qolgan». Bular – kelishik qo‘shimchasini olgan holda ravishga ko‘chgan so‘z.

Valentlik tushunchasi

Hozirgi tilshunoslikda gap bo‘laklari iyerarxiyasida faqat kesimning eng muhim bo‘lak, gapning struktur asosi deb qaralishi gapni tashkil etuvchi sintaktik elementlar o‘rtasidagi sintaktik aloqalar xarakteriga ham ta’sir qiladi. Har qanday gap predikat va uning argumentlari munosabati birligi sifatida qaralar ekan, predikat va uning argumentlari o‘rtasidagi sintaktik munosabatlar tushunchasi hozirgi tilshunoslikda morfologik munosabatlar tiplarining boshqaruv, moslashuv, bitishuv kabi tahlili orqali emas, balki valentlik tushunchasi orqali belgilanadi. Sintaktik aloqa reallashgan valentlik sifatida qaraladi. Keyingi yillarda valentlik nazariyasi o‘z muammosi jihatidan tilshunoslikda alohida soha sifatida ajrala boshladi. Bu termin aslida 1948-yilda S.D.Kaynelson tomonidan qo‘llanilgan edi.

U so‘zning gapda ma’lum tarzda reallashishi va boshqa so‘z bilan ma’lum kombinatsiyaga kirishishi xususiyatini uning sintaktik valentligi hisoblaydi. S.D.Kaynelson bilan bir vaqtda, ammo o‘zi mustaqil holda fransuz tilshunosi L.Tener ham 50-yillarda tobelik grammatikasi doirasida valentlik nazariyasini ishlab chiqdi. Uning bu nazariyasi jahon tilshunosligiga keng tarqaldi, yangi-yangi tadqiqotlar vujudga kelishiga turtki bo‘ldi. Tener o‘z grammatikasida gapni struktur jihatdan tahlil qilish jarayonida uning struktur asosi sifatida fe’lga tayandi.

Fe’l valentligi nazariyasiga ko‘ra gap strukturasini predikat (ko‘p hollarda fe’ldan ifodalangan) belgilaydi va u o‘ziga muvofiq kategorial belgilarga o‘rin ochadi. Gapning barcha komponentlari fe’l-predikat valentligi munosabatiga ko‘ra u bilan bog‘liq argument, aktant, agens, lokalis, kengaytiruvchilar kabi terminlar bilan bir-biridan farqlandi.

Valentlik nazariyasining mukammallashishida nemis tilshunosi G.Xelbig katta rol o‘ynadi. U nemis tili materiallari asosida fe’lning valentlik modellarini ishlab chiqdi va *“Nemis tilidagi fe’llar valentliklari va distributsiyasi lugati”* ni tuzdi.

Dastlab valentlik nazariyasi valentlik tabiatini turlicha hal qiladigan ikki yo‘nalish – forma va semantik yo‘nalishga ajraldi. Keyingi davrlarda bu ikki yo‘nalish orasida o‘zaro yaqinlashish, so‘z shakl va so‘z leksik ma’nosi valentliklari o‘rtasidagi munosabatning mavjudligiga e’tibor berilmoqda.

Hozirda valentlikning asosan ikki turi sintaktik (konfiguratsion) va semantik turlari ko‘p tilshunoslarga yaxshi tanish. Sintaktik valentlik – ma’lum mustaqil so‘zning boshqa mustaqil so‘z bilan birika olish imkoniyati va semantika bilan ham bog‘liq. Semantik valentlik leksemalar o‘rtasidagi turli sintagmatik munosabatlarni ko‘rsatadi. Masalan, o‘zbek tilida *“sayramoq”* leksemasi qushlarning ovoz chiqarishi ma’nosini ifodalaydi. Shuning uchun u faqat kushlar so‘zi bilan bog‘lana oladi. Ammo u *“ko‘p gapirish”* ma’nosi bilan “odam” sinfiga kiruvchilar bilan ham birikadi. Bu sayramoq leksemasining semantik valentligidir.

Sintaktik valentlik leksema valentligining ikkinchi tomoni, u hokim mavqedagi leksemaning tobe leksemalarni o‘ziga tortish uchun ularning ma’lum sintaktik shaklda – kelishik, ko‘makchi, ravishdosh, sifatdosh qo‘shimchasini olishini talab qilishi. Bu hokim leksemaning sintaktik mavqei, grammatik shakli tomonidan belgilanadi. Masalan, Xatni yozdi, Xat yozildi sintaktik qurilmalaridagi (xatni) va (xat) so‘zlarining grammatik shakli bosh so‘zning grammatik shakli bilan belgilangan[2].

Sintaktik valentlik lug‘aviy valentlik bilan zich bog‘langan, biri o‘zgarsa, ikkinchisi ham o‘zgaradi, biri yo‘qolsa, ikkinchisi ham bo‘lmaydi. Masalan, yetakla leksemasi jo‘nalish kelishigidagi otni o‘ziga tortmaydi (Halimga yetakla tarzida). Lekin yetakla leksemasi orttirma nisbat shakli *-t* bilan shakllansa, (Halimga)

so‘zshakli unga tortila oladi. Bunda yetakla leksemasining sintaktik valentligi o‘zgardi.

Lug‘aviy va sintaktik valentlik o‘zaro bog‘liq bo‘lsa-da, ular tez-tez o‘zgarishga uchrab turadi. Masalan, Ashulachi *“hangradi”* birikuvida sintaktik valentlik o‘zgarmasdan voqelangan, ammo lug‘aviy valentlik yangilangan holda yuzaga chiqqan. Bunday hol ko‘chma ma’no yuzaga chiqqanidan dalolat beradi. Xat yozildi, Xatni yozdi birikuvlarida ma’noviy valentlik joyida, biroq sintaktik valentlik o‘zgargan. Bu turlicha sintaktik shakllar voqelanganligini ko‘rsatadi. Biroq aytilgan har ikkala o‘zgarish ham bir yoqlama – yo ma’noviy, yo sintaktik tabiatga ega. Agar o‘zgarish birdaniga ikki tomonlama – ham lug‘aviy, ham sintaktik bo‘lsa-chi? Bunda derivatsiya sodir bo‘ladi. Dalil sifatida ochmoq fe’lining lug‘aviy va sintaktik valentliklarini aniqlash orqali bunga amin bo‘lish mumkin:

1. Tergovchi jinoyatni ochdi.

2. Jinoyat tergovchi tomonidan ochildi.

3. Amerikani Vespuchchi ochdi.

4. Gul ochildi.

1- va 2-misolda ochmoq fe’li valentliklarini to‘ldiruvchi so‘z (tergovchi) ning ma’nosi o‘zgarmagan, lekin sintaktik shakli (tergovchi – bosh kelishikda, tergovchi tomonidan – ko‘makchili) o‘zgargan. Demak, ochmoq leksemasining ma’noviy valentligi o‘zgarishsiz saqlangan, sintaktik valentligi o‘zgargan. Bu o‘zgarish ochmoq fe’liga *-il* majhul nisbat qo‘shimchasi tomonidan kiritilgan. 3-misolda esa fe’lning sintaktik valentligi o‘zgarishsiz qolgani holda ma’noviy valentlik o‘zgarishga uchragan. Chunki Amerika so‘zining ma’nosi ochmoq fe’li ma’nosiga muvofiq kelmaydi. Ammo sintaktik shakli mos. 4-misolda esa ham ma’noviy, ham sintaktik valentliklar butkul o‘zgargan. Chunki gul so‘zi ochmoq fe’liga ma’noviy jihatdan mos emas. Bundan tashqari, birikuvchi so‘zning sintaktik shakli (gul – bosh kelishik) ochmoq fe’lga *-il* qo‘shimchasi qo‘shilishi natijasida o‘zgarishga uchragan.

Demak, ma’lum bo‘ladiki, leksemaning lisoniy valentligi ikkiyoqlama, ma’noviy tomoni unga qanday leksema birika olishi me’yori, chegarasini ko‘rsatsa, sintaktik tomoni birikuvchi so‘zning qanday grammatik qo‘shimchani olishini belgilaydi.

Xulosa

Ma’lumki, har bir milliy til o‘z taraqqiyotida mukammallashib boradi: uning grammatik va fonetik normalari barqarorlashadi, lug’at tarkibi esa boyib boradi.

Shuning natijasida til eng murakkab fikrlarni chiroyli va har xil formada ifodalash xususiyatiga ega bo‘ladi.

O‘zbek tilida teng va tobe bog‘lanish, shuningdek bog’lovchi vositalar an’anaviy tilshunoslikda shakliy-funksional tomondan ancha o‘rganilgan. Bu sohada ko‘pgina maqolalar, monografiya va qo‘llanmalar nashr qilingan. Ammo ulardagi barcha holatlar aniq, ular oxirigacha o‘z yechimini topgan deyish qiyin. Biz kichik tadqiqotimizda sintaktik aloqa turlari haqidagi nazariy qarashlarni umumlashtirish, munosabat bildirishni ma’qul deb bildik.

Aslida ba’zi mustaqil, so‘zlar tilning taraqqiyoti va tarixiy takomillanishi protsessida so‘zlar birikmasi tarkibida qo‘shni so‘zlarga qo‘shilib, enkliza holatida qo‘shni so‘zga tobelanib, ularga biror yo‘sinda singib ketib, shaklan va mazmunan turlicha o‘zgarishlarga uchrab, natijada, mustaqil shaklini, muayyan ma’nosini yo‘qotib o‘zi qo‘shilgan so‘zlar uchun bir affiks vazifasini bajara boshlaydi va shu holatda affiksga aylanadi, degan turlicha ilmiy nazariya’ning haqiqat ekanini tasdiqlovchi bir qancha asosli dalillar mavjud.

Prezdidentimiz aytganidek, “O‘zlikni anglash, milliy ong va tafakkurning ifodasi, avlodlar o‘rtasidagi ruhiy-manaviy bog‘liqlik til orqali namoyon bo‘ladi, ona tili bu millatning ruhidir”.

Tilimizdagi so‘zlar va muayyan grammatik format (affikslarning har biri til va tafakkurning zo‘r ehtiyojlari, hayotning keskin talablari bilan til taraqqiyoti protsessida biror konkret so‘z shaklidan paydo bo‘lgan: -ni, -ning, -ga, -da, -dan kabi kelishik affikslari aslida azaliy emas, ularning har birining o‘ziga xos muayyan tarixi va konkret rivojlanish qonuni borki, ilmiy analiz yo‘li bilan har bir so‘z va har bir affiksning paydo bo‘lish tarixi, rivojlanish qonunlari, taraqqiyot yo‘llarining ochilishi, aniqlanishi mumkin[3].

Aslida ba’zi mustaqil, so‘zlar tilning taraqqiyoti va tarixiy takomillanishi protsessida so‘zlar birikmasi tarkibida qo‘shni so‘zlarga qo‘shilib, enkliza holatida qo‘shni so‘zga tobelanib, ularga biror yo‘sinda singib ketib, shaklan va mazmunan turlicha o‘zgarishlarga uchrab, natijada, mustaqil shaklini, muayyan ma’nosini yo‘qotib o‘zi qo‘shilgan so‘zlar uchun bir affiks vazifasini bajara boshlaydi va shu holatda affiksga aylanadi, degan turlicha ilmiy nazariya’ning haqiqat ekanini tasdiqlovchi bir qancha asosli dalillar mavjud.

Mustaqillik yillarida o‘zbek tilshunosligi jahon tilshunosligining ilg‘or g‘oya va metodlarini o‘zlashtirgan holda uni yangi ilmiy dunyoqarashlar, zamonaga munosib talqinlar, fanning ilg‘or nazariyalari, manbalari bilan boyitishga harakat qilmoqda. Shuning natijasi sifatida A.Nurmonov va N.Mahmudovlar “O‘zbek tili nazariy grammatikasi”ni nashr qildilar. Bu ham ilgarigi sintaktik qarashlarga yangi izlanishlar samarasi sifatida katta hissa bo‘lib qo‘shildi va bu sohani jahon tilshunosligidagi chuqur sintaktik ma’lumotlar materiallar bilan to‘ldirdi.

O‘zbek tilshunosligida so‘nggi yillarda olib borilayotgan tadqiqotlarda sintaktik aloqalar haqidagi nazariy qarashlar, munosabatlar ham o‘zgarayotganligini aytib o‘tish o‘rinlidir. Ayniqsa kesimning asosiy markaziy bo‘lak ekanligi, ismlar va kelishiklar.

Kelishikli munosabatlar, ya’ni moslashuv va boshqaruv munosabatlarining aniq qoliplari ishlab chiqilishi til korpusi uchun juda katta nazariy va amaliy ahamiyatga ega. Quyida bu ahamiyatlar bir necha asosiy yo‘nalishda bayon qilinadi:

1. Sintaktik tahlil avtomatlashtiriladi. Agar boshqaruv va moslashuv munosabatlarining qoidalari va qoliplari aniq belgilansa:

- Til korpusidagi har bir gap avtomatik ravishda pars qilinadi (ya’ni, grammatik tuzilmasi aniqlanadi).
- Gapdagi bog‘lanishlar (so‘zlar orasidagi sintaktik munosabatlar) kompyuter tomonidan aniqlanadi. Bu korpusni sintaktik teglash (annotatsiya qilish) imkonini beradi. *Misol: “Kitobni o‘qidim” gapida ot – fe’l munosabati boshqaruv orqali aniqlanadi (o‘qimoq – -ni kelishigi). Qolip: ot + -ni + fe’l = Tushum kelishigi vositasida boshqaruv.*

2. Mashinali tarjima sifatini oshiradi. Mashinali tarjima tizimlari uchun:

Boshqaruv va moslashuv munosabatlarini tushunish — so‘zlar qanday bog‘lanishini bilish demak. Bu orqali so‘z tartibi, ot-sifat, fe’l-to‘ldiruvchi kabi birliklar tarjimada to‘g‘ri aks ettiriladi.

3. Tilshunoslik tadqiqotlari uchun asos.

- Korpus asosida statistik tahlil qilish mumkin bo‘ladi: qaysi boshqaruv/moslashuv turlari ko‘proq uchraydi, qanday holatda ishlatiladi.
- Yangi til modellari (masalan, o‘zbek tili uchun LLMlar) qurishda bu strukturaviy axborotdan foydalanish mumkin.

4. Til o‘rgatishda foydali vosita. Ayni qoliplar asosida interaktiv mashqlar, sintaktik konstruktorlar yaratish mumkin. Bu talabalarga tilning ichki tizimini tushunishda yordam beradi.

5. Tilni rasmiylashtirish (formalizatsiya) imkoniyati. Tildagi boshqaruv va moslashuv munosabatlari aniq formulalar ko‘rinishida ifodalansa, ular kompyuterga tushunarli bo‘ladi. Bu esa grammatik qoidalarni algoritmik tarzda yozishga imkon beradi.

Foydalanilgan adabiyotlar:

1. A. G‘ulomov, M. Asqarova. Hozirgi o‘zbek adabiy tili. Sintaksis. Toshkent. O‘qituvchi. 1987.

2. G. Abdurahmonov, A. Sulaymonov, X. Xoliyorov, J. Omonturdiev. Hozirgi o‘zbek adabiy tili. Sintaksis. Toshkent. O‘qituvchi. 1979.
3. N. Mahmudov, A. Nurmonov. O‘zbek tilining nazariy grammatikasi. Toshkent. O‘qituvchi. 1995.
4. I. A. Karimov. Yuksak manaviyat – yengilmas kuch. – T.,2010, 83-b
5. A. N. Kononov. Grammatika sovremennogo tureykogo literaturnogo yazıka. M. 1956, 78-97-betlar.
6. N.A. Baskakov. Karakalpakskiy yazık, 1 chast 1 izd-vo AN Moskva, 1956, 196-197 betlar. Shu avtor, Osootnosheniyalichnıx i ukazatelnıx mestoimenii v tyurskix yazıkax. Dokladı i soobsheniya AN. 1. 1956, 135-bet.
7. S.E. Malov, Pamyatniki drevnego-tyurskoy pismennosti izd. AN M. 1951, 187-198 betlar.
8. Prof. V.V. Reshetov. Ob odnom uzbekskom padeje, Tyurkologigicheskiy sbornik. 1 izd AN M. 1956 176-bet.
9. G. I. Ramstedt, Vvedenie v altayskoe yazıkoznanie, izd İn. Lit.
10. E. V. Sevortyan, Kategoriya padeja. İssledovaniya po sravnitelnoy grammatike tyurskix yazıkov, N. 1956, 45-57 betlar.
11. A. M. Sherbak, Grammaticheskiy ocherk yazıka tyurskix tekstov. XX-XIII v.v. iz vostochnogo Turkestana, izd AN M. 1961, 78-91 betlar.
12. Prof. A. K. Borokov, İz materialov dlya istorii uzbeskogo yazıka, Tyurkologicheskiy sbornik İzd-vo AN M. 1956 74-78 betlar.
13. A.G. G’ulomov, O‘zbek tilida kelishiklar, O‘zFa asarlari Toshkent, 1941, Filologiya I Ikitob 5-85 bet.
14. Faxri Kamol, O‘zbek orfografiyasi haqida O‘quvpeddavnashr, Tosh. 1957 32-35 bet.
15. Sh. Shoabdurahmanov, O‘zbek adabiy tili va o‘zbek xalq shevalari, Tosh1962. 141g’150 bet.
16. F. Abdullayev, Kelishik affikslarining genezisiga doir. “O‘zbek tili va adabiyoti masalalari” jurnali, 5-son 1961 27-bet.
17. K. Karimov, Kategoriya padeja v yazıke “Kutadgu bilig”, Avtoref. Kand. Diss. Tosh 1962.

II SHO‘BA. KORPUS LINGVISTIKASI

KORPUS MATNLARINI PYTHON TILI VOSITASIDA TEMATIK MODELLASHTIRISH

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Narzillo Aloyev Raxmatilloyevich,
ToshDO‘TAU tayanch doktorant
vip.alayev@gmail.com

Annotatsiya. Tematik modellashtirish – bu hujjatlar to‘plamini tahlil qilish, ulardagi so‘z va so‘z birikmalarni aniqlash hamda hujjatlar to‘plamini eng yaxshi tavsiflovchi so‘z guruhlari va shunga o‘xshash frazalarni avtomatik ravishda klasterlashga asoslangan nazoratsiz mashinali o‘rganish usuli. Har kuni yuzlab, hatto minglab mijozlar bilan muloqot qiladigan kompaniyada, ijtimoiy tarmoqlardagi xabarlar, elektron pochta xabarlari, chatlar, ochiq so‘rovnomalar va boshqa shu turdagi ma'lumotlarini tahlil qilish ancha murakkab jarayon. Sun'iy intellekt vositalari yordamida matn tahlili tilni tabiiy ravishda qayta ishlash uchun turli xil usullar yoki algoritmlardan foydalanadi. Ulardan biri mavzu tahlili - matnlardan mavzularni avtomatik aniqlash uchun ishlatiladi. Mavzuni tahlil qilish modellaridan foydalangan holda, tashkilotdagi katta hajmdagi matnli ma’lumotlarni qayta ishlash vazifalari mashinalarga yuklatilishi mumkin. Agar mashina har kuni mijozlar katta sondagi so‘rovlarini saralab bera olsa, tashkilot xodimlarining qancha vaqt tejashi va muhimroq vazifalarga sarflashi mumkinligini ta`kidlash zarur. Ushbu maqolada mavzuni tahlil qilish usullarining zamonaviy usullari tavsifi, matematik asosi va Python tilidagi tadbig‘i ko‘rib chiqiladi. Tematik modellashtirish – bu “nazoratsiz” mashinali o‘rganish usuli bo‘lib, dataset treningini talab qilmaydi.

Annotation. Topic modeling is an unsupervised machine learning technique based on analyzing a set of documents, identifying words and phrases in them, and automatically clustering groups of words and similar phrases that best describe the set of documents. In a company that communicates with hundreds or even thousands of customers every day, analyzing social media messages, emails, chats, open surveys and other such data is a complex process. Text analysis using artificial intelligence tools uses different techniques or algorithms to naturally process language. One of them is topic analysis, which is used to automatically identify topics from texts. Using topic analysis models, large text data processing tasks in an organization can be outsourced to machines. It's worth noting how much time an organization's employees can save and spend on more important tasks if a machine

can sort through a large number of customer requests every day. This article describes modern methods of topic analysis, mathematical basis and implementation in Python language. Topic modeling is an "unsupervised" machine learning technique that does not require dataset training.

Аннотация. Тематическое моделирование – это метод машинного обучения, основанный на анализе набора документов, выявлении в них слов и фраз и автоматической кластеризации групп слов и похожих фраз, которые лучше всего описывают набор документов. В компании, которая ежедневно общается с сотнями или даже тысячами клиентов, анализ сообщений в социальных сетях, электронных писем, чатов, открытых опросов и других подобных данных является сложным процессом. Анализ текста с использованием инструментов искусственного интеллекта использует различные методы или алгоритмы для естественной обработки языка. Один из них – тематический анализ, который используется для автоматического определения тем из текстов. Используя модели тематического анализа, большие задачи по обработке текстовых данных в организации можно передать машинам. Стоит отметить, сколько времени сотрудники организации могут сэкономить и потратить на более важные задачи, если машина сможет обрабатывать большое количество запросов клиентов каждый день. В этой статье описаны современные методы тематического анализа, их математическая основа и реализация на языке Python. Тематическое моделирование – это метод машинного обучения, который не требует обучения набору данных.

Kalit so‘zlar: *Tematik modellashtirish, tematik modellar, Python, LSA, LSI, LDA, NMF, HDP, nazoratsiz mashinali o‘qitish usullari, til korpusi.*

Kirish

Tematik modellashtirish – bu hujjatlar to‘plami uchun klaster so‘zlarini aniqlash uchun matn ma’lumotlarini avtomatik ravishda tahlil qiladigan mashinani o‘rganish usuli. Bu usul “nazoratsiz” mashinali o‘rganish sifatida tanilgan bo‘lib, avvaldan odamlar tomonidan tasniflangan teglar yoki o‘quv ma'lumotlarining oldindan belgilangan ro‘yxatini talab qilmaydi [1].

Tematik modellashtirish o‘quv ma’lumotlarini talab qilmagani uchun, ushbu usul ma’lumotlarni tahlil qilishni boshlashning tez va oson yo‘lidir. Shu sababli, tashkilotlar o‘z faoliyalari samaradorligini oshirish maqsadida mavzuni tasniflash modelini o‘rganishga vaqt sarflashni afzal ko‘rishadi.

Bugungi kundagi tematik modellashtirishning ommabop algoritmlariga *yashirin semantik tahlil (Latent Semantic Analysis, LSA), yashirin semantik indeksatsiya (Latent Semantic Indexing, LSI), ierarxik Dirixle jarayoni*

(Hierarchical Dirichlet Process, HDP), yashirin Dirixle taqsimoti (Latent Dirichlet Allocation, LDA) va *manfiy bo‘lmagan matritsa faktorizatsiyasi (Non-negative Matrix factorization, NMF)* kabi usullarini misol sifatida keltirish mumkin [2,3,4]. Ular orasida LSA amalda aniqroq va samarali natijalarni ko‘rsatgan va shuning uchun keng miqyosida qo‘llaniladi [5,6]. Ushbu maqolada yuqorida keltirilgan tematik modellashtirish usullarining barchasini birma-bir ko‘rib chiqiladi va Python tili vositalari orqali tadbiq qilinadi.

Yashirin semantik tahlil (Latent Semantic Analysis, LSA) usuli

Yashirin semantik tahlil – bu hujjatlar to‘plamidagi yashirin munosabatlarni kam o‘lchamli fazoda aniqlashga harakat qiladigan matematik usul. LSA ma’nosi yaqin bo‘lgan so‘zlar o‘xshash matn qismlarida *paydo bo‘lishini taxmin (tarqatish gipotezasi)* qiladi.

Har bir paragrafda so‘zlar sonini o‘z ichiga olgan matritsa (satrlar noyob so‘zlarni, ustunlar esa har bir paragrafni ifodalaydi) matnning katta qismidan hosil qilinadi va o‘xshashlik strukturasini saqlab qolgan holda qator(satr)lar sonini kamaytirish uchun *singulyar qiymat dekompozitsiyasi (singular value decomposition, SVD)* deb ataladigan matematik usuldan foydalaniladi [3,7]. LSA usulida har bir hujjatni boshqalardan ajratilgan holda ko‘rib chiqish o‘rniga, munosabatlarni aniqlash uchun barcha hujjatlarni va ulardagi shartlarni ko‘rib chiqiladi.

SVDdan quyi darajali matritsaga yaqinlashish masalasini hal qilish uchun foydalanish mumkin. So‘ngra ushbu matrisadan *termin-hujjat matritsalarini* aniqash mumkin. Buning uchun quyidagi uch bosqichli amallarni bajarish lozim [8,9]:

1. Berilgan **C** uchun **SVD** matrisasini $\boldsymbol{C} = \boldsymbol{U} \sum \boldsymbol{V^T}$ ko‘rinishda hosil qilamiz.
2. $\sum$ diagonalidagi **r−k** eng kichik singulyar qiymatlarni nolga almashtirish natijasida hosil bo‘lgan $\sum \boldsymbol{k}$ matritsasini hosil qilish.
3. $\boldsymbol{C_k} = \boldsymbol{U} \sum \boldsymbol{kV^T}$ni **C** ga **k**-darajada yaqinlashtirish.

Bu yerda, **C** – *termin-hujjat matrisasi,* $\boldsymbol{U}$, $\sum$ va $\boldsymbol{V^T}$ qiymatlar **SVD** matrisalari.

Yuqorida keltirilgan nazariy ma’lumotlar asosida Python tili vositalari yordamida **LSI** uchun **scikit-learning** modulidan foydalanamiz.

Scikit-learning modulida LSA usuli uchun o‘lchamlarni kamaytirish SVD metodi yordamida amalga oshiriladi.

```
from sklearn.decomposition import NMF, LatentDirichletAllocation, TruncatedSVD
from sklearn.feature_extraction.text import CountVectorizer
NUM_TOPICS = 4
```

```
# Hujjatni tokenlar matritsasiga aylantirish
vectorizer = CountVectorizer(min_df=5, max_df=0.9,
                             stop_words='english', lowercase=True,
                             token_pattern='[a-zA-Z\-][a-zA-Z\-]{2,}')
data_vectorized = vectorizer.fit_transform(title)
# SVDdan foydalaniv Latent Semantic Indexing (LSI) modelni shakllantirish
lsi_model = TruncatedSVD(n_components=NUM_TOPICS)
lsi_Z = lsi_model.fit_transform(data_vectorized)
print(lsi_Z.shape)
```

_____

(3000, 4)

```
def print_topics(model, vectorizer, top_n=10):
    for idx, topic in enumerate(model.components_):
        print("Topic %d:" % (idx))
        print([(vectorizer.get_feature_names_out()[i], topic[i])
                for i in topic.argsort()[:-top_n - 1:-1]])

print("LSI Model:")
print_topics(lsi_model, vectorizer)
print("=" * 20)
```

_____

LSI Model:

Topic 0:

[('new', 0.9423293654141083), ('york', 0.09876242707814353), ('samsung‘, 0.08674228078722532), ('study', 0.0842158513506398), ('says', 0.07742874309109753), ('google', 0.07544052179967121), ('apple', 0.06923485166706239), ('galaxy', 0.06400938270176393), ('report', 0.05169112595076477), ('cancer', 0.04687394655373266)]

Topic 1:

[('google', 0.6199432211963671), ('samsung‘, 0.4647407644932722), ('galaxy', 0.3972519251332346), ('says', 0.17182596426071345), ('apple', 0.14386302171460943), ('android', 0.1330713392779825), ('glass', 0.13117330837697994), ('tab', 0.1099072145269534), ('report', 0.10445918022761816), ('price', 0.08276757970152882)]

Topic 2:

[('samsung‘, 0.5275265496715218), ('galaxy', 0.459192654236729), ('tab', 0.11338869931862162), ('price', 0.08361848415269445), ('apple', 0.051806901002287985), ('note', 0.045800336897351294), ('specs', 0.04483063275553376), ('india', 0.042755633537156765), ('mini', 0.038437773223479306), ('features', 0.03624060134992984)]

Topic 3:

[('says', 0.7334439981745634), ('ebola', 0.2912417585498625), ('study', 0.2819044681057816), ('health', 0.12002575609088922), ('world', 0.11769503771591194), ('west', 0.11704154406432216), ('outbreak', 0.10062689666445736), ('virus', 0.09807160654546011), ('report', 0.09094477004970669), ('million', 0.08399790424182825)]

==========

Yuqoridagi ro‘yxatda LSI modeli bo‘yicha mavzular keltirilgan.

```
from sklearn.manifold import TSNE
# NLTK
from nltk.tokenize import RegexpTokenizer
from nltk.stem.snowball import SnowballStemmer
from nltk.corpus import stopwords
import re

# Visualizatsiya
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import matplotlib
%matplotlib inline
import seaborn as sns

# Bokeh
from bokeh.io import output_notebook
from bokeh.plotting import figure, show
from bokeh.models import HoverTool, CustomJS, ColumnDataSource, Slider
from bokeh.layouts import column
from bokeh.palettes import all_palettes
output_notebook()
```

Keyingi qadamda LSI modelini visualizatsiya qilamiz va model bo‘yicha bir-biriga yaqin so‘zlar va hujjatlarni ko‘rib chiqamiz.

```
import pandas as pd

from bokeh.io import push_notebook, show, output_notebook
from bokeh.plotting import figure
from bokeh.models import ColumnDataSource, LabelSet
output_notebook()

svd = TruncatedSVD(n_components=100)
documents_2d = svd.fit_transform(data_vectorized)

df = pd.DataFrame(columns=['x', 'y', 'document'])
df['x'], df['y'], df['document'] = documents_2d[:,0], documents_2d[:,1], range(len(data))

source = ColumnDataSource(ColumnDataSource.from_df(df))
labels = LabelSet(x="x", y="y", text="document", y_offset=8,
                  text_font_size="8pt", text_color="#555555",
                  source=source, text_align='center')

plot = figure(width=600, height=600)
plot.circle("x", "y", size=12, source=source, line_color="black", fill_alpha=0.8)
plot.add_layout(labels)
show(plot, notebook_handle=True)
```

Xulosa. Ushbu maqolaning asosiy g‘oyasi keng qo‘llaniladigan tematik modellashtirish usullarini amalga oshirish va taqqoslasdan iborat. Olingan

natijalarga ko‘ra NMF usuli eng yuqori muvofiqlik ballini bergan bo‘lsa-da, LDA eng ko‘p qo‘llaniladigan usul hisoblanadi va izchil deb hisoblanadi. Chunki LDA usuli ko‘proq "muvofiq" mavzularni taqdim etishi mumkin. Mavzu ehtimoli har bir hujjatda o‘zgarmas bo‘lishi kerak bo‘lsa, NMF yaxshiroq ishlaydi. Boshqa tomondan, HDP usulidan kamroq foydalaniladi, chunki mavzular soni oldindan aniqlanmagan va shuning uchun ushbu usul kamdan-kam qo‘llaniladi. Albatta, olingan natijalar berilgan namunaviy korpus ma'lumotlariga bog‘liq.

Foydalanilgan adabiyotlar:

1. Elov B., Alayev R., Aloyev N. (2024). Tematik modellashtirishning zamonaviy usullari. Digital transformation and artificial intelligence, *2*(1), 8–16. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i12
2. Elov.B., Alayev N. Matnlarini tematik modellashtirish va tasniflash usullari. barqarorlik va yetakchi tadqiqotlar onlayn ilmiy-amaliy jurnali. Vol. 3 No. 12 (2023). 263-276
3. Elov B., Aloyev N., Yuldashev A. (2023). SVD va NMF metodlari orqali tematik modellashtirish. O‘zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 2(6). 55-66
4. Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. International Journal of Advanced Computer Science and Applications, 6(1). https://doi.org/10.14569/ijacsa.2015.060121
5. Tao, R., Wei, Y., & Yang, T. (2021). Metaphor Analysis Method Based on Latent Semantic Analysis. Journal of Donghua University (English Edition), 38(1). https://doi.org/10.19884/j.1672-5220.202010087
6. Darmalaksana, W., Slamet, C., Zulfikar, W. B., Fadillah, I. F., Maylawati, D. S. adillah, & Ali, H. (2020). Latent semantic analysis and cosine similarity for hadith search engine. Telkomnika (Telecommunication Computing Electronics and Control), 18(1). https://doi.org/10.12928/TELKOMNIKA.V18I1.14874
7. Ke, Z. T., & Wang, M. (2022). Using SVD for Topic Modeling. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2022.2123813
8. Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. ACM Computing Surveys, 54(10). https://doi.org/10.1145/3507900
9. Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. EAI Endorsed Transactions on Scalable Information Systems, 7(24). https://doi.org/10.4108/eai.13-7-2018.159623

ӘОЖ 811.111

КОРПУСТЫҚ ЛИНГВИСТИКАНЫҢ ДАМУЫНДАҒЫ ПРОФЕССОР АСҚАР ЖҰБАНОВТЫҢ РӨЛІ

Миров М.О,
Филология ғылымдарының кандидаты, доцент
mirov-agu@bk.ru
Турсынова М.А,
Магистрант
malika.tursynova.00@mail.ru
Қ. Жұбанов атындағы Ақтөбе өңірлік университеті
Қазақстан, Ақтөбе

Аннотация. Бұл мақалада корпустық лингвистиканың қазақ тіл білімінде қалыптасуына, дамуына айрықша еңбек сіңірген профессор Асқар Жұбановтың еңбектері жан -жақты талданады.

Abstract: This article provides a comprehensive analysis of the scientific works of Professor Askar Zhubanov, who made a significant contribution to the formation and development of corpus linguistics in Kazakh linguistics. His research on the creation of a national corpus, the development of language models, and computational linguistics is examined.

Аннотация. В данной статье всесторонне анализируются научные труды профессора Аскара Жубанова, внесшего значительный вклад в формирование и развитие корпусной лингвистики в казахском языкознании. Рассматриваются его исследования в области создания национального корпуса, разработки языковых моделей и компьютерной лингвистики.

Кілт сөздер: *корпустық лингвистика, ұлттық корпус, тілдік модель, компьютерлік лингвистика, мәтіндік корпус, қолданбалы тіл білімі, машиналық аударма, формалды модельдер, мәтін лингвистикасы.*

Корпустық лингвистика тіл арқылы қоршаған ортаны, адамзат қасиеттерін, қоғамдық құбылыстар мен жаратылысты танып білудің ғылыми-теориялық негіздерін қарастырумен қатар, қазіргі заман талабына сай математикалық, формальды – логикалық, статистикалық әдістерді пайдалана бастады.

Корпустық лингвистика тіл білімінің танымдық бағыттары мен аспектілерін әр түрлі қырынан ғылыми негіздейді және адамның таным процесіне қатысты білімін тереңдетіп, тәжірибесін молайтады.

Осындай қажеттіліктен туындаған салаға қатысты прфессор Асқар Құдайбергенұлы Жұбановтың «Основные принципы формализации содержания казахского языка» атты док- торлық монографиясы қазіргі уақыттағы қолданбалы лингистикадағы маңызы зор еңбек деп тануға болады.

Қазіргі уақытта тіл білімі мен корпустық лингвистиканың бір-бірімен байланысын көрсете отырып, компьютерлік лингвистика саласы арқылы ақпараттық – танымдық жүйесін қалыптастырып, осы мәселенің танымдық аспектілеріне байланысты ғылыми еңбектерді, концептуалды пікірлерді, көзқарастарды талдап, талқылай отырып, корпустық лингвистика арқылы зерттеу және де қол жеткізген материалдарды ана тіліміздің материалдары арқылы қолданысқа енгізіле бастады.

Сонымен бірге корпустық лингвистикаға қатысты ұғымдарды, терминдерді, метатіл түсініктерін меңгеру қажеттігі оларды нақты мысалдар мен деректер арқылы талдап көрсету, оны теориялық жағынан қарастыра отырып, ой, сана, тіл тұтастығы арқылы тілдік қатынасты жете меңгеру, таным процесін жан-жақты іске асыруға болатындығын көреміз. Себебі, корпустық лингвистика-компьютерлік лингвистиканың бір саласы [1].

Корпустық лингвистика тілді ақпараттандыру және түрлі техникалар мен технологияларды пайдалану барысында зерттеуді қажет ететін маңызды ғылым саласы. Лингвистикалық корпустардың құрастырылуы мен оны пайдаланудағы жалпы ұстанымдары-корпустық лингвистикасаласының зерттеу аймағын қамтиды. Жалпы лингвистикалық немесе тілдік корпус дегеніміз - нақтыланған тілдік мәселелерді компьютермен шешуге арналған тілдік ақпараттар ауқымы.

А.Жұбанов компьютерлік лингвистика қазақ тіл білімінде қалыптаса қоймаған ХХ ғасырдың 60-70 жылдардың өзінде үлкен көрегендікпен аталған саланың болашақтың саласы екенін дәл болжаған. Қазақ тілінің ұлттық корпусын құрудың, онда түрлі белгіленімдерге негізделген алгоритмдік жүйені енгізудің маңыздылығын да атап көрсеткен. Ғалым сөз мағынасын бағ-дарламалаудың алғашқы сатысы саналатын лексика-семантикалық белгіленімге қатысты: «Лексика-семантикалық белгіленім – өте күрделі мәселе. Ол тіпті әлем тілдерінде де әлі күнге дейін шешімін таппаған», деп пікір білдірген.

Қазақ тіл білімінің қолданбалы саласын дамытуға, дәлірек айтқанда, қазақ цифрлық тіл білімін, қазақ инженерлік лингвистикасын, корпустық лингвистиканы қалыптастыруда, айтарлықтай қомақты үлес қосқан ғалым, есімі бүкіл түркі жұртына аты әйгілі маман, филология ғылымдарының докторы, профессор Асқар Құдайбергенұлы Жұбанов еңбектерінің маңызы зор деп білеміз. Асқар Жұбанов- есімі елге танылған біртуар тұлға, профессор

Құдайберген Жұбановтың бел баласы, сүт кенжесі. Асқар Жұбановтың математик маман болып, еңбек жолын жаңа ғана бастаған кезеңіне ойша зер салар болсақ, сол тұста ғылым салаларындағы ізденістердің алға басуы әртүрлі ғылым салаларының да тоғысып, заман талабына қарай жаңа бағыттағы зерттеулерді талап ете бастағанын байқатқан еді. Бұл жылдары ел басына төнген талай қиындықтардың бұлты сейіліп, өнер-мәдениет, ғылым мен техниканың дамуы жаңаша сипат алып, елдің де еркін тыныстай бастаған кезі еді. Осы кезде тілге де математика жағынан келіп, тілдік бірліктердің сандық өлшемін есептеп, оны алгоритмдік негізде жүйелеуге, мәтінді электронды есептеуіш машина арқылы қарастырып, аударма жасау ісін де компьютерлік техника көмегімен жүзеге асыруға бағытталған жаңа ғылыми сала ретінде «қолданбалы тіл білімі» ғылымының аясында «математикалық лингвистика» бой көтере бастаған болатын [2.108]. Осы тұста жас математик Асқардың бір-бірінен алыс жатқан екінші бір сала - лингвистикаға келуін бір жағы күтпеген жағдай десек, екіншіден тағдырдың жазуы деуге болады. Себебі әкесі профессор Құдайберген Жұбанов өзінің қысқа өмірінде тілдің сан-салалы бағытына қалам тербеген болса, ғылым дамуына байланысты дүниеге келген қазіргі тіл білімінің жаңа салаларының алғашқы ұшқындары да ғалым Құдайберген Жұбановтың еңбектерінен бастау алып жатқанын бүгінгі жас ғалымдар өз зерттеулерінде көрсетіп келеді. Осындай жаңаша зерттеу әдістерінің бірі, тілдік бірліктерді санмен байланыстыра зерттеудің алғашқы үлгісі Құдайберген Жұбанов еңбектерінен де табылған болатын [2.109]. Кезінде Қ.Б.Бектаев, С.Мырзабеков сияқты т.б. алғашқы тілші-статист ғалымдар өздерінің еңбектерінде Қолданбалы тіл білімінің «Статистикалық лингвистика» деп аталып жүрген жаңа саласының бастауында профессор Құдайберген Жұбановтың тұрғанын арнайы атап көрсеткен болатын.

Алғашқы уақытта тілді тек тілдік тұрғыдан зерттеу керек деген көзқарастағылар тілді математикалық есептеулермен байланыстыра қарастырудың қажеттілігін дұрыс сезіне алмай жатқан кезінде, яғни, бүгінгі таңда керекті ақпараттарды алу, не өз ақпаратымызды екінші бір жерге жіберу сияқты толып жатқан барлық іс-әрекетімізді автоматты түрде үйде отырып-ақ жүзеге асыратынымызды ойласақ, бүгінде цифрлық технология деген жаңа атаумен бүкіл адамзаттың қажеттілігіне айналып отырған осы салаға қарай алғашқылар қатарында Асқар Жұбановтың да бұрылуын ғылыми көрегендік, батыл қадам, ғылымдағы сирек құбылыс деп те бағалауға болады. Себебі жаңа салаға сол саланың мамандары да батыл бара бермейтінін ескерсек, тіл саласына математик маманның келуі, сонымен бірге өз жұмысында бір елге жүк болатындай пайдалы іс жасау, екінің бірі қолынан келе бермейтін шаруа деп есептейміз. Сонымен бірге әкесі Құдайберген Жұбановтың өзінің өлшеусіз ғылыми мәнді еңбектерімен қазақ тіл білімін биікке көтергенін, одан кейін

туған ағасы ғалым Есет Жұбановтың да қазақ фольклорының тілін зерттеуді алғашқы болып қолға алып, өзінің өзгеге ұқсамайтын терең білімімен ғылымға өшпестей үлес қосып кеткенін ескерсек, қазақ тіл білімінде алғашқы болып жаңа бір саланың іргесін қалап, оның ғылыми нәтижелерін ел игілігіне ұсынып, соңына ерген ізбасар ғалым шәкірттеріне жол көрсете білуі Асқар Жұбановтың да ғылымдағы өзіндік дара жолын көрсетеді.

Жұбанов Асқар Құдайбергенұлы 1937 жылғы күз айының 13 қарашасында Алматы қаласында дүниеге келді [3.382]. Әкесі Құдайберген Жұбанов солақай саясаттың құрығына іліккен кезде, дүние есігін ашқанына үш ай болған Асқар Жұбанов туған қаласы Алматыдан жер аударылғандай, әке-шешесінің туған жері Ақтөбедегі Мұғалжар өлкесінен бір-ақ шығады. Қдайберген Жұбановтың жөргектегі нәрестесі Асқардың жарық дүниеге келгеніне небәрі үш-ақ күн болғанда «жапон шпионы» деген жалған жаламен қамауға алынған сәтінен бастап, інілерінің тағдырына алаңдап, ағалық көмек көрсетіп, ақыл-кеңес беріп келген Ахмет Жұбанов болатын. Ол 1949 жылы бала Асқарды Алматыға алдырып, №18 қазақ орта мектеп-интернатының төртінші сыныбына оқуға береді. Бір жыл өте үлкен ағасы Ақырап 1950 жылы Алматыдағы көркем-сурет училищесін бітіріп, жолдама бойынша Петропавл қаласына жұмысқа тұрып, анасы мен бауырларын көшіріп алған кезде, отбасының сүт кенжесі Асқар да осы басқа жерде оқуын жалғастырады. Шыр етіп жарық дүниенің есігін ашқаннан тағдырдың тепкісін көрген кенже бала Асқар өзінің екі аға, үш апасы, «халық жауының әйелі» атанған Раушан ансымен бірге сол кезеңдегі қызыл саясаттың кесірінен әке есімі «ақталғанша» талай жерде қиындықты көріп өседі.

Бала кезінен техника, математика бағытындағы салаларға ерекше қызығушылық танытып өскен Асқар Құдайбергенұлы орта мектепті бітірген соң 1955 жылы Алматы қаласындағы Тау-кен (политехникалық) институтына ең жоғары балл жинап, оқуға түсіп тұрған сәтінде, сол қатал заманның «Халық жауының баласы!» деген айыптауы тағы да алдын орап, білім қуған талапты жастың арманының тауы шағылған еді. Қай заманда да үлкен жүректі, адал азаматтар болды десек, алаштың атақты перзенттерінің бірі, сол кезде ҚазПИ-дің ректоры, Кеңестер Одағының Батыры Мәлік Ғабдуллиннің: «Сенің әкең – ұлы кісі» деп, сол сәтте жас жігіттің жарым көңіліне жарық нұр сепкендей болып, бауырына басып, 1955 жылы Абай атындағы ҚазПИ-дің физика-математика факультетіне оқуға түсуіне қол ұшын бергенін Асқар Жұбанов әр кез айтып отыратын [4.6].

Әке жолын қуып, әке рухына табына жүріп, тіл саласына келген математика маманы Жұбанов Асқар Құдайбергенұлы 1967-1970 жылдары Қазақ ССР Ғылым академиясы Тіл білімі институтының күндізгі бөлімінің аспиранты болып, Минск қаласындағы Шет тілдер институты жанында

ғылыми тәжірибеден өтіп, Мәскеу-Ленинградтың осы саладағы ғалымдарымен байланыс жасап тұрады. Тіл білімі институтындағы тілді зерттеуде математика мүмкіндігін пайдалануды іске асыратын жаңа бағыттағы «Статистика-лингвистикалық зерттеу және автоматтандыру» тобының кіші, кейіннен аға ғылыми қызметкері бола жүріп, тілдің сөз байлығын жаңа қырынан зерделеу жолында лингвистикалық сөздіктерді сандық және сапалық белгілеріне қарай математикалық статистика тәсілімен зерттеуді қолға алған алғашқы белгілі математик - ғалым Қалдыбай Бектаевпен бірлесе еңбек етеді. Асқар Құдайбергенұлы 1973 жылы физика-математика ғылымдарының кандидаты, филология ғылымдарының докторы Қ.Б.Бектаевтың және академик І.Кеңесбаевтың ғылыми кеңесшілігімен «ЭЕМ көмегімен қазақ мәтінін статистикалық зерттеу» атты тақырыпта кандидаттық диссертациясын қорғап, қазақ тілін математикалық әдіспен, яғни статистика мен ықтималдылықтар теориясы әдістерімен зерттеу жұмыстарын жүргізді [3.382]. Қазақша мәтіндерді электронды есептеу машинасының көмегімен зерттеп, тұңғыш рет сөзформа, негіз сөз, әріп, әріп тіркесі, т.б. тілдік бірліктердің автоматтанған жиілік сөздіктерін құрастырудың алгоритмі мен бағдарламасын жасап шығарады. Осындай зерттеулер нәтижесінде заманымыздың заңғар жазушысы Мұхтар Әуезов «Абай жолы» эпопеясында 465691 сөзқолданысты пайдаланғанын, онда сөз саны – 16983, ал сөзформасы – 61424 екенін ғылыми түрде анықтайды[5.68]. Ал 2002 жылы А. Байтұрсынұлы атындағы Тіл білімі институтының «Қолданбалы лингвистика» бөлімінің меңгерушісі қызметін атқара жүріп, сол мекеменің ғылыми кеңесінде «Қазақ мәтіні мазмұнын формалдаудың негізгі ұстанымдары» атты тақырыпта докторлық диссертация қорғап, филология ғылымдарының докторы ғылыми атағын алады. Ол қазақ тіл білімінің корпустық лингвистика деп аталған жаңа бағыттағы саласының бірден-бір білікті маманы ретінде өзіне ізбасар ғалым-шәкірттер дайындаумен шұғылданып, 1991 жылдан бастап А. Байтұрсынұлы атындағы Тіл білімі институтының «Қолданбалы лингвистика» бөлімінің меңгерушісі, кейіннен жаңаша аталымдағы «Компьютерлік лингвистика» ғылыми тобының бас ғылыми қызметкері міндеттерін атқарды.

Асқар Жұбанов өз кезінде «М.О. Әуезовтің толық шығармалар жинағы мәтінін ЭЕМ-ге енгізуге дайындау және сөзнұсқағыш-сөздіктерді құрастыру», «М. Әуезовтің 20 томдық шығармалары бойынша автоматтанған картотекалық қор жасау», «М. Әуезовтің шығармалары бойынша сөзформалар мен сөздердің жиілік сөздіктерін ЭЕМ арқылы құрастыру» атты ғылыми тақырыптарға жетекшілік еткен еді. «Профессор» ғылыми атағын алған ғалым Асқар Жұбанов корпустық лингвистиканың еліміздегі алғашқы ірі маманы ретінде әл-Фараби атындағы Қазақ ұлттық университетінің филолог магистранттары мен студенттеріне математикалық және компьютерлік лингвистика пәндері бойынша дәріс беріп, осы салаға арналған жоғары оқу

орындарына арналған оқулықтарын дайындап шығарады [6]. Корпустық лингвистика маманы ғалым Асқар Жұбановтың жетекшілігімен осы саланың 5 кандидаттық және бірнеше магистрлік диссертациясы қорғалды.

Корпустық лингвистиканың білгір маманы ретінде Асқар Жұбанов «Квантитативная структура казахского текста»(1987), «Куманша-қазақша жиілік сөздік» (құрастыруш. бірі. -1978), «М.О.Әуезовтің 20 томдық шығармалар текстерінің жиілік сөздіктері» (құрастыруш. бірі. -1995), «Основные формализации содержания казахского текста»(2002), «Қолданбалы лингвистика: қазақ тілінің статистикасы»(2004), «Қолданбалы лингвистика: формалды модельдер»(2006), «Қолданбалы тіл білімінің мәселелері»(2008), «Корпустық лингвистика»(2017), т.б. еңбектер жазып қалдырды [3.385]. Ғалым Асқар Жұбанов математика маманы ретінде тілге де математика жағынан келіп, тілдік бірліктердің сандық өлшемін есептеп, оны алгоритмдік негізде жүйелеуге, мәтінді электронды есептеуіш машина арқылы қарастырып, аударма жасау ісін де компьютерлік техника көмегімен жүзеге асыруға бағытталған жаңа ғылыми сала ретінде «қолданбалы тіл білімі» ғылымының аясында «математикалық лингвистика», «статистикалық лингвистика» деп түрліше аталған салаға көп үлес қосты. «Статистикалық лингвистика» - тілдің грамматикалық категорияларын, тілдегі барлық деректерді сандық тұрғыда тексеріп, тіл элементтерін өлшеудің сандық мәселелерімен айналысатын қолданбалы тіл білімінің бір жаңа саласы. Тілдегі жиілік сөздіктердің түрлерін, кері әліпби-жиілік сөздік, сөзнұсқағыш әліпби-жиілік сөздік, мәтін мен оның жиілік сөздігі бірліктерімен арақатынасы, сөздерді компьютер арқылы алудың біріккен және іріленген алгоритмін жасау, цифр заңы және оны қазақ мәтіні бойынша түзілген жиілік сөздіктерге қолдану, мәтін және оның жиілік сөздігі ішіндегі сөздің ақпараттық сипаттамасын беру, лингвистикалық болжамды тексеру критерийін анықтау, т.б. мәселелерді қарастырады. Ғалымның осы статистика мәселесі жөнінде жазған оқулығы бүгінде жастардың игілігіне айналды. Қазақ цифрлық тіл білімін қалыптастыруда ғалым А.Қ.Жұбанов еңбектерінің маңызы зор. Асқар Құдайбергенұлы өзінің зерттеу нәтижелерін ғалым шәкірті А. Жаңабековамен бірлесіп жазған «Корпустық лингвистика» атты монографиясында баяндады.

Көпшілікке жете таныс емес «формалды логика» дегенді тіл ғылымына қарай бұру, қазақ тілін зерттеуде математикалық әдістерді қолдануда, сол арқылы «машиналық аударма» жасаумен айналысатын ғылыми саланың дүниеге келіп, оның статистика, формалды модельдер, мәтін лингвистикасы, қолданбалы лингвистика, компютерлік лингвистика, корпустық лингвистика деген атаулармен дамуында осы бағыттың алғашқы мамандардың бірі ретінде Асқар Жұбановтың көп еңбегі сіңген еді [7]. Қазіргі таңда қазақ тіл білімінде аталған саланың сандық технологиялармен үндесе жаңа қырынан дами бастағанын ескерсек, Асқар Жұбанов еңбектерінің заман көшімен бірге ілесіп,

жаңа леппен дамып бара жатқанын көріп, ғалымның осы жаңа саланың алғашқы бастаушылардың бірі ғана емес, оны әрі қарай алып барушы да болғанын мақтан етеміз.

Қазіргі ғылыми-техникалық үдерістер заманында ақпарат көлемінің бұрын-соңды болмаған дәрежеде артуына байланысты қажетті мәліметтерді тез тауып алу бағытындағы ізденістер Асқар Құдайбергенұлының компьютерлік лингвистика саласындағы зерттеулерінде өз шешімін тауып, ақпаратты іздестіру, оны табу әрекеттері, аталған қызметтерді жүзеге асыратын құралдары мен әдістері, яғни ішкі жүйелерді жабдықтау түрлері, ақпарат іздестіру тілінің грамматикасы, қазақ тілінің компьютерлік қорын құрастыру, т.б. теориялық мәселелер қарастырылды. Бүгінгі таңда осы аталған саланың ғылым қуған жас мамандары Асқар Құдайбергенұлының зерттеулері арқылы корпустық лингвистиканың ғылыми негіздерімен танысып, корпустық технологиялардың мәнін түсініп, корпустармен жұмыс істеу дағдысын меңгеріп жатыр. Ғалым Асқар Жұбановтың жетекшілігімен қазақ тілінің жиілік сөздіктері де шығарылып, корпустық лингвистиканың жас мамандары да ғылыми жетістіктерімен тілімізді жаңа бағытта зерттеуге үлес қосып келеді [8]. «Корпустық лингвистика компьютерлік технологияларды қолдана отырып, тілдік мәліметтердің лингвистикалық корпусын құрастыру мен оны қолданудың жалпы принциптерін зерттеумен айналысады. Ол мынадай үш түрлі тәсілді қолданады: 1. Корпус ішінен тілге қатысты мәліметтерді автоматты түрде таңдап алу; ақпаратты өңдеу; өңделген мәліметтерді тексеру және оларға түсініктеме беру»,-деп түсіндірген еді [9,17]. Ғалымдардың көрсетуіне қарағанда, осы тәсілдер алдағы уақытта да жан-жақты зерттеулермен жалғаса бермек.

Қазақтың «тектілік» деген түсінігін дәлелдеп, халқымыздың мақтанышына айналған «Жұбановтар әулеті» деген қымбат атты жоғары ұстаған ғалым ағамыз Асқар Жұбанов өзінің көзі тірісінде әке-шешесінің туған жері Ақтөбе, Мұғалжар өлкесінің жұртшылығына сыйлы, әкесінің есімін иеленген Құдайберген Жұбанов атындағы Ақтөбе өңірлік мемлекеттік университетінің профессор-оқытушыларына жақсы таныс, тығыз қарым – қатынас орната білген ардақты азамат еді. Университеттің ғылыми-педагогикалық өміріне өзінің ғылыми ақыл-кеңестерімен қол ұшын беріп, жастардың ғылыми шығармашылығына қолдау көрсетіп жүретін. Асекең университет жастарымен келген сайын кездесіп, оларға ғылыми тұрғыда ағалық, аталық ақылын айтып, қазіргідей жастарға барлық жағдайлар жасалып жатқан кезеңде қиындықтан қашпай еңбектенуге, оқу-білімге зейін салып, ғылыммен шұғылдануға шақырушы еді. Сондай кездесулердің бірінде: «Өмірдің жақсы да, жаман да жақтарын басымнан кешірдім, бірақ оның қиын тұстарының бәрі де әкемнің нақақтан «халық жауы» аталуымен

салыстырғанда түк те емес», - деп, өткен күндерді еске ала отырып, бақытты заманның бақытты жастарына батасын беріп, университетіміздің жақсы істеріне куә болып, жеткен биіктеріне балаша мәз болып отыратын... Ақтөбе жеріне әр келген сайын университеттің қазіргі білім беру ісіндегі заман талаптарына сай жұмыс жүргізіп, көш басындағы алдыңғы оқу орындары қатарынан көрініп келе жатқанына қуанып, өзінің ғалымдық, азаматтық келбетімен білімді жастарға рухани күш сыйлайтын. Университетте «Жұбановтану» саласындағы жас зерттеушілерге де да бағыт-бағдар беріп отыратын. Жұбанов Асқар Құдайбергенұлы өзінің ғалымдық еңбегі үшін «Қазақстан Республикасының ғылымын дамытуға сіңірген еңбегі үшін» (2013) медалі, Қ.Жұбанов атындағы Ақтөбе өңірлік мемлекеттік университеті Ғылыми кеңесінің шешімімен «Жұбанов» медалімен (27.11.2017 ж. №3 хаттама), Қазақстан Республикасының «Құрмет грамотасы» (5.12.2017 ж.) , «Қазақстан Республикасының ғылымын дамытуға сіңірген еңбегі үшін» (2017) медалі, «Еңбек ардагері» медалімен (2018) марапатталды. Ахмет Байтұрсынұлы атындағы Тіл білімі институты Қолданбалы лингвистика бөлімінде тапжылмай көп жылдар бойы меңгерушілік қызмет атқараған профессор Асқар Құдайбергенұлы Жұбанов Қолданбалы лингвистика саласының Статолингвистика, Формалды модельдеу, Компьютерлік лингвистика сияқты салалары бойынша зерттеулер жүргізе жүріп, және осы салалар бойынша Жоғары оқу орындарында дәріс бере жүріп, Қазақ тілінің ұлттық корпусын жасау қажеттігін де ертеден-ақ көрегендікпен аңғарған болатын.

Мұны ғалым А.Жұбановтың: «Әлемдік корпустық лингвистиканың даму сипаты – ұлттық толық мәтіндерді арнайы зерттеу нысаны етіп алуды қажет етуде. Қазақ мәтіндерінің стилдік, құрылымдық, мағыналық, функционалдық және т.б. сипаттарын анықтау өзекті мәселелердің бірі болып отыр. Сондықтан автоматтанған қазақ тіліндегі мәтіндер корпусының компьютер- лік базасын құру да ғылыми және практикалық жағынан аса құнды мәселе», – деуінен көреміз. Қолданбалы лингвистика бөлімінің зерттеу бағытын осы мәселеге бұрды.

Цифрлы технологияларды қолдану арқылы тілдік материалдарды автоматтандыру маңыздылығы бүгінде тіл білімінің «Компьютерлік лингвистика», оның ішінде «Корпустық лингвистика» сияқты салаларының дамуына, қалыптасуына мүмкіндік берді.

Жалпы қазақ тілінің ұлттық корпусын құрастыру мәселесі жиілік сөздік құрастыру, картотекадан басталған еді. Бүгінде бұл іспен Мақсатты қаржыландыру жобасы аясында А.Байтұрсынұлы атындағы Тіл білімі институт ұжымы айналысып келеді. Бұл Ұлттық корпус мәселесінде ғалым А.Қ. Жұбанов корпустарды бөліп, жіктеп қарастырған болатын. Соның бірі,

қолданбалы лингвистика бөлімі айналысып отырған параллель корпусы бар, тарихи корпус, іскери, ресми және т.б. осындай ішкорпустармен қоса Асқар Жұбанов айтып кеткендей қазіргі баспа ісін автоматтандыру, яғни, бұл да корпустың бір бөлігіне кіретін маңызды мәселелер екенін, сол уақытта 2013 жылдары «Компьютерлік лингвистикаға кіріспе» атты еңбегін шығарып, оның бір тарауына «баспа ісін автоматтандырудың лингвистикалық негіздері» деп кезеңдерге бөліп қарастырып кеткен болатын.

Қорыта айтқанда, қазақ тіл білімі ғылымының жарық жұлдызы атанған Құдайберген Жұбановтың даңқына сай ұл болып, «Әкеден ұл тумас па, әке жолын қумас па?» деп дана қазақтың әр қариясы арман еткендей ғылымдағы толағай табыстарымен танылған Жұбанов Асқар Құдайбергенұлының артында ғылыми мол мұрасы, жоғары оқу орындарына арналған оқу құралдары, үлгі-өнеге берген өзінің ізін қуған талантты шәкірттері қалды. Ахмет Байтұрсынұлы атындағы Тіл білімі институтының бас ғылыми қызметкері ретінде зерттеу жүргізен корпустық лингвистика саласындағы ғылыми еңбектері кейінгі ұрпаққа рухани мол мұра деп білеміз.

Пайдаланылған әдебиеттер

1. Жұбанов А.Қ., Жаңабекова А.Ә. Корпустық лингвистика «Қазақ тілі» баспасы, – Алматы, 2017.
2. «Құдайберген Жұбанов және қазақ совет тіл білімі». - Алматы: «Ғылым», 1990. -368 б.
3. Құдайберген Жұбанов. (Ұлы дала тұлғалары). (жалпы редакциясын басқарған ф.ғ.д., профессор М.Малбақов). – Алматы: «Шафран» ЖШС, 2013.- 528 б.
4. Оразбаева Ф.Ш. Жұбанов тағылымының ұлағаты // «Жұбанов тағылымы-IX» дәстүрлі халықаралық ғылыми конференция материалдары жинағы. - Ақтөбе: Қ.Жұбанов атындағы Ақтөбе ӨМУ, 2017.- 5-9- б.
5. Жұбанов А.Қ. Қолданбалы лингвистика: Қазақ тілінің статистикасы. Оқу құралы – Алматы: «Қазақ университеті», 2007. -210 б.
6. Жұбанов А.Қ. Компьютерлік лингвистикаға кіріспе: Оқу құралы – Алматы: «КИЕ лингвоелтану инновациялық орталығы», 2013. -204 б.
7. Жұбанов, А. Қ. Қолданбалы лингвистика: формалды модельдер : Оқу құралы. - Алматы : Қазақ университеті, 2006. - 280 б.
8. Жұбанов А., Жаңабекова А., Б.Д. Карбозова, А.Қ. Қожахметова Қазақ тілінің жиілік сөздігі. –Алматы. «Қазақ тілі» баспасы, 2016. – 792 б.

Жұбанов А., Жаңабекова А. Корпустық лингвистика. - Алматы: «Қазақ тілі» баспасы, 2017. - 336 б.

UDK: 004.65, 8, 81’2, 81’28

XORAZM SHEVASIDAGI MATNLARNI O‘ZBEK ADABIY TILIGA O‘GIRISH MASHINALI TARJIMA MODELINI QURISH UCHUN PARALLEL KORPUS YARATISH

Sharipov Maqsud Siddiqovich,
dotsent
maqsbek72@gmail.com
Urganch davlat universiteti

Kurbanova Lola Ulug‘bek qizi,
o‘qituvchi
lolakompyuterlingvistikasi@gmail.com
Urganch davlat universiteti

Annotatsiya. Ushbu maqolada Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga mashinali tarjima qilish imkonini beruvchi tizimni qurish uchun zarur bo‘lgan **parallel korpus yaratish** masalasi yoritilgan. Dialektal tilni standart tilga moslashtirishda dastlabki va muhim bosqich sifatida sheva-adabiy til juftliklaridan iborat parallel korpus tuzildi. Korpusni shakllantirishda Xorazm shevasida yozilgan ijtimoiy tarmoq matnlari, og‘zaki xalq ijodi namunalarining matnga ko‘chirilgan variantlari, hamda hududiy OAV materiallari asos sifatida tanlandi. Har bir shevada berilgan gapga mos holda o‘zbek adabiy tilida qo‘lda tarjima tayyorlandi. Korpus XML formatida belgilandi va u mashinali tarjima modellarini kelgusida o‘rgatish uchun tayyorlangan ma’lumotlar bazasini tashkil etadi. Tadqiqot natijasi sifatida mashinaviy tarjima va dialektologik izlanishlar uchun ochiq parallel resurs yaratildi.

Abstract. This article discusses the creation of a **parallel corpus** necessary for developing a system capable of performing machine translation from the Khorezm dialect into the Uzbek literary language**.** As an initial and essential stage in adapting dialectal language to the standard form, a parallel corpus consisting of dialect–standard sentence pairs was compiled. In the process of forming the corpus, texts written in the Khorezm dialect were collected from social media**,** transcribed samples of oral folk literature, and regional media sources. Each dialectal sentence was manually translated into its equivalent in the literary Uzbek language. The corpus was annotated in XML format and serves as a structured database prepared for training machine translation models in the future. As a result of the study, an open-access parallel resource was created to support further research in machine translation and dialectology.

Аннотация. В данной статье рассматривается вопрос создания **параллельного корпуса**, необходимого для построения системы машинного

перевода, переводящей тексты на хорезмском диалекте на узбекский литературный язык. На первом и важном этапе адаптации диалектной речи к стандартному языку был сформирован параллельный корпус, состоящий из пар предложений на диалекте и их ручных переводов на литературный узбекский язык. При формировании корпуса в качестве основного источника использовались тексты на хорезмском диалекте из социальных сетей, записанные версии устного народного творчества, а также материалы региональных СМИ. Каждому диалектному предложению был сопоставлен эквивалент на литературном языке, подготовленный вручную. Корпус был размечен в формате XML и представляет собой базу данных, предназначенную для последующего обучения моделей машинного перевода. В результате исследования был создан открытый параллельный ресурс, предназначенный для задач машинного перевода и диалектологических исследований.

Kalit so‘zlar: *Xorazm shevasi, o‘zbek adabiy tili, parallel korpus, mashinali tarjima, dialektologiya, NLP.*

Zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasining jadal rivojlanishi natijasida mashinali tarjima texnologiyalari dunyo tillarining turli variantlari, shevalari va lahjalarini avtomatik tarzda standart tilga konvertatsiya qilish imkonini bermoqda. O‘zbek tilida ham hududiy shevalarning ko‘pligi, ularning leksik, fonetik va grammatik xususiyatlari tilshunoslar, tarjimonlar va NLP mutaxassislari uchun yangi ilmiy vazifalarni yuzaga keltirmoqda.

O‘zbekiston hududida keng tarqalgan Xorazm shevasi o‘zining boy iboralari, o‘ziga xos talaffuz va uslubiy vositalari bilan ajralib turadi. Ayni paytda, rasmiy yozishmalar, ta’lim tizimi va ommaviy axborot vositalarida adabiy tilning ustuvorligi sababli, Xorazm shevasida yaratilgan matnlarning adabiy tilga avtomatik tarjima qilinishi dolzarb masalaga aylanmoqda. Biroq, ushbu jarayonni amalga oshirish uchun eng asosiy shartlardan biri bu – **sheva-adabiy til juftliklaridan iborat parallel korpus mavjud bo‘lishidir.**

Ushbu maqolada Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga tarjima qilishga mo‘ljallangan mashinali tarjima modelini qurish uchun zarur bo‘lgan **parallel korpusni yaratish bosqichlari,** manbalar tanlovi, matnlarni normalizatsiya qilish, XML formatida belgilash, va kelgusidagi tadqiqotlar uchun yaroqli ma’lumotlar bazasini shakllantirish masalalari keng yoritiladi. Korpusda aks ettirilgan matnlar turli janrlarga oid bo‘lib, ijtimoiy tarmoqlar, og‘zaki ijodiy materiallar va mahalliy nashrlardan olingan. Har bir sheva gapiga mos tarzda adabiy tarjima qo‘lda tuzilgan bo‘lib, bu o‘zbek tilida ilk marta dialektal parallel korpus yaratish yo‘lidagi muhim qadamlardan biridir. [2].

Tabiiy tilni qayta ishlash (NLP) va mashinali tarjima sohalarida dunyoda keng ko‘lamli tadqiqotlar olib borilmoqda. Ayniqsa, shevalar va lahjalardan adabiy tilga tarjima qilishga doir tadqiqotlar oxirgi yillarda dolzarb mavzulardan biriga aylangan. Arab, xitoy, hind va nemis tillarida bu borada keng qamrovli korpuslar yaratilgan va maxsus mashinali tarjima modellarida testdan o‘tkazilgan. Biroq, o‘zbek tilida dialektologik yondashuv asosidagi NLP ishlanmalari hanuz yetarli emas.

E. Kuriyozov, U. Salaev, va S. Matlatipov[3] tomonidan olib borilgan boshqa bir tadqiqotda o‘zbek tilidagi **matnlar asosida tasniflash datasetlari, nomuhim so‘zlar to‘plamlari, lemmatizatsiya modullari,** va **sentiment tahlil korpuslari** ishlab chiqilgan. Ular ko‘pincha umumiy adabiy tilga asoslangan bo‘lib, dialektal variantlar, xususan Xorazm shevasi uchun maxsus parallel korpus yaratilmagan. [1]

Shuningdek, **Madatov, Bekchanov va Vičič** [4] o‘z tadqiqotlarida o‘zbek tilida stop-so‘zlar va ularni aniqlash bo‘yicha statistik tahlillarni taqdim etishgan. Biroq, bu ishlar asosan yozma adabiy tilga asoslangan bo‘lib, shevalar, og‘zaki nutq va mintaqaviy lahjalarni qamrab olmaydi.

Xalqaro tajribaga e’tibor qaratilsa, arab tilining misrlik, livanlik va boshqa lahjalari uchun yaratilgan parallel korpuslar yordamida **sheva–standart til mashinali tarjimasi** muvaffaqiyatli amalga oshirilmoqda. Bu modeldan ilhomlanib, o‘zbek tilining Xorazm shevasini o‘rganish va tarjima qilishda ham xuddi shunday parallel juftliklar asosida korpus yaratish samarali yechim sifatida ko‘rilmoqda.

Xulosa qilib aytganda, mavjud adabiyotlar o‘zbek tili uchun korpuslar, morfologik va sintaktik modellar, lemmatizatsiya va transliteratsiya tizimlari borasida asos yaratib bergan bo‘lsa-da, **dialektal matnlar asosida parallel korpus tuzish** va ularni mashinali tarjima uchun tayyorlash borasida hali ilmiy bo‘shliq mavjud. Mazkur tadqiqot aynan shu bo‘shliqni to‘ldirishga qaratilgan bo‘lib, Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga o‘girishda birinchi parallel korpus namunasi bo‘lib xizmat qiladi.

Ushbu tadqiqotning asosiy metodologik yondashuvi – Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga mos tarzda juftlab, mashinali tarjima modellarini qurish uchun tayyor parallel korpus yaratishdan iborat. Tadqiqot quyidagi bosqichlarda amalga oshirildi:

Korpusni shakllantirish uchun Xorazm shevasida yozilgan turli janrdagi va shakldagi matnlar yig‘ildi. Asosiy manbalar quyidagilardan iborat:

- **Ijtimoiy tarmoqlardagi matnlar** (Telegram, Facebook postlari, izohlar);

- **Xalq og‘zaki ijodi namunalari** (ertaklar, latifalar, maqollar va hikoyalar);
- **Hududiy nashrlar va intervyular** (Xorazm viloyatidagi gazeta maqolalari va bloglar);
- **O‘zaro suhbatlardan olingan transkriptlar** (dialektal og‘zaki matnlarning yozma ko‘rinishlari).

Yig‘ilgan Xorazm shevasidagi matnlarning har bir gapiga mos tarzda o‘zbek adabiy tilidagi tarjimalar qo‘lda tayyorlandi. Tarjimalar ona tili bo‘yicha mutaxassislar va Xorazm shevasini chuqur biladigan tarjimonlar tomonidan amalga oshirildi. Har bir gaplik juftlikda semantik moslik va grammatik to‘g‘rilikni ta’minlashga alohida e’tibor qaratildi.

Matnlar va ularning tarjimalari XML formatida belgilanib, korpus holatiga keltirildi.

XML formatdan foydalanish:

- matnlarning strukturaviy tahlilini osonlashtirish;
- hujjatning uslubi, turi, manbasi, yozilgan sanasi, muallifi kabi **meta-ma’lumotlarni** saqlash;
- mashinali tahlilga tayyor shaklda saqlash imkonini beradi.

Misol uchun:

```
<CORPUS NAME="Xorazm-Uzbek">
  <TEXT ID="001" TYPE="og‘zaki" STYLE="sheva">
    <AUTHOR>Anonim</AUTHOR>
    <SOURCE>Telegram</SOURCE>
    <DATE>2024-12-01</DATE>
    <CATEGORY>Folklor</CATEGORY>
    <CONTENT TITLE="Qishloq hikoyasi">Opom galdi, bog‘ tarafni
sirip qo‘ydim.</CONTENT>
    <STANDARD>Onam keldi, hovlini supurib qo‘ydim.</STANDARD>
  </TEXT>
</CORPUS>
```

Korpusdagi har bir sheva-adabiy gaplik juftlik .tsv (tab bilan ajratilgan qiymatlar) yoki .csv formatda saqlandi. Bu format mashinali tarjima modellarini (masalan, Transformer, mBART) o‘rgatishda keng qo‘llaniladi.

Format namunasi:

sheva_matni	adabiy_matn
Doyim oytdi, boroli.	Amakim aytdi, boraylik.

Ma’lumotlar ustidan qo‘lda va dasturiy tozalash ishlari olib borildi:

- noto‘g‘ri imloviy yozuvlar tuzatildi;
- orfografik va grammatik xatoliklar tuzatildi;
- takrorlanuvchi yoki kontekstdan tashqaridagi gaplar chiqarib tashlandi;
- semantik jihatdan nomuvofiq juftliklar belgilandi va ajratib qo‘yildi.

Umumiy natija sifatida, taxminan **10 000 dan ortiq gaplik parallel korpus** yaratildi. Ushbu korpus mashinali tarjima modellarini keyingi bosqichda o‘rgatish uchun tayyor holatga keltirildi. Korpus o‘zbek tilida ilk marta hududiy dialekt (Xorazm shevasi) asosida tayyorlangan parallel resurs sifatida ilmiy va amaliy ahamiyatga ega.

Ushbu tadqiqot doirasida Xorazm shevasini o‘zbek adabiy tiliga tarjima qilish imkonini beruvchi mashinali tarjima modelini yaratishga tayyorlov bosqichi sifatida **parallel korpus** shakllantirildi. Korpus yaratish jarayoni davomida to‘plangan materiallar, ularning tahlili va tuzilgan juftliklarning umumiy statistikasi quyida bayon etiladi.

Yig‘ilgan matnlar **leksik boyligi, morfologik xususiyatlari** va **sintaktik tuzilmasi** jihatidan adabiy o‘zbek tilidan ancha farq qilishi kuzatildi.

Bu xususiyatlar tarjima modelini qurishda e’tiborga olinadigan muhim lingvistik ko‘rsatkichlardir.

Yaratilgan parallel korpus quyidagi asosiy sonli ko‘rsatkichlarga ega:

Ko‘rsatkich	**Qiymat**
Jami gaplar soni	10 000+
O‘rtacha gap uzunligi	8–12 so‘z
Foydalanilgan manbalar	Telegram, Facebook, matbuot
Adabiy tarjima usuli	Qo‘lda, mutaxassislar tomonidan
Saqlash formati	XML, TSV
Belgilov formatlari	<TEXT>, <AUTHOR>, <STANDARD> va h.k.

XML formatda korpusni belgilash natijasida har bir matnga oid meta-ma’lumotlar (muallif, janr, manba, sana, sarlavha) ham aniqlashtirilib saqlandi.

Yaratilgan parallel korpus quyidagi jihatlarda muhim ahamiyat kasb etadi:

- **Mashinali tarjima modellarini o‘qitish** uchun tayyor va tozalangan resurs sifatida xizmat qiladi.
- **Dialektologik tadqiqotlar**da shevalarning morfologik va sintaktik tuzilmasini o‘rganish imkonini beradi.

- O‘zbek tilining **hududiy variantlari bo‘yicha lingvistik vositalar** yaratishga asos bo‘ladi.
- **Raqamli lingvistika va ta’lim tizimi**da og‘zaki va dialektal matnlarni standart tilga moslashtirishda qo‘llanishi mumkin.

Xulosa tarzida, tadqiqot natijasida o‘zbek tilida ilk bor Xorazm shevasidan adabiy tilga o‘girishga mo‘ljallangan **parallel korpus** yaratildi. Bu korpus nafaqat mashinali tarjima modellarini yaratish, balki dialektologiya, leksikografiya va NLP tadqiqotlari uchun keng imkoniyatlar eshigini ochadi.

Ushbu tadqiqot natijasida Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga avtomatik tarzda tarjima qilish imkonini beruvchi mashinali tarjima tizimini yaratish uchun zarur bo‘lgan **parallel korpus** muvaffaqiyatli shakllantirildi. Korpus turli manbalardan yig‘ilgan dialektal matnlar va ularning qo‘lda tuzilgan adabiy tarjimalaridan iborat bo‘lib, mashinali tarjima modellarini o‘qitish uchun mustahkam poydevor yaratdi. Korpusning XML va TSV formatlarida tuzilgani uni keyingi bosqichda avtomatlashtirish va modellashtirish jarayonlarida yengillik yaratadi. Bu loyiha o‘zbek tilining hududiy variantlarini standartlashtirish va NLP yondashuvlari orqali integratsiyalashishiga xizmat qiladi.

Xorazm shevasi bilan cheklanib qolmasdan, boshqa hududiy shevalar (masalan, Farg‘ona, Qashqadaryo, Andijon) asosida ham parallel korpuslar yaratish zarur. Bu butun o‘zbek tilining dialektal xaritasini qamrab olgan mashinali tarjima tizimlarini ishlab chiqishga yo‘l ochadi.

Yaratilgan parallel korpusni asos qilib olib, mashinali tarjima modellarini (masalan, Transformer, mBART, mT5) o‘qitish va baholash bo‘yicha eksperimentlar o‘tkazish tavsiya etiladi. Bu orqali korpusning amaliy samaradorligi va NLP vositalarida qo‘llanilish darajasi aniqlanadi.

Yaratilgan korpusni ochiq manbada joylashtirib, ilmiy hamjamiyat va til texnologiyalari bilan shug‘ullanuvchi mutaxassislar uchun foydalanishga taqdim etish kerak. Bu nafaqat ilmiy almashinuvni kuchaytiradi, balki korpusning doimiy boyitilishini ham ta’minlaydi.

Foydalanilgan adabiyotlar:

1. E. Kuriyozov, U. Salaev, S. Matlatipov, and G. Matlatipov, “Text classification dataset and analysis for Uzbek language,” arXiv preprint arXiv:*2302.14494*, 2023.
2. Sharipov M.S., Kurbanova L.U., Qurbanova R.U. O‘zbek tili korpusi dasturiy ta’minotini yaratish // CTCL-2023.
3. Analysis for Uzbek language,” arXiv preprint arXiv:*2302.14494*, 2023.

4. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” in *CEUR Workshop Proceedings*, 2022, pp. 126– 136. [Online]. Available: www.scopus.com

5. M. Sharipov, E. Kuriyozov, O. Yuldashev, and O. Sobirov, “UzbekTagger: The rule-based POS tagger for Uzbek language,” arXiv preprint arXiv:2301.12711, 2023.

6. K. Madatov, S. Bekchanov, and J. Vičič, “Accuracy of the Uzbek stop words detection: a case study on” School corpus»,” arXiv preprint arXiv:2209.07053, 2022.

7. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of stopwords extracted from Uzbek texts,” Data Brief, vol. 43, p. 108351, 2022.

8. X. Madatov, M. Sharipov, and S. Bekchanov, “O‘zbek tili matnlaridagi nomuhim so‘zlar,” *Computer linguistics: problems, solutions, prospects*, vol. 1, no. 1, 2021.

9. M. Sharipov and O. Sobirov, “Development of a rule-based lemmatization algorithm through Finite State Machine for Uzbek language,” arXiv preprint arXiv*:2210.16006*, 2022.

10. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a morphological and syntactic tagged corpus for the Uzbek language,” arXiv preprint arXiv:2210.15234, 2022.

11. E. Kuriyozov, S. Matlatipov, M. A. Alonso, and C. Gómez-Rodriguez, “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek,” in Human Language Technology. Challenges for Computer Science and Linguistics: 9th

12. Language and Technology Conference, LTC 2019, Poznan, Poland, May 17–19, 2019, Revised Selected Papers, 2022, pp. 232–243.

13. U. Salaev, E. Kuriyozov, and C. Gómez-Rodriguez, “A machine transliteration tool between Uzbek alphabets,” in CEUR Workshop Proceedings, 2022, pp. 42–50. [Online]. Available: www.scopus.com

14. M. Sharipov and U. Salaev, “Uzbek affix finite state machine for stemming,” arXivpreprint arXiv:2205.10078, 2022.

15. Tiedemann J. “Parallel Data, Tools and Interfaces in OPUS,” in Proc. of the 5 th International Workshop on Language Resources and Evaluation, 2012, pp. 221-225.

16. Goldwasser D., “Statistical Machine Translation: A Survey,” Journal of Computer Science, vol. 56, № 3, 2013, pp. 450-463.

“ABDURAUF FITRAT KORPUSI” LEKSIKASINING STATISTIK TAHLILI

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU
Alimbekova Mavjuda Xalimjon qizi,
tayanch doktorant
alimbekovamavjuda18@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada “Abdurauf Fitrat korpusi” matnlarining leksikasi va ularning qatlamlari statistik tahlili yoritilgan. Korpus matnlari asosida so‘zshakllar korpus yordamida shakllantirilib, ular asosida lemmalar qo‘lda lemmatizatsiya qilingan va qo‘lda leksik qatlamiga ko‘ra teglangan. Teglangan lemmalar bazasi korpusga yuklanib, muallif ijodida qo‘llangan o‘z va o‘zlashgan qatlam so‘zlari aniqlandi. Muallif ijodida qo‘llangan so‘zshakllarning leksik qatlami har bir asarda ko‘rsatilgan. Ular uslub va janrlar kesimida statistik tahlil qilindi.

Abstract. This article covers the lexicon of the texts of the “Abdurauf Fitrat Corpus” and the statistical analysis of their layers. Based on the corpus texts, word forms were formed using the corpus, and lemmas were manually lemmatized and manually tagged according to the lexical layer. The base of tagged lemmas was loaded into the corpus, and the words of the original and borrowed layer used in the author’s work were identified. The lexical layer of word forms used in the author’s work is indicated in each work. They were statistically analyzed in terms of style and genre.

Аннотация. В данной статье рассматривается лексикон текстов “Корпуса Абдурауфа Фитрата” и статистический анализ их слоев. На основе текстов корпуса были сформированы словоформы с использованием корпуса, а леммы были вручную лемматизированы и вручную размечены в соответствии с лексическим слоем. База размеченных лемм была загружена в корпус, и были выявлены слова исходного и заимствованного слоя, используемые в работе автора. В каждой работе указан лексический слой словоформ, используемых в работе автора. Они были статистически проанализированы с точки зрения стиля и жанра.

Kalit so‘zlar: *Abdurauf Fitrat korpusi, lemma, leksik qatlam, o‘z qatlam, o‘zlashgan qatlam.*

Kirish. Fitrat asarlari tilining xususiyatlari, leksikasini tadqiq etish va lug‘atlarini yaratish tilshunoslikning muhim vazifalaridan biridir. Abdurauf Fitrat

milliylikka katta e’tibor bergan va bu asarlarida sezilarli darajada aks etgan. Sof turkiy va shevaga xos so‘zlardan foydalanish bilan birga, so‘z yasalishiga ham e’tibor berib, tilimizga yot bo‘lgan – o‘zlashma qo‘shimchalardan ham foydalanmaslikka harakat qilgan. Shu sababli Abdurauf Fitrat asarlari tilini tadqiq etish muhim sanaladi. Fitrat asarlari korpusini yaratish o‘sha davr tilini tadqiq etish, lug‘atlarini yaratish, so‘zlarda kuzatiladigan lingvistik hodisalarni qiyosiy-tarixiy usulda o‘rganishga yordam beradi. Har bir tilning leksikasi ikkiga: o‘z qatlam va o‘zlashma qatlamga bo‘linadi. O‘zbek tili lug‘at boyligi turkiy, arabiy, forsiy, yevropaviy qatlamlardan iborat. O‘zbek tilining o‘z qatlamini o‘zbek va boshqa turkiy tillar uchun umumiy bo‘lgan, qadimgi davrlardan buyon ishlatilib kelinayotgan so‘zlar, o‘zbek shevalariga xos so‘zlar, o‘zbek tilining ichki imkoniyati asosida yasalgan so‘zlar tashkil qiladi.

Adabiyotlar sharhi

Dunyo tilshunosligida korpus yordamida muallif asarlari tilining tadqiqiga bag‘ishlangan ko‘plab ishlarni uchratish mumkin. O.V.Kukushkina, A.A.Polikarpov, E.V.Surovsevalar tomonidan yaratilgan A.P.Chexov badiiy asarlari tilining chastotali grammatik-semantik lug‘ati [1] bu sohadagi tadqiqotlarning yorqin namunasidir. Dostoyevskiyning muallif korpusi asosida bir qator lug‘atlar, jumladan, muallifning chastotali lug‘ati [2] va A.Ya.Shaykevich tomonidan yaratilgan ideomatik so‘zlar bazasi [3] yaratilgan.

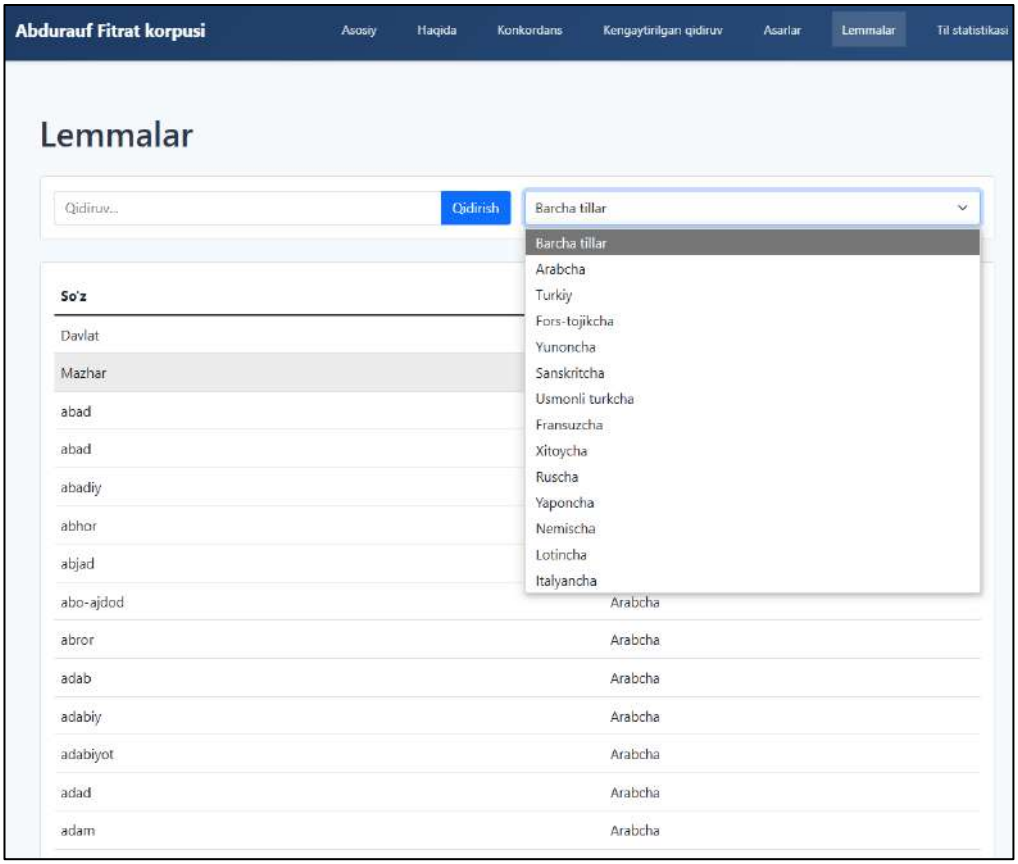
O‘zbek tilshunosligida Sh.Hamroyeva o‘zbek tili mualliflik korpusining lingvistik asoslarini yaratdi va Abdulla Qahhor mualliflik korpusiga asos soldi [4]. N.G‘ulomovaning “Alisher Navoiy muallif korpusi va uning semantik teglar ma’lumotlar bazasi (“Badoye’ ul-vasat” devoni asosida)”[5] nomli tadqiqoti hamda M.A.Abjalova hammuallifligida yaratilgan Alisher Navoiy, N.S. G‘ulomova, Sh.M. Sa’dullayeva muallif korpusi [6] bu sohada yaratilayotgan ishlar uchun namuna bo‘lib xizmat qiladi.

Tadqiqot metodologiyasi

Ushbu tadqiqot ishida statistik tahlilo metodidan foydalanildi. Abdurauf Fitrat asarlarida qo‘llangan so‘zshakllar korpus yordamida aniqlanib, ular asosida lemmalar qo‘lda shakllantirildi. Lemmalarning leksik qatlami o‘rganilib, qo‘lda teglandi va “Abdurauf Fitrat korpusi”ga yuklandi. Olingan natijalar statistik tahlil qilindi.

Tahlil va natijalar

“Abdurauf Fitrat korpusi” [7] yaratilib, teglangan lemmalar bazasi korpuysga yuklandi. Korpusning “Lemmalar” bo‘limida umimiy qidiruv va tillar bo‘yicha qidiruv imkoniyatlari yaratildi. (1-rasmga qarang).



1-rasm. “Abdurauf Fitrat korpusi”da lemmalar qidiruvi

Fitrat ijodida jami 9157 ta lemma aniqlandi. Ulardan sof turkiy lemmalar soni 5192 ta ekani aniqlandi. O‘zlashgan qatlamdan eng ko‘pi arabcha so‘z 2505 ta, fors-tojikcha so‘z 1039 tani tashkil etdi. Shuningdek, yunoncha, lotincha, italyancha, ruscha, xitoycha, nemischa, sanskritcha so‘zlar ham mavjud:

1) yunoncha: *akademiya, anatomiya, aristokrat, astronomiya, azot, daftar, difteriya, dio‘lo‘g, do‘qtur, drama, energiya, falsafa, fantaziya, faylasuf, faytun, geometriya, gigiena, gram, hipnoz, injil, jo‘g‘rofiya, kapitalizm, kapitol, koinot, komediya, lirik, lirika, mikroskop, lirizm, mo‘no‘lo‘g, neft, papirus, parazit, piragramma, plato‘nizm, poetik, po‘p, sistema, teatr, ritm, telefon, teleg‘rof, tip, tragediya, angel;*

2) fransuzcha: *adres, bilet, bonka, burjuaziya, dinamit, ekspluatasiya, idealizm, jandarm, jurnal, kabinet, kamzul, magazin, marsh, miliard, milio‘n, mo‘da, olmon, opportunist, opportunizm, partiya, qo‘mita, raman, ratsionalizm, reforma, reformistik, ro‘l, ro‘mantizm, ro‘mon, ro‘montizm, ro‘montik, sentimentalizm, shovinizm, simvolist, simvolizm, so‘fa;*

3) nemischa: *german, nemis, kapitalist, lampa, mashruta, soldat, tans, yarmarka;*

4) lotincha: *agent, agrar, detsimetr, do‘kto‘r, element, fabrika, fakt, familiya, federatsa, feo‘dal, feo‘dalizm, forma, general, imperator, individ, instinkt, institut,*

iyul, jinerol, kamandir, kanal, kislorod, klassik, klassistizm, kollektiv, kolxoz, komisiya, komissar, kommunist, ko‘nsul, ko‘piya, mandat, massa, material, materiyol, materialist, metrika, miqro‘sqo‘f, mobilizm, mo‘dernizm, muzey, naturalizm, nota, no‘mir, no‘yabr, ognistizm, operasiya, pochta, protestant;

5) italyancha: *banka, gazet, gazit, opera;*

6) ruscha: *etnograficheskaya, gosudarstvennogo, nauk, bolshevik, bronepoyezd, cholvor, choynak, choynak, elektrik, fabrikant, futurist, galos, grajdan, harakteristika, idealistik, ideolog, imperialist, istikan, kalisa, karbon, kartochka, komik, kopiya, marksizm, mashina, mikrob, miqro‘b, mistik, muzika, obraz, ochirat, o‘ris, o‘rus, patnus, pechatniy, plan, poyezd, proletariat, rabo‘chi, rumka, samovar, shapka, shkaf, shlyapa, skoriy, skripka, slovo, spiska, stansiya, sud, syezd, ustul, uyezd, valisapid, velosiped, vintovka, xaraktaristika, zakon;*

7) usmonli turkcha: bebak, eksik, gizlamak, indirmoq, kandi, moso, orqadosh, o‘rtu, saqin, sayg‘i, sevdik, sodacha, terazi, yumruq.

8) xitoycha: *choy;*

9) sanskritcha: *baxshi, rupiya, buddizm, dakki, fo‘ta, varat.*

10) inglizcha: *boykot, folklor, gurji, ingliz, park.*

11) mo‘g‘ulcha: *anfar, bo‘sag‘a, inju, juvanvar, lochin, manglay, mergan, o‘lja, qurultoy, suyurg‘ol, tengri, ulus, yosog‘.*

O‘zlashgan so‘zlardan eng ko‘pi arabcha qatlamga mansub, ular 2506 tani tashkil qiladi. Fors-tojikcha so‘zlar miqdori 1042 tadan iborat.

Ushbu teglangan 9157 ta lemma “Abdurauf Fitrat korpusi”ga yuklandi va bu teglangan lemma orqali korpusi matnlaridagi 60185 ta so‘zshaklning qatlami aniqlandi, statistikasi chiqarildi (1-jadvalga qarang). 1-jadvalda faqat lemmalarning leksik qatlami bo‘yicha statistikasi ko‘rsatilgan. So‘zshakllar bo‘yicha leksik qatlam statistikasi har bir asarda alohida ko‘rsatilgan.

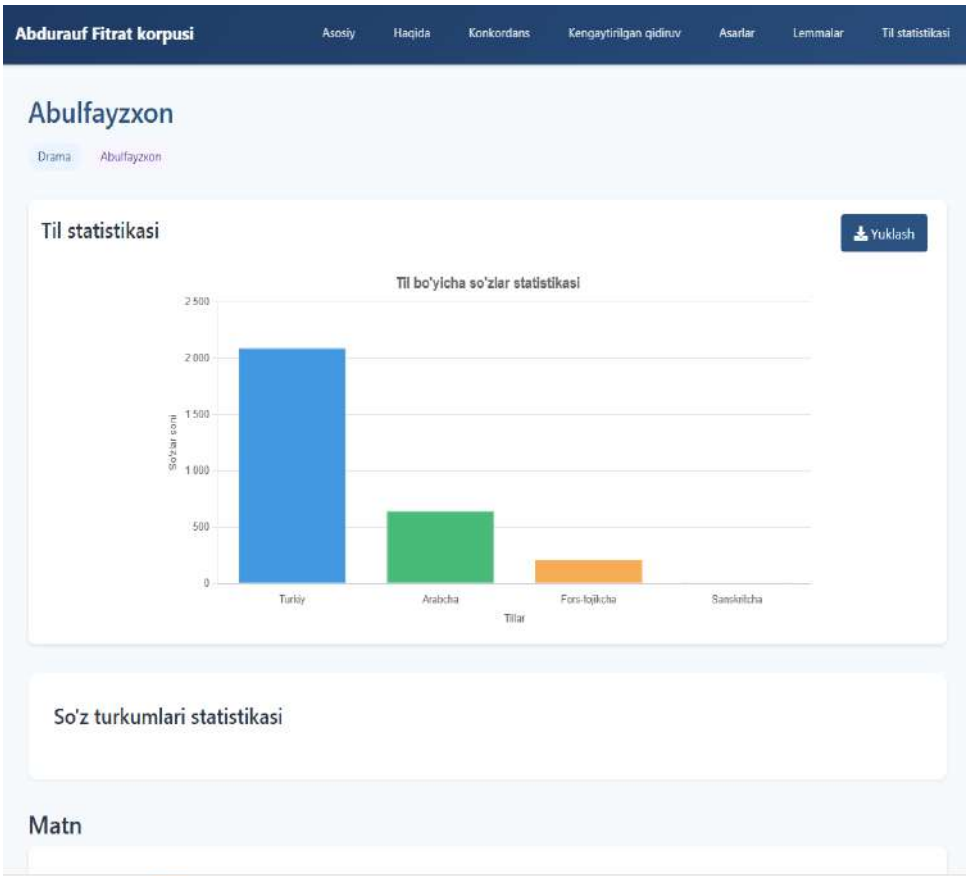
1-jadval. “Abdurauf Fitrat korpusi” lemmalarida leksik qatlam statistikasi

Til	**Miqdor**	**Foiz**
Turkiy	5192	56,7%
Arabcha	2506	27,4%
Fors-tojikcha	1042	11,4%
Noma’lum qatlam	118	1,3%
Lotincha	66	0,7%
Ruscha	66	0,6%
Yunoncha	53	0,4%
Fransuzcha	39	0,2%
Mo‘g‘ulcha	13	0,2%
Usmonli turkcha	14	0,1%
Nemischa	9	0,1%
Sanskritcha	6	0,1%

Inglizcha	5	0%
Italyancha	4	0%
Xitoycha	1	0%
Yaponcha	1	0%

Ushbu jadval ma’lumotlaridan ko‘rinib turibdiki, turkiy lemmalarning soni 5192 tani tashkil etgan. O‘zlashgan qatlamga oid so‘zlarning asosiy qismini arabcha va fors-tojikcha so‘zlar tashkil qilgan. Ruscha, lotincha, yunoncha, fransuzcha lemmalar terminlar, ilm-fan taraqqiyoti natijasida kirib kelgan so‘zlardir. Nemischa, italyancha, sanskritcha, xitoycha, yaponcha so‘zlar juda kam miqdorni tashkil etadi. Leksik qatlami aniqlanmagan so‘zlar 118 tani tashkil etdi. Ular “noma’lum qatlam” nomi ostida birlashtirildi.

Fitrat asarlarining qaysi janrlarida o‘z va o‘zlashgan qatlam so‘zlarining ulushi qancha ekani korpus yordamida aniqlandi. Har bir janrning so‘zshakllari bo‘yicha leksik qatlami aniqlanib, bar graph ko‘rinishida statistikasi chiqarildi. (2-rasmga qarang).



2-rasm. “Abulfayzxon” asarida qo‘llangan so‘zshakllar leksik qatlami

Ushbu rasmda ko‘rsatilgani kabi muallifning har bir asaridagi so‘zshakllar va ularning leksik qatlami diagramma shaklida berilgan. Korpusga yuklangan muallifning barcha asrlaridagi til statistikasi asosida janrlar va uslublar kesimida leksik qatlamning ifodalanishi statistikasi ko‘rsatilgan. (2-jadvalga qarang). Bunda o‘z qatlamga oid so‘zlarni turkiy so‘zlar deb nomladik va o‘zlashma qatlamga oid so‘zlardan eng ko‘p ko‘rsatkichga ega bo‘lgan arabcha, forscha so‘zlar olingan. Boshqa kam sonli so‘zlardan iborat qatlamlar olinmadi.

2-jadval. “Abdurauf Fitrat korpusi”da so‘zshakllarning leksik qatlami janr va uslublar kesimida

Badiiy uslub matnlarida leksik qatlamning aks etishi			
janrlar	Turkiy	Arabcha	Fors-tojikcha
she'r	364	60	29
g'azal	8		1
mansura	206	34	6
qissa	3694	2780	732
hikoya	1728	581	217
drama	12 696	3402	1172
jami	18696	6857	2157
Ilmiy uslub matnlarida leksik qatlamning aks etishi			
ocherk	1278	591	157
ilmiy maqola	7317	4715	1144
darslik	1726	358	66
majmua	296	153	18
risola	5249	3699	780
qo'llanma	6571	2630	669
jami	22437	12146	2834
Publitsistik uslub matnlarida leksik qatlamning aks etishi			
publitsistik maqola	1683	985	126
maktub	564	393	37
risola	14086	10928	2495
munozara	2168	2410	574
jami	18501	14716	3232

Badiiy uslub janrlarida muallif turkiy so‘zlardan ko‘p foydalangan. Arabcha va fors-tojikcha so‘zlar boshqa uslubdagi ushbu qatlam so‘zlariga nisbatan ancha kam qo‘llanilgan. Badiiy uslub matnlarida turkiy so‘zlar 18696 ta, arabcha so‘zlar 6857 ta, fors-tojikcha so‘zlar 2157 tani tashkil qiladi. Badiiy asar xalq tilining boyligini, go‘zalligini ifodalovchi uslubdir. Unda xalq shevalari, xalq tiliga xos bo‘lgan ifodalar ko‘pligi bilan boshqa uslubdan ajralib turadi. Fitratning badiiy uslubda yozilgan asarlarida turkiy so‘zlarning ko‘pligidan ham bilish mumkin. Badiiy uslub janrlaridan qissa janrida muallifning “Hind sayyohi bayonoti” qissasi nomli asari mavjud. Ushbu asarda badiiy uslubning boshqa janrlariga nisbatan

arabcha so‘zlar ko‘proq qo‘llanilgan. Buning sababi qissaning qahramoni tilidan hadislar, oyatlar keltirilgan o‘rinlar ko‘p uchrashidadir.

Ilmiy uslub janrlarida qo‘llanilgan o‘z qatlamga mansub so‘zlarning soni 22437 tani tashkil qiladi. Arabcha so‘zlar 12146 ta, fors-tojikcha so‘zlar esa 2843 ta. Ilmiy uslub matnlarida arabcha, fors-tojikcha o‘zlashmalarning ko‘p uchrashining asosiy sababi arabcha terminlarning mavjudligi, Fitrat tomonidan o‘rganilgan shoir-yozuvchilarning ijodida qo‘llanilgan so‘zlarning ko‘pligidir.

Publitsistik uslub janrlarida 18501 ta turkiy so‘z, 14716 ta arab tilidan o‘zlashgan so‘z, 3232 ta fors-tojikcha so‘zlar qo‘llanilgan. Ushbu uslub matnlarida arabcha so‘zlar boshqa uslub matnlaridagiga nisbatan ko‘proq qo‘llanilgan. Publitsistik uslubda yozilgan “Oila”, “Rahbari najot” nomli asarlarida arabcha kalimalar, oyatlar, hadislar ko‘p qo‘llanilgan. Millat ravnaqi uchun, oila mustahkamligi uchun yagona najot yo‘li islom dini ekani yoritilgan ushbu asarlarda arabcha so‘zlarning ulushu katta bo‘lishi tabiiy. Munozara janrida yozilgan “Hindistonlik bir farangi ila buxorolik mudarrisning jadid maktablari xususinda qilg‘an munozarasi” nomli asarida boshqa janrlardan farqli tarzda arabcha so‘zlar turkiy so‘zlardan ko‘p qo‘llanilgan. Turkiy so‘zlar 2168 ta, arabcha so‘zlar 2410 ta, fors-tojikcha so‘zlar 574 tani tashkil qilgan. Ushbu asarda hindistonlik sayyoh va buxorolik mudarris suhbatida Qur’on oyatlari, hadislar, arabcha jumlalar ko‘p qo‘llanilgan. Bu arabcha so‘zlarning ulushi turkiy so‘zlardan ko‘p bo‘lishiga sabab bo‘lgan.

Xulosa. Abdurauf Fitrat ijodida jami 9157 ta leksik birlikdan foydalangani ma’lum bo‘ldi. Ulardan o‘z qatlamga mansub lemmalarning soni 5192 tani tashkil etadi. O‘zlashgan qatlamga oid so‘zlarning asosiy qismini arabcha va fors-tojikcha so‘zlar tashkil qiladi. Arabcha so‘zlar 27.4 ta, fors-tojikcha so‘zlar 11.4 foiz, boshqa talam so‘zlari 5 foizni o‘z ichiga oladi. Ruscha, lotincha, yunoncha, fransuzcha lemmalar terminlar, ilm-fan taraqqiyoti natijasida kirib kelgan so‘zlardir. Nemischa, italyancha, sanskritcha, xitoycha, yaponcha so‘zlar juda kam miqdorni tashkil etadi.

Badiiy uslub janrlarida muallif turkiy so‘zlardan ko‘p foydalangan. Arabcha va fors-tojikcha so‘zlar boshqa uslubdagi ushbu qatlam so‘zlariga nisbatan ancha kam qo‘llanilgan. Badiiy uslub matnlarida turkiy so‘zlar 18696 ta, arabcha so‘zlar 6857 ta, fors-tojikcha so‘zlar 2157 tani tashkil qiladi. Ilmiy uslub janrlarida qo‘llanilgan o‘z qatlamga mansub so‘zlarning soni 22437 tani tashkil qiladi. Arabcha so‘zlar 12146 ta, fors-tojikcha so‘zlar esa 2843 ta. Ilmiy uslub matnlarida arabcha, fors-tojikcha o‘zlashmalarning ko‘p uchrashining asosiy sababi arabcha terminlarning mavjudligi, Fitrat tomonidan o‘rganilgan shoir-yozuvchilarning ijodida qo‘llanilgan so‘zlarning ko‘pligidir. Publitsistik uslub janrlarida 18501 ta turkiy so‘z, 14716 ta arab tilidan o‘zlashgan so‘z, 3232 ta fors-tojikcha so‘zlar qo‘llanilgan.

Foydalanilgan adabiyotlar:

1. Кукушкина О.В., Суровцева Е.В., Лапонина Л.В. Частотный грамматико-семантический словарь языка художественных произведений А.П.Чехова с электронным приложением. – М.: МАКС Пресс, 2012. – 571 с.
2. Заботкиной В.И. “Методи когнитивного анализа семантика слова компютерно-корпусной подход”, Москва: Язика словянской культури, п. 348, 2015.
3. Словар язика Достоевского. Идеоглоссарий.// Российская академия наук институте руского язика им В.В.Виноградова, 2008.
4. Hamroyeva Sh.M. “O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari”, Buxoro, – 259 bet, 2018.
5. G‘ulomova N.S. “Alisher Navoiy mualliflik korpusi va uning semantik teglari bazasini yaratish (“Badoye’ ul-vasat” devoni asosida)”, Toshkent, – 190 bet, 2022.
6. https://navoiykorpusi.uz/
7. https://fitratkorpusi.uz/

QORAQALPOQ TILI SO‘ZLARINI LEMMALASH UCHUN LINGVISTIK MANBA YARATISH

Yusupov Firnafas
Texnika fanlari nomzodi, dotsent
firnafas@mail.ru
TATU Urganch filiali

Sharipov Maqsud Siddiqovich
Texnika fanlari nomzodi, dotsent
maqsbek72@gmail.com
Urganch davlat universiteti

Urazbayeva Nilufar Oybek qizi
urazbayevanilufar724@gmail.com
UrDU magistranti

Annotatsiya. Mazkur maqolada qoraqalpoq tilidagi so‘zlarni lemmalash uchun zarur bo‘lgan lingvistik manbani yaratish masalasi ko‘rib chiqiladi. Unda qoraqalpoq tilining leksik xususiyatlari, affiksal tizimi, shuningdek, so‘zlarning lemmalash jarayonidagi o‘rni tahlil qilinadi. Leksik tuzilmalarni aniqlash, lug‘aviy bazani shakllantirish lemmalashning muhim bosqichlari sifatida ko‘rsatiladi. Mualliflar qoraqalpoq tiliga oid leksik va grammatik axborotlarni tizimlashtirish orqali, tabiiy tilni qayta ishlash, mashina tarjimasi va sun’iy intellekt sohalarida foydalanish mumkin bo‘lgan lingvistik manbani ishlab chiqish zarurligini asoslab beradilar. Tadqiqot natijalari qoraqalpoq tilining raqamli muhitda rivojlanishiga hissa qo‘shadi.

Annotation. This article discusses the issue of creating a linguistic resource necessary for the lemmatization of words in the Karakalpak language. It analyzes the lexical features and affixal system of the Karakalpak language, as well as the role of words in the lemmatization process. Identifying lexical structures and building a lexical database are presented as important stages of lemmatization. The authors justify the need to develop a linguistic resource that can be used in natural language processing, machine translation, and artificial intelligence by systematizing lexical and grammatical information related to the Karakalpak language. The research results will contribute to the development of the Karakalpak language in the digital environment.

Аннотация. В данной статье рассматривается вопрос создания лингвистического ресурса, необходимого для лемматизации слов каракалпакского языка. Анализируются лексические особенности, аффиксальная система каракалпакского языка, а также роль слов в процессе

лемматизации. Определение лексических структур и формирование лексической базы представлены как важные этапы лемматизации. Авторы обосновывают необходимость разработки лингвистического ресурса, который можно использовать в области обработки естественного языка, машинного перевода и искусственного интеллекта, путем систематизации лексической и грамматической информации, относящейся к каракалпакскому языку. Результаты исследования внесут вклад в развитие каракалпакского языка в цифровой среде.

Kalit so‘zlar. *Qoraqalpoq tili, korpus, korpus tahlili, lingvistik manba, lemma, lemmalash, algoritm.*

Hozirgi kunda qoraqalpoq tilidagi lemmalash jarayoniga oid ilmiy tadqiqotlar va manbalar yetishmayapti. Shu sababli, qoraqalpoq tilidagi so‘zlarni lemmalash uchun lingvistik manba yaratish zarurati tug‘iladi. Ushbu maqolada qoraqalpoq tili so‘zlarini lemmalash uchun kerakli lingvistik manbaning yaratilishi jarayoni, uning ahamiyati, Tabiiy tilni qayta ishlash (ing. Natural Language Processing – NLP) muammolarini yechishga yo‘naltirilgan resurslarga e’tibor qaratish va kutilayotgan natijalari haqida batafsil ma’lumot beriladi. Lemmalash so‘zlarning asosiy shaklini aniqlash va ularga ishlov berish jarayonini ifodalaydi, bu esa to‘liq lingvistik ma’lumotlar bazasini talab etadi. Mazkur ma’lumotlar bazalari korpus deb nomlanadi va u yozma matnlarning istalgan to‘plamidan iborat bo‘lishi mumkin [1]. Lingvistik tahlil dasturlari ushbu korpuslardan tezkor qidiruv va tahlilni amalga oshirish orqali samarali natijalarga erishadi. Maqsadimiz qoraqalpoq tilining raqamli muhitda rivojlanishini ta’minlash, uning lingvistik xususiyatlarini o‘rganish va xalqaro tilshunoslik tadqiqotlariga hissa qo‘shishdir.

Qoraqalpoq tili – turkiy tillar qatoriga kiradi. 1 mlnga yaqin insonlar ushbu tilda so‘zlashadilar. Qoraqalpog‘iston Respublikasida istiqomat qiluvchi qoraqalpoq xalqining ona tilidir. Qoraqalpoq tili – O‘zbekiston Respublikasining tarkibidagi Qoraqalpog‘iston Respublikasining rasmiy tilidir. Bu to‘g‘risida maxsus qonun chiqarilgan. Qonunda qoraqalpoq tili Qoraqalpog‘istonning barcha hududida davlat tili sifatida foydalanishga imkon berilgan. Qoraqalpoq tili - turkiy tillarning qipchoq guruhiga mansub tillardan; qozoq va no‘g‘ay tillari bilan birgalikda qipchoq tillarining qipchoq-no‘g‘ay guruhchasini tashkil etadi. Qoraqalpog‘iston Respublikasining davlat tili (o‘zbek tili bilan birga). Asosan, Qoraqalpog‘istonda, shuningdek, Xorazm, Navoiy, Buxoro viloyatlari hamda Qozog‘iston va Turkmanistonning unga qo‘shni hududlarida, Rossiya Federatsiyasi va Afg‘onistonda tarqalgan. Qoraqalpoq tilida so‘zla-shuvchilarning umumiy soni 883 ming kishidan iborat (o‘tgan asrning 90-yillar o‘rtalari). Qoraqalpoq tili, asosan, 2 ta: shimoli-sharqiy va janubi-g‘arbiy lahjalarga bo‘linadi; bu lah-jalar fonetik jihatdan o‘zaro farqlanadi. Adabiy Qoraqalpoq tili 20-asrning 1-yarmida shimoli-

sharqiy lahja asosida shakllangan. Qoraqalpoq tili boshqa turkiy tillarning aksariyati kabi tipologik xususiyatlari bilan ajralib turadi [2, 3, 4].

Qoraqalpoq tilining o‘ziga xos belgilari: unlilar ohangdoshligi - singarmonizm mavjud, Atlarshmz (otlarimiz). kunler - (kunlar); umumturkiy ch undoshi sh bilan; sh esa s undoshi bilan almashadi, qash (qoch), tas (tosh), bas (bosh). Ayrim so‘zlarda g‘ undoshi o‘rnida v; g undoshi o‘rnida y ishlatiladi, tav (tog‘), tiy (teg) va boshqa Qoraqalpoq tilida, boshqa ko‘pchilik turkny tillarda bo‘lganidek, umumturkiy so‘zlardan tashqari, arab, eroniy, rus tillaridan o‘zlashgan so‘zlar ham anchagina [3, 4].

Hozirgi kunda, tillarning kompyuter yordamida ishlashini o‘rganuvchi sohalarda, lemmalash muhim ahamiyatga ega. Bu maqolada, qoraqalpoq tilida lemmalash algoritmini yaratish uchun zarur bo‘lgan lingvistik manbalarni ishlab chiqish masalasi ko‘rib chiqiladi. Qoraqalpoq tili, turkiy tillar oilasiga mansub bo‘lib, o‘zining leksik, sintaktik va morfologik xususiyatlari bilan ajralib turadi. Bu tilda ko‘plab o‘zgartirishlar va afikslar mavjud, ular so‘zning ma’no va grammatik shaklini o‘zgartiradi. Qoraqalpoq tilining leksikologiyasini o‘rganish, lemmalash algoritmini yaratishda asosiy bosqichdir. Qoraqalpoq tilida fe’llar ko‘p hollarda zamon va shaxs affikslariga ega. Shuningdek, nomlar uchun ko‘plik, egalik, to‘g‘ridan-to‘g‘ri va bilvosita qo‘shish affikslaridan foydalaniladi. So‘zning har bir shakli o‘zining lemmataga keltirilishini talab qiladi, shuning uchun qoraqalpoq tilining morfologik xususiyatlarini batafsil tushunish, algoritmning samaradorligini oshiradi [4, 5].

Lemmalashish (yoki lemmatization) – bu so‘zning morfologik shakllarini uning lemmaga (ya’ni, so‘zning ildiz shakliga) keltirish jarayonidir. Masalan, turk tilida “yazacak”, “yazdı”, “yazıyor” kabi shakllar bir xil lemmaga, ya’ni "yaz" ga keltiriladi. Qoraqalpoq tilida ham so‘zlar turli morfologik shakllarda bo‘lishi mumkin, shuning uchun lemmalash jarayoni ushbu tilning to‘g‘ri tahlil qilinishi uchun zarurdir.

- Inglizcha. *“ran”, “runs”, “running”* -> *“run”*
- O‘zbekcha. *“yozdi”, “yozadi”, “yozyapti”, “yozgan”* -> *“yozmoq”*
- Qoraqolpoqcha. *“yazdi”, “ yazaman”, “yazish” – “yazmaq”*
- Turkcha. *“koşacak”, “koşuyor”, “koşdı” – “koş”*

Lemmalash algoritmini yaratish uchun tilga oid katta hajmdagi lingvistik manbalarni tayyorlash zarur. Bu manbalar quyidagi turlardan iborat bo‘lishi mumkin:

Qoraqalpoq tilidagi so‘zlarning leksik tuzilishini o‘z ichiga olgan lug‘atlar. Ular so‘zlarning lemmallash uchun asosiy manba bo‘lib xizmat qiladi. Buning uchun

Qoraqalpoq tilining 7 tomlik izohli lug‘atini misol keltirishimiz va undan foydalanishimiz mumkin.Ushbu izohli lug‘atda har bir harf uchun alohida ta’rif berilib o‘tilgan va har bir so‘zning ikki va undan ortiq shakllari ham ko‘rsatilgan. Bu lug‘atdagi har bir harfdagi so‘zlarni alohida fayl shaklida yig‘ib chiqishni talab qiladi.

Qoraqalpoq tili uchun lemmalashni amalga oshirishda lingvistik manba yaratishning muhim ahamiyati bor. Bu manbalar tilni avtomatik qayta ishlashda, mashina o‘rganish va sun’iy intellekt tizimlarini yaratishda yordam beradi. Lemmalashish uchun quyidagi manbalarni yaratish mumkin:

Lingvistik lug‘at: Qoraqalpoq tilida lingvistik lug‘at yaratish uchun - so‘zlarning ildiz shakllarini har xil qo‘shimchalardan holi tarzda ko‘rsatib berish. Bu lug‘at tilning leksik va lug‘aviy shakli haqida ma’lumot beradi. Misol uchun, “yozgansizlar” so‘zining qo‘shimchasiz shakli “yoz”, “kitobdagi” so‘zining lemmasi esa “kitob” ekanligi - bularning barchasi lug‘atda o‘z aksini topadi.

Leksik analizator: Lemmalashish jarayonini avtomatlashtirish uchun leksik analizator yaratish zarur. Bu tizim, kiruvchi matndagi har bir so‘zni analiz qilib, uni lemmaga keltiradi. Masalan, "kitoblar" so‘zi "kitob" lemmaga, "yazaman" so‘zi esa "yaz-" ildiziga keltiriladi. Bunday analizator dasturlari Qoraqalpoq tilining xususiyatlarini hisobga olgan holda ishlab chiqilishi kerak.

Sentiment tahlilini amalga oshirish: Lemmalashish orqali tilda ifodalangan xabar yoki fikrlarni o‘rganish imkoniyati oshadi. Masalan, “tezroq” va “tezlik” so‘zlarining lemmalashidan keyin, bir xil semantik ma’no (tezlik) o‘rganilishi mumkin. Bu esa sentiment tahlili kabi sohalarda foydali bo‘ladi.

Tegishli korpuslar yaratish: Qoraqalpoq tilini lemmalash uchun maxsus til korpuslari kerak bo‘ladi. Bunday korpuslar tilning tabiiy materiallarini o‘z ichiga oladi va so‘zning kontekstual o‘zgarishlarini tahlil qilishda yordam beradi. Korpusdagi so‘zlar, iboralar va ularning grammatika qoidalari lemmalashish jarayonini to‘g‘ri amalga oshirishda asos bo‘lib xizmat qiladi.

Xulosa qilib shuni aytish lozimki, lingvistik muammolarni hal qilishda, tarjima dasturlari ishlab chiqishda lingvistik manbaning o‘rni katta. Ushbu maqolada aynan qoraqalpoq tili uchun lingvistik manba yaratishning nazariy va amaliy ma’lumotlari berib o‘tildi. Qoraqalpoq tili uchun lemmalash algoritmini yaratish, turkiy tillarning morfologiyasini va lingvistik xususiyatlarini chuqur tushunishni talab qiladi. Tabiiy tilni qayta ishlashning samaradorligini oshirish uchun qoraqalpoq tiliga oid to‘liq va to‘g‘ri lingvistik manbalar yaratish zarur. Shuningdek, kompyuter lingvistik yondoshuvlar, masalan, mashina o‘rganish va statistik metodlar yordamida, lemmatizatsiya jarayonini yanada takomillashtirish mumkin.

Bu tadqiqot, qoraqalpoq tilining tabiiy tilni qayta ishlash sohasida muvaffaqiyatli rivojlanishiga hissa qo‘shishi mumkin. Lemmalash algoritmining samaradorligi, tilning tabiiy shakllaridan foydalanishda va avtomatik tarjimada katta yordam beradi. Kelajakda ushbu manbani mukammallashtirib, uni turli xil janrdagi matnlarni atroflicha tahlil qilish uchun asosiy manbaga aylantirmoqchimiz.

Foydalanilgan adabiyotlar:

1. Baskakov N. A. Karakalpakskiy yaznk, t. 1 – 2, M. – L., 19 5 1-52;
2. Berdimuratov Ye., Hazirgi zaman qaraqalpoq tilinip lsksikologiyasm, Nokis, 1968;
3. Baskakov N. A., Vvedeniye v izucheniye tyurskix yazmkov, 2-izd., M.. 1969;
4. Berdimuratov Ye. Umarov A. Bekimbetov P., Karakalpakskiy yazj (fonetika, morfologiya, sintaksis), Nukis. 1983.
5. Birgit N. Schlyter. Language Policies in Present-Day Central Asia(ingl.)// International Journal on Multicultural Societies. -UNESCO, 2001. -Iss. Vol. 3, No. 2. -P. 130. -ISSN 1817-4574. Архивировано 10 aprel 2018 года.
6. Қарақалпак тилинин түсиндирме сөзлиги.
7. O‘zbekiston milliy ensiklopediyasi (2000-2005).

“O‘ZBEK BOSHLANG‘ICH SINF MAKTAB KORPUSI”DAN BAZIS SO‘ZLARNI AVTOMATIK AJRATIB OLISH

Madatov Xabibulla Axmedovich,
Fizika-matematika fanlari nomzodi, dotsent,
habi1972@mail.ru
Urganch davlat Universiteti

Xajibaeva Surayyo Maxmudjonovna
o‘qituvchi
surayyo.khajiboyeva@gmail.com
Urganch davlat Universiteti

Xujamov Elyor Jumanazarovich
UrDU talabasi

Annotatsiya – bazis so‘zlarni ajratib olish masalasi til o‘rganuvchilar, maktab o‘quvchilari, shuningdek, lingvistlar uchun ham bir qator qulayliklar yaratadi. Ushbu maqolada “O‘zbek boshlang‘ich sinf maktab korpus”idan bazis so‘zlarni ajratib olish masalasi ko‘rib chiqilgan. Bazis so‘zlar to‘plami, shunday to‘plamki, ushbu to‘plam yordamida o‘zbek tilidagi barcha so‘zlarni ifodalash mumkin bo‘lsin. Ya’ni, bazis so‘zlar o‘zbek tilining asosini tashkil etuvchi so‘z yasalishi va lingvistik tahlil jarayonlarida ishlatiladigan so‘zlardir. Odatda, bunday so‘zlarga kundalik muloqotning asosini tashkil etuvchi umumiy otlar, fe’llar, sifatlar va boshqa nutq qismlari kiradi. Bundan tashqari, bazis so‘zlar asosan ilmiy va badiiy adabiyotlarda qo‘llaniladi. Ushbu ishda yuqori chastotali so‘zlarni aniqlash va solishtirish usullari qo‘llanilgan. Biz o‘zbek tili sinonimlarining izohli lug‘ati va “O‘zbek boshlang‘ich sinf maktab korpus”i asosida yaratilgan yuqori chastotali so‘zlarni aniqlash va bazis so‘zlarni ajratib olish usulini taqdim etamiz.

Abstract – the issue of extracting basis words creates a number of conveniences for language learners, schoolchildren, as well as linguists. This article presents a set of basis words, such a set that can be used to represent all words in the Uzbek language. That is, basis words are words used in the processes of word formation and linguistic analysis that form the basis of the Uzbek language. Typically, such words include common nouns, verbs, adjectives and other parts of speech that form the basis of everyday communication. In addition, basis words are used mainly in scientific and literary literature. In this work, high-frequency method and comparison method were used. We present a method for identifying high-frequency words and extracting basis words, created on the basis of the Explanatory Dictionary of Uzbek Synonyms and Uzbek Primary School Corpus.

Аннотация – вопрос извлечения базовых слов создает ряд удобств для изучающих язык, школьников, а также лингвистов. В данной статье

представлен набор базовых слов, такой набор, с помощью которого можно выразить все слова узбекского языка. То есть базовые слова — это слова, используемые в процессах словообразования и лингвистического анализа, составляющие основу узбекского языка. Обычно к таким словам относятся нарицательные существительные, глаголы, прилагательные и другие части речи, составляющие основу повседневного общения. Кроме того, основные слова в основном используются в научной и художественной литературе. В данной работе используются методы высокочастотной идентификации и сравнения. Мы представляем метод выявления высокочастотных слов и извлечения базовых слов, созданный на основе Толкового словаря узбекских синонимов и Корпуса узбекского языка для начальной школы.

Kalit so‘zlar – *bazis so‘z, sinonimlar lug‘ati, maktab korpusi, yuqori chastotalar metodi, agglyutinativ til.*

I. KIRISH

O‘zbek tili go‘zal va juda boy til. Bugungi kunda, o‘zbek tilini axborot kommunikatsion texnologiyalardan foydalanib, turli xil online izohli lug‘atlar yaratish, tarjima masalalari, umuman, olganda tabiiy tilni qayta ishlash masalalari ustida turli xil ishlar olib borilmoqda. Hissiyotlarni tahlil qilish, nomuhim so‘zlarni topish, matnlarni tasniflash, lemmatizatsiya masalasi, niqoblangan tilni modellashtirish (MLM) kabi jarayonlar o‘zbek tilini tahlil qilishning masalalari hisoblanadi.

“Hozirgi kunda mamlakatimizda yashayotgan, O‘zbekistonimizni yagona va ahil oila deb biladigan turli millat va elatlar vakillari ham o‘zbek tilini o‘rganishga katta qiziqish va istak bildirayotganlari ayniqsa e’tiborlidir”, – deya ta’kidlaydi davlatimiz rahbari Shavkat Mirziyoyev. Ta’kidlanishicha, bugungi kunda jahon miqyosida 50 millionga yaqin kishi o‘zbek tilida so‘zlashadi. Ko‘plab mamlakatlarda bu go‘zal va jozibali til katta ishtiyoq bilan o‘rganilmoqda. Bundan tashqari, Prezidentimizning 2020-2030 yillarda o‘zbek tilini rivojlantirish va til siyosatini takomillashtirish konseptsiyasida xorijliklar uchun o‘zbek tilini o‘rgatuvchi dasturlar yaratish asosiy yo‘nalishlardan biri sifatida ta’kidlangan. Shunga asoslanib, o‘zbek tilini o‘rganuvchi mustaqil o‘rganuvchilar uchun yengillik yaratish bizning asosiy maqsadlarimizdan biridir.

O‘zbek tilini o‘rganuvchilar uchun tilni darajalarga bo‘lish asosida o‘rganish g‘oyasini ilgari sursak, har bitta darajada o‘rganuvchi bilishi zarur bo‘lgan so‘zlar lug‘ati mavjud bo‘lishi zarur. Ya’ni, o‘rganuvchi birinchi darajadagi aytaylik sodda so‘zlarni yod olsa va o‘rgansa, keyin o‘ziga mos va qiziqarli adabiyotni tanlay va o‘qiy oladi, shu asosida matn tuzadi, nutqida bayon qiladi, eshitsa anglaydi va javob qaytaradi. Masalan, 2-sinf o‘quvchisi uchun “kitob” so‘ziga izoh berish zarur bo‘lsa, unga “varaqlardan iborat o‘quv vositasi; darslik, ertak kabi turlari mavjud” deb

ta’riflanadi. Undan katta sinflarda esa “ma’lum matnli varaqlardan iborat, juzlab tikilgan, muqovalangan, hajmi 48 sahifadan kam bo‘lmagan bosma (qadim qo‘lyozma ham) asar” deya ta’rif beriladi. Shunday qilib, bosqichma-bosqich yangi va xilma-xil so‘zlar bilan uning lug‘at boyligi oshib boradi. Ammo, birdaninga o‘z darajasiga mos bo‘lmagan so‘zlarni yod oldirish yoki ulardan foydalanish ko‘nikmasini shakllantirishga harakat qilinsa, o‘rganuvchi ushbu tilni o‘rganishdan zerikadi, so‘zlardan foydanish ham qiyin kechadi. Natijada, tilni yaxshi o‘rgana olmaslik holati kelib chiqadi. Maktab o‘quvchilarida esa, adabiyotga nisbatan qiziqish pasayadi. Yuqoridagi fikrlardan kelib chiqib, o‘zbek tilida bazis so‘zlarni topishning ahamiyatli tomoni shundaki, o‘zbek tili darajalari asosida tayyorgan lug‘at bazis so‘zlarga asoslangan holda bo‘lishi zarur. Ya’ni, ushbu lug‘atni yaratishni bazis so‘zlarni kiritishdan boshlash kerak.

Yana bitta ahamiyatli tarafi shundaki, bazis so‘zlar WordNet asoslarini qurish uchun zarur bo‘ladi. WordNet so‘zlarni semantik munosabatlarga, shu jumladan sinonimlar, giponimlar va meronimlar bilan bog‘laydigan so‘zlar orasidagi semantik munosabatlarning leksik ma’lumotlar ombori. Ya’ni, har bitta sinonimlar, giponimlar va meronimlar asosini bazis so‘zlar tashkil qiladi.

II. BOG‘LIQ ISHLAR

Ushbu maqola [1] ta’lim sifatini oshirish uchun muhim bo‘lgan boshlang‘ich sinf o‘quvchilarining intellektual salohiyatiga moslashtirilgan ta’lim korpusini yaratishni o‘rganadi. Darsliklarning o‘quvchilarning intellektual qobiliyatlari bilan mos kelmasligi tushunishni qiyinlashtiradi va o‘rganishga bo‘lgan qiziqishni kamaytiradi. Talabalarning intellektual salohiyatini baholash uchun korpusdan foydalangan holda, tegishli o‘quv materiallarini o‘quv jarayoniga integratsiyalash, faollikni oshirish mumkin. Tadqiqotda ushbu muammoni samarali hal etish maqsadida O‘zbekiston Respublikasi maktabgacha va maktab ta’limi vazirligi tomonidan tasdiqlangan 35 ta o‘zbek tilidagi darsliklardan ishlab chiqilgan korpusga e’tibor qaratilgan. Nomuhim so‘zlarni filtrlash matn so‘rovlarini qayta ishlashda hal qiluvchi protsedura bo‘lib, ular keng ma’lumotlar to‘plamlari ichida ma’lumotlarni qidirish uchun ishlatiladi. Bu jarayon semantik ma’noni yo‘qotmasdan qidiruv maydonini qisqartirish imkonini beradi. Faqat grammatik rolga ega bo‘lgan va so‘rovning axborot mazmuniga hissa qo‘shmaydigan nomuhim so‘zlar, shunga qaramay, so‘rovning umumiy murakkabligiga hissa qo‘shadi. Ushbu muammoni hal qilish uchun ishlatiladigan mavjud matematik modellar tabiiy tillarning barcha toifalariga taalluqli emas[2]. Tabiiy tilni qayta ishlashda ma’lumotlarni yig‘ish va matnni tahlil qilish jarayonida nomuhim so‘zlarini bilish kerak. Bu usul agglyutinativ xususiyatga ega bo‘lgan tillarga nisbatan qo‘llaniladi. O‘zbek tili ham agglyutinativ til hisoblanadi. Bu ish[3] o‘zbek matnlaridagi nomuhim so‘zlarni yoki ko‘pchilik nomuhim so‘zlarni baholash usulini aniqlashga bag‘ishlangan. Axborot

almashuvi va axborotlar tez sur’atlar bilan o‘sib borayotgan bugungi davrda biz turli sohalarga oid barcha ma’lumotlarni o‘qish imkoniyatiga ega emasmiz. Muayyan sohaga oid ma’lumotlarni o‘qish va tahlil qilish ham ko‘p vaqt talab etadi. Ushbu muammoni hal qilish uchun maqolada matnni umumlashtirish taklif etiladi [4]. Unda o‘zbek tili uchun tajriba sifatida umumlashtirish topshirig‘i mavjud bo‘lib, metodologiya TF-IDF algoritmi asosida matnni abstraktlashtirishga asoslangan. O‘zbek tili bilan bog‘liq ishlar va manbalar nuqtai nazaridan shuni e’tirof etish kerakki, uning tabiiy tilni qayta ishlash (NLP) nuqtai nazaridan kam resursli til maqomiga ega bo‘lishiga qaramay, so‘nggi paytlarda ilmiy-tadqiqot ishlari jadal sur’atlarda bormoqda. Bu tilning NLP qamroviga ko‘plab yangi qo‘shimchalar, jumladan transliteratsiya vositalari[5], hissiyotlarni tahlil qilish vositalari[6] va o‘zbekcha matnni umumlashtirish algoritmi[7] qo‘shilishi natijasida paydo bo‘ldi.

O‘zbekiston aholisi orasida 2 millionga yaqin kishi resurslari kam bo‘lgan qoraqalpoq tilida so‘zlashadi, bu esa nomuhim so‘zlarini ajratib olishda turli qiyinchiliklar tug‘dirmoqda. Ushbu muammo ushbu maqolada tasvirlangan ma’lumotlar to‘plami tomonidan hal qilinadi[8]. Shunga qaramay, u o‘zbekcha nomuhim so‘zlari ro‘yxati yordamida sinovdan o‘tkazildi. Mahalliy olimlar o‘zbek tilidagi matnlarning til korpusini yaratish borasida ham muayyan ishlarni amalga oshirdilar. Ushbu maqolada [9] biz o‘zbek tilining sintaktik va morfologik tegli korpusini yaratish uchun “Nutqning bir qismi” (POS) va sintaktik teglar to‘plamini ishlab chiqish orqali bu bo‘shliqni to‘ldirishga harakat qildik [10]. Tabiiy tilni qayta ishlash texnologiyasidan samarali foydalanish kam resursli tilda so‘zlashuvchi aholi uchun juda muhimdir. Aytish joizki, o‘zbek tili uchun Aspect-Based Sentiment Analysis (ABSA) vositalarini yaratish uchun ommaga ochiq, yaxshi o‘rnatilgan lingvistik manbalar mavjud emas. Ushbu ishda yuqorida aytib o‘tilgan bo‘shliqni yopish maqsadida UzABSA, birinchi yuqori sifatli izohli ABSA ma’lumotlar to‘plami taqdim etilgan. O‘zbek tili aglyutinativ til bo‘lganligi uchun qo‘shimchalar soni ko‘p bo‘lib, qo‘shimchalar qo‘shilishi bilan so‘zlar yasaladi. Shu sababli biz so‘zning o‘zagini topishda bir qancha muammolarga duch kelamiz. Ushbu maqolada[11] o‘zbek tilidagi oddiy so‘z shakllarining hech qanday ma’lumotlar bazasini o‘z ichiga olmasdan, affikslarni ajratish usuli bilan o‘zbekcha so‘zlarga o‘zak yasash metodologiyasi taklif qilingan.

III. NAZARIY QISM

Ushbu bo‘limda o‘zbek tili sinonimlarining izohli lug‘ati va O‘zbek boshlang‘ich sinf maktab korpusi(O‘BSMK) yordamida bazis so‘zlarni ajratib olish hamda uning bosqichma-bosqich metodologiyasi haqida ma’lumot berilgan.

A. Ma’lumotlarni tavsiflash va to‘plash

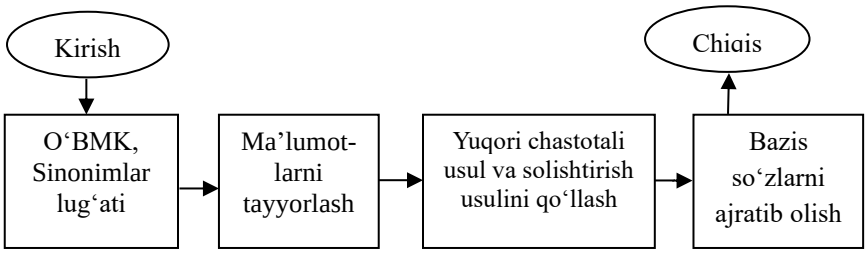
Juda ko‘plab xorijiy tillarda, jumladan, ingliz tilida ham tilni bilish darajasi orqali til o‘rganuvchilar ma’lum bir toifalarga bo‘linadi. Ularning har bir toifasiga qarab so‘zlar bazasi majud bo‘lib, ular yordamida o‘z fikrini, yozishda matn mazmunini to‘liq ifoda qila oladi. Masalan, ingliz tilida ushbu darajalar A1, A2, B1, B2, C1, C2 yoki beginner, pre-intermediate, intermediate, upper-intermediate, advanced, mastery kabilar ko‘rinishida ifodalanadi. O‘zbek tilida tilni bilish darajalari uchun so‘zlar bazasini shakllantirish masalasini ko‘rib chiqdik. Ushbu jarayonni 1-,2-,3-,4- sinf o‘quvchilari uchun mos holatga keltirdik. Ya’ni, birinchi sinf o‘quvchisi nechta so‘z(token)ni bilishi zarur degan savolga javob berdik.

Buning uchun, bizga avvalo asosiy elementlardan biri bo‘lgan maktab korpusi zarur bo‘ldi. Ushbu jarayon ustida O‘zbekiston Respublikasi Maktabgacha va maktab ta’limi vazirligi tomonidan tasdiqlangan 35 ta boshlang‘ich sinf darsliklari qo‘l mehnati asosida tokenlarga ajratildi, turli xil belgilar olib tashlandi va so‘zlar unikal holatga keltirildi. Keyin ularning chastotalari aniqlangan holatda korpusga joylandi.

Yana bitta asosiy elementlardan biri bu sinonimlar lug‘atidir. Biz buning uchun Azim Hojayevning “O‘zbek tili sinonimlar lug‘ati” kitobini oldik, unda sinonimlar bizga kerakli holatda berilgan. Ya’ni, ular sinonimlar guruhlarga ajratilgan shaklda berilgan. Ushbu kitobni pdf formatdan txt formatga o‘tkazdik. Ushbu fayl ustida tozalash ishlari bajarildi, so‘zlar unikal holatga keltirildi.

B. Metod algoritmi

Ushbu ishda biz bazis so‘zlar jadvalini tuzish modelini taklif qildik, u quyidagi bosqichlarni o‘z ichiga oladi: (1) O‘BSMK, Sinonimlar lug‘ati olish; (2) Ma’lumotlarni tayyorlash; (3) Yuqori chastotali usul yordamida yuqori chastotali so‘zlarni topish va solishtirish usuli yordamida korpus uchun so‘zlar guruhini ajratish; (4) Bazis so‘zlar sifatida maksimal chastotali so‘z yoki so‘zlar tanlanadi



Shakl 1. Taklif etilayotgan modelning umumiy ko‘rinishi

C. Ma’lumotlarni tayyorlash.

Har bir sinf uchun bazis so‘zlarning alohida lug‘atini yaratish masalasini ko‘rib chiqildi. O‘zbek tilida O‘BSMK uchun 1,2,3 va 4-sinflar uchun tokenlar

to‘plamini yaratildi. Ushbu korpusda so‘zlar chastotalari bilan keltirildi. Sinonimlar lug‘atida sinonimlar guruhlarga bo‘linadi. Masalan, chiroyli, go‘zal, sohibjamol so‘zlari bir guruhga mansub so‘zlardir. Bir guruhdan so‘z olinib, uni O‘BSMK bilan taqqoslandi. Ushbu lug‘atda a - so‘z bilan boshlangan barcha so‘zlarning chastotalari qo‘shilgan so‘zning chastotasi. Shu tarzda sinonimlar lug‘atidan bir guruhga mansub so‘zlar ichidan eng katta chastotali so‘z olinadi va yangi lug‘at tuziladi. Agar bir nechta maksimal va teng chastotali so‘zlar bir guruhga kirsa, ularning barchasi yangi lug‘atga kiritilgan. Yuqoridagi amallar yordamida bazis so‘zlarni alohida ro‘yxatlar(1-sinf, 2-sinf, 3-sinf, 4-sinf)ga yozib qo‘yamiz.

Olingan natijalar 1-sinf, 2-sinf, 3-sinf va 4 sinflar kesimida [https://zenodo.org/records/13340539] sillkasiga yuklangan.

IV. OLINGAN NATIJALAR

BAZIS SO‘ZLAR SONI

Boshlang‘ich maktab uchun sinflar kesimida bazis so‘zlar soni			
1-sinf	***2-sinf***	***3-sinf***	***4-sinf***
2336	2462	4186	5112

I jadval sinflar kesimida darsliklarda qo‘llanilgan bazis so‘zlarning umumiy sonini ko‘rsatadi.

Yuqoridagi jadvalda keltirilgan natijalarga ko‘ra, 1 sinf maktab o‘quvchisi 2336 ta bazis so‘zlarni bilish orqali gapira oladi, eshitib tushuna oladi, insho yoza oladi va o‘qish ko‘nikmasi shakllanadi. O‘quvchi bu bazis so‘zlarni o‘rganib boshqa so‘zlarni ham ifodalay oladi. Bu 2336 ta bazis so‘zlar O‘BSMKsidagi 1 sinflar uchun 3177 ta lemmalarning ichidan yuqori chastotali so‘zlarni aniqlash orqali kelib chiqqan. 2 sinf uchun 5119 ta lemmalarning ichidan 2462 ta bazis so‘z, 6710 lemmalarning ichidan 4186 ta bazis so‘z, 8029 ta lemmalarning ichidan 5112 ta bazis so‘z mos ravishda ajratib olingan.

V. XULOSA

Xulosa qilib aytganda, har bir sinf maktab o‘quvchilari va o‘zbek tilini o‘rganuvchilar uchun bazis so‘zlar asos bo‘lib xizmat qiladi. Biz bazis so‘zlarni boshlang‘ich sinf maktab darsliklaridagi lemmalar va o‘zbek tili sinonimlar lug‘atidan foydalangan holda yuqori chastotali so‘zlarni topish metodi va solishtirish metodini qo‘llab keltirib chiqardik.

Foydalanilgan adabiyotlar:

1. K. A. Madatov, S. Sattarova, Creation of a Corpus for Determining the Intellectual Potential of Primary School Students, *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, Altai, Russian Federation, 2024, pp. 2420-2423

2. Madatov, K.; Bekchanov, S.; Vičič, J. Automatic Detection of Stop Words for Texts in the Uzbek Language. Preprints 2022, 2022040234.

3. Madatov, K., Bekchanov, S., & Vičič, J. (2022). Accuracy of the Uzbek Stop Words Detection: a Case Study on “School Corpus” CEUR Workshop Proceedings, 3315, 107–115.

4. K. A. Madatov and S. K. Bekchanov, "The Algorithm of Uzbek Text Summarizer," 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, 2024, pp. 2430-2433, doi: 10.1109/EDM61683.2024.10615191.

5. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A Machine Transliteration Tool Between Uzbek Alphabets,” *CEUR Workshop Proceedings*, vol. 3315, pp. 42–50, 2022.

6. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” CEUR Workshop Proceedings, vol. 3315, pp. 126–136, 2022.

7. K. A. Madatov and S. K. Bekchanov, "The Algorithm of Uzbek Text Summarizer," 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2430-2433, 2024.

8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of Karakalpak language stop words,” *Data in Brief*, vol. 48, pp. 109111, June 2023.

9. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language,” *CEUR Workshop Proceedings*, vol. 3315, pp. 93–98, June 2022.

10. S. Matlatipov, J. Rajabov, E. Kuriyozov, and M. Aripov, “UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language,” Proceedings of the 3rd Annual Meeting of the ELRA-ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2024) at LREC-COLING 2024, pp. 394–403, 2024.

11. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” CEUR Workshop Proceedings, vol. 3315, pp. 137–144, 2022.

O‘ZBEK TILI KORPUSI UCHUN TURIZMGA OID MATNLARNING LINGVISTIK BAZASINI YARATISH ASOSLARI

Adizova Nilufar Istamovna,
f.f.f.d., dotsent
nilufaradizova8@gmail.com
Buxoro davlat universiteti

Annotatsiya. Jahon tilshunosligida turizmga oid matnlar va ularning lingvistik, leksik-semantik, diskursiv tadqiqiga doir korpusga asoslangan ilmiy izlanishlarga XX asrning ikkinchi yarmidan boshlab e’tibor kuchaydi. Sun’iy intellekt va turli tizimdagi axborot texnologiya vositalarining hayotimizdan keng o‘rin ola boshlashi bugun bu kabi ishlarning yanada muhim ahamiyat kasb eta boshlaganini ko‘rsatdi. Zotan, turizm dunyo hamjamiyatida mamlakatlar iqtisodiyoti, milliy madaniyati, tarixi hamda ijtimoiy hayotining ajralmas qismi, insoniyat faoliyatining barcha yo‘nalishlariga bevosita daxldor, eng tez rivojlanayotgan, daromadli sohalaridan biridir. Ushbu maqolada turizmga oid matnlar bazasini yaratish asoslari hamda ularning mazmun-mohiyatiga ko‘ra tasniflanishi haqida fikr yuritilgan bo‘lib, Buxoro viloyati misolida ayrim misollar keltirilgan.

Аннотация. Во второй половине XX века в мировой лингвистике усилилось внимание к корпусным научным исследованиям, посвящённым лингвистическому, лексико-семантическому и дискурсивному анализу текстов, связанных с туризмом. Активное внедрение искусственного интеллекта и информационных технологий в повседневную жизнь подчёркивает значимость таких исследований в наше время. Туризм стал неотъемлемой частью экономики, национальной культуры, истории и социальной жизни стран мира и представляет собой одну из самых быстроразвивающихся и прибыльных отраслей, охватывающих все сферы человеческой деятельности. В данной статье рассматриваются основы создания базы туристических текстов и их классификация по содержательно-смысловому признаку. В качестве примера приведены некоторые тексты, относящиеся к Бухарской области.

Annotation. In the second half of the 20th century, global linguistics saw an increased focus on corpus-based scientific research dedicated to linguistic, lexico-semantic, and discourse analysis of texts related to tourism. The active integration of artificial intelligence and information technologies into everyday life highlights the relevance of such research today. Tourism has become an integral part of the economy, national culture, history, and social life of countries around the world, and it represents one of the fastest-growing and most profitable industries, encompassing

all areas of human activity. This article explores the foundations for creating a database of tourism-related texts and their classification based on content and meaning. As an example, several texts related to the Bukhara region are presented.

Kalit so‘zlar: *turistik matnlar, turistik birliklar, konsept, lingvokulturologiya, madaniy obyektlar, turizm, ziyorat turizmi, sayohat, ichki turizm.*

Har qanday davlat, millat o‘zining boy o‘tmishi, farovon turmushi, mustahkam iqtisodiyoti, taraqqiy etayotgan sanoati bilan dunyo hamjamiyatida bo‘y ko‘rsatishga intiladi. Jumladan, O‘zbekiston ham zamonaviy turizm industriyasini rivojlantirish uchun barcha zarur manbalarga ega. Buyuk Ipak yo‘li ustida joylashgan vatanimiz qulay tabiiy iqlim sharoitlariga, boy tarixiy, madaniy merosga va, ayni paytda, ham ichki, ham xalqaro turizmni rivojlantirish uchun yuqori salohiyatga ega. Xalqlar o‘rtasida iqtisodiy, ijtimoiy va madaniy aloqalarning samarali yo‘lga qo‘yilishida insoniyat faoliyatining eng tez rivojlanayotgan sohalaridan biri sifatida turizm sohasining til vositalari orqali ta’riflanishi va tasvirlanishi beqiyos ahamiyat kasb etadi. Ushbu sohada olib boriladigan ishlar, o‘z navbatida, tilning leksik tarkibini boyitish va kengaytirishga ham jiddiy ta’sir ko‘rsatadi.

Ma’lumotlarga qaraganda, dunyoda turizm chuqur tarixiy ildizlarga ega bo‘lgan hodisa, uning ilk uchqunlari dastlab, antik va undan keyingi davrlardan boshlangan. Turizm paydo bo‘lishining dastlabki davrlarida uning asosiy sayohatchilari olimlar, savdogarlar, badavlat kishilar, elchilar bo‘lgan. Ular yangi-yangi yerlarni o‘zlashtirish, moddiy manfaatdorlik, to‘kin va farovon hayotni izlash bahonasida dunyoning eng rivojlangan sivilizatsiya markazlari, iqlimi, tabiati go‘zal bo‘lgan joylari, yuksak madaniyatga ega davlatlarga sayohatlarni amalga oshirganlar.

Keyingi davrlarda sport turizmi, elchilik aloqalari, xalqaro hamkorlik, savdo-sotiq, tovar ayirboshlash, joylarni o‘zlashtirish kabilar bilan bir qatorda muqaddas e’tiqodlar tufayli musulmonlarning Makka, Madina, Buxoro va boshqa shaharlarga, xristianlarning Quddus, Misr, Rim va Arab mamlakatlariga ziyoratlari asosida diniy turizmning ham shakllanganiga guvoh bo‘lish mumkin. Ilk uyg‘onish (Renessans) davrida ta’limiy, ma’rifatga asoslangan sayohatlar ham olib borilgan.

Xo‘sh, “turizm” o‘zi nima? “O‘zbek tilining izohli lug‘ati” da “turizm”ga “*tourisme~tour* (fr.) – *aylanish, sayr, sayohat qilish, dunyoni ko‘rish, bilish, o‘rganish, dam olish maqsadida tashkil etiladigan sayr-sayohat, sayyohlik – sayohat qilish faol dam olish turlaridan biri sifatida inson faoliyatining turli sohalari bilan chambarchas bog‘liq bo‘lgan ikki qarama-qarshi – milliy va xalqaro munosabatlar ta’sirida rivojlanadigan jarayon”, deb berilgan [10]*.

BMTning Turizm tashkiloti ta’rifida esa quyidagicha aks etgan: turizm ijtimoiy, madaniy va iqtisodiy hodisa bo‘lib, bu odamlar o‘zlarining odatdagi muhitidan tashqaridagi mamlakatlar yoki joylarga shaxsiy yoki ish/kasbiy maqsadlar uchun safar qilishi. Shuning uchun, u mehmonlarning faoliyati bilan ham bevosita bog‘liq [3].

Umuman olganda, turizm sohasi inson faoliyatining turli jabhalari bilan o‘zaro bog‘liq holda milliy va xalqaro jarayonlarning ta’sirida vujudga keladi hamda ushbu soha doirasida qo‘llanadigan o‘ziga xos milliy leksik birliklar tizimi shakllanadi.

L.M.Goncharova fikricha, “turizm” tushunchasi muayyan til sohiblari ongida mavjud bo‘lgan “sayohat” lingvomadaniy konsepti bilan uzviy bog‘liq holda predmetlarga emas, balki muayyan voqealarga oid murakkab mental xususiyatli konseptdir. Kontsept – bu har qanday madaniyat uchun muhim va qadrli bo‘lgan, u yoki bu obyekt haqidagi bilimlarni o‘z ichiga olgan va tilda uni nomlovchi va tasvirlovchi leksik birliklar to‘plamiga asoslangan milliy xususiyatga ega bo‘lgan mental ta’limotdir [4]. U muayyan bir millatning madaniy xotirasiga asoslangan bo‘lib, keng semantik maydonga ega bo‘lgan ijod mahsullari(fenomen) dir.

Har qanday millatning o‘ziga xos madaniy merosi, qadriyat va an’analari, urf-odatlari, xulq-atvor normalari hamda milliy xususiyatlari uning tili orqali aks etadi. Shuning uchun ham til va madaniyatni bir-biridan ayro tasavvur etib bo‘lmaydi. Til nafaqat millatning milliy-madaniy qiyofasini aks ettiruvchi, ifodalovchi, balki ularni asrash va rivojlantirishga xizmat qilish asosida boshqalarga yetkazuvchi muhim vosita hamdir.

Dunyo tilshunosligida turizm sohasi bo‘yicha lingvistik tadqiqotlar uning ekstralingvistik va lingvistik omillarining o‘zaro ta’sirini aniqlashga qaratilgan bo‘lib, tahlillar lingvokulturologik, antropotsentrik, pragmatik va kognitiv-diskursiv yo‘nalishlarda amalga oshirilgan. Lingvistik jihatdan leksik xususiyatlari esa, asosan, *terminologik, leksikografik va diskursiv* tomondan tahlil qilingan [8]. Biroq o‘zbek tilida turizmga oid matnlar hali korpus asosida tizimli ravishda o‘rganilmaganligi hamda turizm doirasida qo‘llanuvchi milliy-madaniy birliklar tadqiqiga ehtiyoj borligiga e’tibor qaratsak, ushbu sohaga oid matnlar hamda leksik birliklarning lingvistik va lingvokulturologik yondashuvlar asosida o‘ziga xos xususiyatlari, milliy-madaniy jihatlarini ochib berish fan oldiga muhim vazifalarni belgilab beradi.

Zamonaviy korpuslar – bu ma’lum bir tilda, elektron shaklda jamlangan matnlar to‘plamiga asoslangan axborot-ma’lumot tizimi [7] bo‘lib, bugungi kunda o‘zbek tili korpusi ham tilning lug‘at va grammatik sathi kabi uning muhim tarkibiy qismiga aylanib bormoqda. Jumladan, turizm sohasiga doir matnlar korpusini yaratish va bu kabi matnlarda aks etgan milliy-madaniy birliklarni leksik-semantik

jihatdan tahlil qilish hamda tasniflashga korpus lingvistikasi sohasining dolzarb vazifalaridan biri sifatida qarash mumkin.

Bugungi kunda sayohatga chiquvchi kishilarning soni kun sayin ortib bormoqda. Dunyo bo‘ylab sayohat qiluvchi kishilar o‘sha mamlakatning tarixi, madaniyati, muzeylari, milliy kiyimlari, diqqatga sazovor joylari bilan yaqindan tanishish, mahalliy aholi bilan muloqot qilish, milliy taomlaridan tatib ko‘rish kabilarni maqsad qiladi. Mahalliy sayohatchilar esa mamlakatning so‘lim go‘shalariga dam olish, zavq olish, milliy taomlardan tanovul qilish maqsadida sayohatlar uyushtirishadi.

Sayohatchilar uchun muayyan obyektlar, tarixiy voqealar, mintaqalar, madaniy xususiyatlar, boy o‘tmishga ega bo‘lgan tarixiy shaharlar, tarixiy shaxslar, milliy kalorit to‘g‘risida keng ma’lumotlarni taqdim etish, ular bilan yaqindan tanishtirish maqsadida turizmga oid matnlarni korpus asosida jamlash sayyohlar tashrifini yanada kengaytiradi hamda mamlakat iqtisodiyoti, moddiy bazasining o‘sishi uchun ham muhim ahamiyatga ega bo‘ladi. Ushbu maqsadni amalga oshirishda sayohat qo‘llanmalari, yo‘riqnomalar, reklama matnlari, katalog va broshyuralar va boshqa shu kabi vositalar turli madaniyat vakillari o'rtasida bog‘lanish o‘rnatuvchi muhim axborot manbayi sifatida aniq maqsad bilan harakat qilish, tashrif buyuriladigan joylar to‘g‘risida aniq tasavvur uyg‘otish, o‘zini erkin va ishonchli his qilish, shu bilan bir qatorda, madaniyat va tilga oid xususiyatlarni ham tarqatish imkonini yaratadi.

Biz biror shaxs, narsa yoki voqea-hodisa haqida ma’lumot berar ekanmiz, avvalo, fikrimizni muayyan matnlar asosida tinglovchiga aniq va ta’sirchan yetkazishga harakat qilamiz. Buning uchun maqsadimizga mos birliklardan foydalanamiz. Shunga ko‘ra, turizmga oid matnlarda, asosan, sayohatbop joylar, turli me’moriy obidalar, inshootlar, tarixiy shaxslar, tarixiy voqealar, milliy-madaniy o‘ziga xosliklar haqida ma’lumotlar keltiriladi. Bular esa bevosita tilda mavjud bo‘lgan o‘ziga xos so‘z va ifoda birliklari orqali aks etadi.

E.Y.Novikova fikricha, sayyohlik matnlari, bir tomondan, turistlarni ma’lum bir mintaqaga jalb qilishga qaratilgan bo‘lib, reklama qilish va ko‘ngilocharlik xususiyatlarini namoyon etsa, ikkinchi tomondan, ularga madaniy jihatdan muhim obyektlar, hodisalar, joylar va shaxslar haqida muhim ma’lumotlarni yetkazish, axborot uzatish kabi vazifalarni bajaradi [6].

Yuqorida aytilgan fikrlardan kelib chiqqan holda ”turistik matnlar” tushunchasini *“muayyan tarixiy joylar, osori-atiqalar, ko‘hna obidalar, ziyoratgohlar, diqqatga sazovor joylar, tarixiy shaxslar, ular yashagan manzil-makonlar, milliy buyumlar, milliy taomlar va boshqa milliy xususiyatlar haqida ma’lumot aks etgan, sayohatchilar diqqatini jalb qilish uchun ularning ta’rif-u tasvirlari keltirilgan matnlar yig‘indisi”*, deb xulosalash mumkin.

Buxoro azaldan madaniyat markazi, ilm-ma’rifat o‘chog‘i, ko‘hna va mo‘jizaviy bir maskan sifatida dunyo ahlining diqqat-e’tiborida bo‘lib kelgan. Buni O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev ham shunday ta’kidlagan edilar: *Buxoroning ko‘hna va boqiy shahar, deb atalishida katta ma’no bor. Bu yerdagi qadimiy obidalar, osori-atiqalar, xalqimizning betakror qadriyatlari, beqiyos an’analari asrlar osha yashab kelmoqda... 1993-yilda Buxoro shahrining tarixiy markazi YUNESKOning Butun jahon merosi shaharlari ro‘yxatiga kiritilgani uning tarixiy jihatdan bebaholigi, butun ma’rifatli dunyo uchun madaniy va sivilizatsion ahamiyatining to‘g‘ridan-to‘g‘ri isboti bo‘ldi...[1]*

Shunga ko‘ra, qadim Buxoroning turizm salohiyatidan kelib chiqqan holda u yerda saqlangan va bugungi kungacha ko‘plab tashrif buyuruvchilarning hayratiga sazovor bo‘lib kelayotgan madaniy meros, osori-atiqalar, tarixiy shaxslar, milliy taomlar va boshqa milliy kolorit asosida aks etuvchi matnlarni quyidagicha tasniflash mumkin. Ushbu matnlar chet ellik hamda mahalliy sayohatchilar uchun yo‘riqnoma, qo‘llanma, ko‘rsatma, tavsiya, targ‘ibot kabi vazifalarni bajaradi:

1) obidalarga oid matnlar;
2) kiyim-kechakka oid matnlar;
3) tarixiy shaxslarga oid matnlar;
4) ziyoratgohlarga oid matnlar;
5) inshootlarga oid matnlar;
6) voqelikka oid matnlar;
7) taomlarga oid matnlar;
8) narsa-buyumlarga oid matnlar;
9) geografik hududlarga oid matnlar;
10) bog‘larga oid matnlar;
11) dam olish maskanlariga oid matnlar;
12) muqaddas qadamjolarga oid matnlar.

Masalan, quyidagi matnda Buxoro geografik hudud sifatida ta’rif etiladi: *Sharq durdonasi sanalmish **Buxoroyi sharif** nafaqat qadimiy obidalari, me’moriy yodgorliklari bilan, balki ilm-fan, madaniyat va ma’rifat tarixi bilan ham mashhurdir. Buxoro **“Qubbat-ul islom”** deb ta’riflangan, payg‘ambarimiz tomonidan **“Fohira”** (ko‘rkam, serhasham-O‘TIL, 363) deb nom olgan va “Yo Ollohim! Buxoroni qiyomatgacha o‘z panohingda asra! Buxoroliklar qalbini taqvo bilan pok etgin, muloyim qil! Ikki dunyoda saodatmand ayla!”, deb* duo olgan *yurtdir. Payg‘ambarimizning bu ezgu tilaklari ijobat bo‘lib, Buxoro islomiy ma’rifat nurlari bilan sayqal topdi, yetti saharif shaharlar qatoridan o‘rin oldi. Shu bilan birga, dunyo musulmonlari ziyorat qilishni orzu qiladigan yetti ulug‘ pirlarga makon bo‘ldi.*

Buxoroning dunyoga mashhur muqaddas qadamjolari bo‘lmish “Yetti pir ziyoratgohlari” haqida ham bilishni xohlovchilar, ushbu manzil-makonlarni tavof qilishni orzu qiladigan kishilar juda ko‘p bo‘lib, ular haqidagi ma’lumotlar yozilgan tarqatma va qo‘llanmalar muhim ahamiyatga ega:

Buxoro shahri Sharqning mashhur qadimiy shaharlaridan biri. Tasavvufning buyuk vakillari bu yerda yashab, diniy va ijtimoiy faoliyat olib borib, muborak Buxoroning vujudga kelishi va gullab-yashnashi, ma’naviy tarbiya va hissiy ruhni yuksaltirishga o‘z hissalarini qo‘shganlar.

Yetti pir *– Buxoro viloyatining turli tumanlarida joylashgan, muqaddas ziyoratgohlar majmuasi, naqshbandiya tariqatining ketma-ket yetti halqasi. Bu halqa Xoja Abdulxoliq G‘ijduvoniydan boshlanib, Xoja Orif Revgariy, Xoja Mahmud Anjir Fag‘naviy, Xoja Ali Romitaniy, Xoja Muhammad Boboyi Samosiy, Xoja Sayyid Amir Kulol orqali Xoja Bahouddin Naqshband bilan tugaydi. Buxoroga ziyoratga kelganlarning ko‘pchiligi o‘z safarlarini Naqshbandiya ta’limotiga asos solgan ma’naviy ustoz Xo‘ja Abduxoliq G‘ijduvoniy maqbarasini ziyorat qilishdan boshlaydilar...*

Xoja Abdulxoliq G‘ijduvoniy *(1103-1179) payg‘ambarning o‘ninchi avlodi bilan ma’nan bog‘liq. U kishi G‘ijduvonda tug‘ilgan. Xojagon tariqati asoschisi. O‘z qarashlari va ta’limoti bilan islom dini ma’rifati va madaniyati taraqqiyotiga beqiyos hissa qo‘shgan mutasavvuf. O‘zlarining hikmatnamo qoidalarida insonning o‘zligini anglashi, Vatanga muhabbati aks etgan hamda Ollohga yaqin bo‘lgan holda foydali mehnat qilish, oila qurish, farzandlar tarbiyasi bilan shug‘ullanish, jamiyatga naf keltirish, hunar o‘rganishg chorlaganlar. Xo‘ja Abdulxoliq G‘ijduvoniy davrida maqbara chillaxona vazifasini o‘tagan bo‘lib, u yerda qirq kunlik jazirama issiq davrda odamlar ibodat qilish uchun kelgan. Vaqt o‘tishi bilan muqaddas ibodat joyi ardoqli ziyoratgohga aylandi. Maqbara bugungi namunaviy shaklga ega bo‘lishi uchun bir necha zamonadan omon o‘tib kelishi kerak bo‘ldi. Hozirda Abduxoliq G‘ijduvoniy maqbarasi qadimgi Sharqning bir qismi, yer va osmon orasidagi rishta hisoblanadi.*

Xulosa qilib aytish mumkinki, mamlakatimizda butun dunyo xalqlari orzu qiladigan darajada katta moddiy va ma’naviy imkoniyatlar mavjud. Ulardan bahramand bo‘lgan holda barcha moddiy va ma’naviy resurslarimizdan xalq farovonligi yo‘lida to‘g‘ri foydalanish, ularni targ‘ib qilish hamda kelgusi avlodlar uchun ham asrab-avaylashga e’tibor qaratish lozim. Zotan, madaniyatni tildan foydalanuvchi kishilar yaratadi. Ba’zi millatlarda odatiy me’yor, madaniy xususiyatni ifodalovchi etalonlar boshqasi uchun oddiy muloqot vositasi sifatida namoyon bo‘ladi. Matnlarda o‘z ifodasini topgan milliy-madaniy birliklar har bir xalqning milliy qarashlari asosida o‘ziga xos tarzda shakllanadi va ular mazkur til sohiblarining milliy-madaniy boyligi deb qaraladi. Shuning uchun ham sayohat

qiluvchi insonning tasavvurida muayyan millatning madaniy-tarixiy xotiralari bilan bog‘liq voqealar jonlanadi va o‘sha mentalitetga oid stereotipik so‘z va ifodalar shakllanadi.

Foydalanilgan adabiyotlar:

1. O‘zbekiston Respublikasi Prezidentining “2017 – 2019-yillarda Buxoro shahri va Buxoro viloyatining turizm salohiyatini jadal rivojlantirish chora-tadbirlari to‘g‘risida” gi PQ-2980-son Farmoni\2017-yil, 19-may.
2. Elena Manca, Francesca Bianchi. Tourism and tourist promotion around the world: a linguistic and socio-cultural perspective. Salento-2013.
3. Goncharova Л.М. Концепт «путешествие» и его межкультурная представ ленность в рекламной коммуникации сферы туризма. НИР. Современная коммуникативистика (№ 1, 2012). Москва.
4. Istamovna, Adizova Nilufar. “Leksik birliklarning milliy o‘zlik ruhini ifodalashdagi roli (Erkin Vohidov dostonlari asosida)” / Conferences. 2021.
5. Новикова Э. Ю. Жанровые, ценностные и транслатологические характеристики международного туристического дискурса в немецкой и русской лингвокультурах. Автореф. дисс. канд. филол. наука. - Уфа – 2019. - С.13
6. Qarshiyev Abduvali Berkinovich , Tursunov Muhammadsolih Sa’din o‘g‘li. O‘zbek tili milliy korpusini loyihalash. “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” xalqaro ilmiy-amaliy konferensiya. Vol. 1 №. 01. 2022. B.94
7. Тайави Я. М. Лингвокультурные реалии во французских туристических путеводителях. Дисс. канд. филол. наук. Воронеж, 2017. -С. 25.
8. Istamovna, A. N. Sayohat turizmiga oid matnlarda milliy-madaniy birliklarning leksik-pragmatik xususiyatlari. “Contemporary technologies of computational linguistics”, 2024. 2(22.04), 479-484-b.

O‘ZBEK TILI KORPUSI UCHUN TEGLASH TIZIMINI ISHLAB CHIQISH MASALASI

Abdullayeva Oqila Xolmo‘minovna,
Filologiya fanlari falsafa doktori (PhD)
abdullayevaoqila@gmail.com
ToshDO‘TAU doktoranti DSc

Annotatsiya. Tabiiy tilni qayta ishlash (NLP)da matnlarni morfologik va sintaktik teglash va teglangan korpuslarni yaratish eng muhim vazifalardan biri hisoblanadi. Bugungi kunda universal POS teglar tizimidan ko‘plab tillarda tegsetlar tizimining asosi sifatida foydalanishib, keyinchalik har bir tilga xos xususiyatlarni ifodalovchi qo‘shimcha teglar bilan kengaytirilgan. Ushbu universal POS teglar tizimidan ko‘p tilli teglangan NLP ma’lumot platformalaridan biri hisoblangan Universal Dependencies (UD) loyihasi ham ishlab chiqildi. Ammo barcha tillar o‘z sintaksisi, morfologiyasi va fonetikasi bilan bir-biridan farqlanadi, bu esa POS teg va sintaktik teglar tizimini ishlab chiqish zaruriyatini keltirib chiqaradi. Mazkur maqolada tegsetlar tizimining, teglash modellarining yaratilishi bilan bog‘liq masalalar muhokama qilingan.

Abstract. In natural language processing (NLP), one of the most crucial tasks is the morphological and syntactic tagging of texts and the creation of tagged corpora. Currently, the universal POS (Part-of-Speech) tag system is widely used as the foundation for tagging systems in many languages, which has been subsequently expanded with additional tags to represent language-specific features. The Universal Dependencies (UD) project, one of the multilingual tagged NLP data platforms, was also developed based on this universal POS tag system. However, all languages differ in their syntax, morphology, and phonetics, necessitating the development of specialized POS tag and syntactic tag systems. This article discusses issues related to the creation of tagset systems and tagging models.

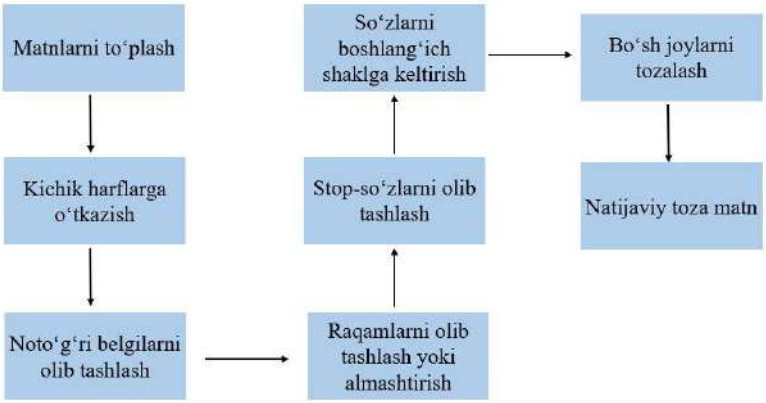
Аннотация. В области обработки естественного языка (NLP) одной из важнейших задач является морфологическая и синтаксическая разметка текстов и создание размеченных корпусов. В настоящее время универсальная система тегов частей речи (POS) широко используется в качестве основы для систем разметки во многих языках, которая впоследствии была расширена дополнительными тегами для отражения специфических языковых особенностей. Проект Universal Dependencies (UD), одна из многоязычных платформ размеченных данных для ОЕЯ, также был разработан на основе этой универсальной системы POS-тегов. Однако все языки различаются по своему синтаксису, морфологии и фонетике, что требует разработки специализированных систем POS-тегов и синтаксических тегов. В данной

статье рассматриваются вопросы, связанные с созданием систем наборов тегов и моделей разметки.

Kalit so‘zlar: *teg, POS teg, sintaktik parsing, parser dasturlar, treebank.*

O‘zbek korpus lingvistikasi sohasidagi so‘nggi tadqiqotlarda o‘zbek tili Milliy korpusini ishlab chiqish, turli maxsus korpus turlarini shakllantirish, korpusda til birliklarini teglash masalasi ko‘rib chiqilmoqda. Korpuslarning yaratilishi bilan bog‘liq tadqiqotlar tahlil qilinganda, ilk morfologik va sintaktik teglash tizimi ishlab chiqilganini kuzatish mumkin. Bu tadqiqotlar XX asrning 50-70-yillariga borib taqaladi. Ilk yirik hajmli til korpusi Broun korpusi ustida amalga oshirilgan va korpus uchun tegsetlar ishlab chiqilgan [1]. Mazkur tadqiqotlar yanada mukammal tegsetlar tizimining, teglash modellarining yaratilishiga zamin bo‘lgan. Keyingi tadqiqotlar tillarning milliy korpuslari yaratilishi va tizimli teglar to‘plamining ishlab chiqilishi bilan bog‘liq bo‘ldi. Shunday tadqiqotlar sirasiga BNC uchun teglashda asos bo‘lib xizmat qilgan CLAWS POS-teglari [2] va statistik hamda qoidaga asoslangan yondashuvlarni birlashtirgan model asosida qurilgan Brill teglarining [3] ustida amalga oshirilgan tadqiqotlar edi. POS teggerlardan keyingi ilmiy izlanishlar morfologik va sintaktik teglarni standartlashtirish va universal teglovchi dasturlarni ishlab chiqish bo‘ldi. Sintaktik teglarni taklif qilgan yirik tizim UD Universal dependency loyihasi bugungi kunda 150 dan ortiq tillar uchun xizmat qilmoqda [4]. So‘nggi tadqiqotlar UD standarti, Universal dependenciesning tahlil qilish modellari, boshqa tillarga moslashuvi, UD grammatikasini ko‘p tilli modellarga integratsiya qilish mavzularida olib borilgan. Agglyutinativ tillarda ham dependency parsing tizimlari ishlab chiqilgan.

Sintaktik parsing va uning algoritmlari bilan ishlashda oddiy matnlar bilan cheklanishning iloji yo‘q. Tabiiy tilga ishlov beruvchi dasturlarni yaratishda xatoliklardan xoli, to‘g‘ri tuzilgan gaplardan iborat toza matnlar to‘plamiga ehtiyoj seziladi. Lekin o‘zbek tilida ishlov berilgan matnlar resursi kam hisoblanadi. Shuning uchun avvalo matnlarni xatoliklardan tozalash va keraksiz belgi hamda kodlarni olib tashlash kerak. Umuman yig‘ilgan matnni tozalash tabiiy tilni qayta ishlashning eng muhim birlamchi jarayoni hisoblanadi. Quyida dasturlar uchun toza matn to‘plamini olishdagi amallarni ko‘rish mumkin.



1-chizma. Matnni dastlabki qayta ishlash bosqichlari

Sintaktik parsing algoritmlarini ishlab chiqishda toza matn bilan cheklanib bo‘lmaydi. Yana qo‘shimcha annotatsiyalash zarur. Parser dasturining maqsadi gap tuzilishini aniqlashdir. Sintaktik tahlil algoritmlari bitta gap ustida amalga oshiriladi. Shuning uchun sintaktik tahlildan oldin bajarilishi kerak bo‘lgan birinchi muhim bosqich – bu gap chegaralarini aniqlashdir. Undan keyingi bosqich sintaktik tuzilmani so‘z yoki tokenlarga (so‘z va tinish belgilari) ajratishdir. Keyingi qayta ishlash – morfologik tahlildir, ya’ni har bir so‘zga asosiy shakl (lemma) va so‘z turkumi, jins, kelishik kabi grammatik ma’lumotlarini belgilovchi morfologik teg qo‘shiladi. Morfologik teg noaniq yoki aniq bo‘lishi mumkin – ikkinchi holatda bu jarayon teglash (tagging) deb nomlanadi. Morfologik tahlil haqidagi ma’lumotlar zamonaviy tahlil vositalari uchun juda muhim hisoblanadi [5]. Biz o‘zbek tili uchun, asosan, https://uznatcorpara.uz/ ga kiritilgan 14 ta POS teglari orqali morfologik tahlil qilmoqdamiz. 12 ta so‘z turkumi yuzasidan POS teglarni kengaytirish va qayta ishlash ustida tadqiqotlar olib borilmoqda. Morfologik va sintaktik teglash tizimi bo‘yicha ko‘plab tadqiqot ishlari olib borilgan. 22 ta tilning o‘zaro o‘xshash turkum va jihatlarini teglovchi universal so‘z turkumlari (Part-of-Speech, POS) to‘plamini ishlab chiqilib, takli qilindi [6]. Bunda aksariyat tillarda mavjud 12 ta so‘z turkumi POS teglari belgilab berilgan. Bunday universal teglar nazoratsiz va tillararo teglash tizimlari ishlab chiqish, turli analizatorlarni (parsers) qurish hamda zarur natijalarni olishda foydalidir. Ayniqsa bu umumiy belgilash tizimiga ega korpuslar mavjud bo‘lmaganda standart yondashuv sifatida qo‘llanilishi va qo‘lda teglanishi mumkin. Mazkur universal tegsetlarning yaratilishi haqida gap ketganda, kelajakdagi sintaktik tuzilmani nazoratsiz o‘rganish bo‘yicha tadqiqotlarni osonlashtirish va eng yaxshi amaliy natijalarni standartlashtirish maqsadida o‘n ikki universal so‘z turkumidan iborat bo‘lgan belgilash tizimi (tagset) taklif qilingani aytiladi. Ushbu belgilash tizimiga qo‘shimcha ravishda, 25 xil daraxtbanki (treebank) belgilash tizimlarini ushbu universal to‘plamga moslashtirish xaritasi ishlab chiqilgan. Natijada, daraxtbanki ma’lumotlari bilan birgalikda ushbu universal belgilash tizimi

va moslashtirish xaritasi 22 ta til uchun umumiy so‘z turkumlarini o‘z ichiga olgan ma'lumotlar to‘plamini hosil qiladi. Ushbu resursdan foydalanishni uchta tajriba orqali namoyish etishgan: turli tillarda belgilash aniqliklarini taqqoslash, standart so‘z turkumlaridan foydalanmaydigan nazoratsiz grammatika induksiya yondashuvini taqdim etish va universal belgilash tizimidan foydalanib, tillarda tobelik munosabatini aniqlovchi analizatorlarni (dependency parsers) yaratish, bu esa zarur natijalarga erishishga imkon berishi ta’kidlangan [6]. Shunday bo‘lsa-da har bir til xususiy va grammatik jihatdan bir-biridan farqliligi sababli universal teglash tizimidan foydalanish doim ham foydali bo‘lmaydi.

O‘zbek tili korpuslari uchun oxirgi tadqiqotlarda teg tizimi taklif qilinganini kuzatish mumkin. Bunda UD (universal dependencies) teglaridan foydalanish holatlari va maxsus tegsetlar ham ishlab chiqilganini kuzatish mumkin. N.Abduraxmonova Protégé texnologiyasi asosida o‘zbek tilidagi sintaktik strukturalarni teglash uchun 18 ta universal teglar tizimi taklif etilgan [7].

1-jadval. Protégé texnologiyasi asosidagi o‘zbek tili sintaktik teglar tizimi

Tag	Name_English	Tag	Name_English
WC	Word combination	**SLP**	Singular personal pronouns
COLC	Collocation	**PPL**	Plural personal pronouns
FP\FCOLC	Free phrase\ Free collocation	**ICN\CPC\T**	Interconnectedness\Complicity
NP	Noun Phrase	**S**	Simple sentence
NA	Noun Adjoinment	**Sub**	Subject
NG	Noun Government	**Obj**	Object
NCS	Noun Collateral subordination	**Attr**	Attributive
VP	Verb Phrase	**Mod**	Modifier
AGRM	Agreement	**Pre**	Predicate

Yana bir guruh olimlar gapning asosini tashkil etuvchi beshta asosiy bo‘lak va gap bo‘lagi bo‘lib kelmaydigan komponentlar uchun sintaktik tegset taklif qilishgan [8]. O‘zbek tilida gaplarda asosiy komponent hisoblanuvchi kesim, ega, aniqlovchi, to‘ldiruvchi, hol va ularning turlari uchun teglar taklif qilingan. Shuningdek, undov va kiritmalar uchun ham teglar mavjud. UD tegsetlari tillarda grammatikani (so‘z va ularning sintaktik bog‘lanishlari) bir xil tarzda belgilash uchun ishlatiladigan bir tizim bo‘lsa-da, ammo tilning o‘z ichki qonun-qoidalarini hisobga olgan holda morfologik va sintaktik teglarning ishlab chiqilishi tabiiydir.

2-jadval. M.Sharipov va boshqa mualliflar tomonidan taklif qilingan sintaktik teglar ro‘yxati

Nomi	Sintaktik teg	Ta’rifi
SUBJECT (ega)	EG	Subject
PREDICATE (kesim)	OK	Noun Predicate
	FK	Verb Predicate

ATTRIBUTE (aniqlovchi)	QA	Genetive Attribute
	SA	Adjectival Attribute
OBJECT (to‘ldiruvchi)	VL	Indirect Object
	VS	Direct Object
MODIFIERS (hol)	VH	Condition Modifiers
	PH	Time Modifier
	OH	Place Modifier
	SH	The Reason Modifier
	MH	The Aim Modifier
EXCLAMATION (undov)	UN	A person or object that is focused on speech
THE ENTRY WORD (kiritma)	KR	The entry word

Yuqorida taklif qilingan tegsetlar biz tadqiqotimiz uchun ishlab chiqqan tegsetlar tizimiga o‘xshash bo‘lsa-da, ammo ba’zi teglarning belgilanishi farqlidir. Qachonki umumiy kelishuvga kelinganda, O‘zbek tili teggeri ishlab chiqilishi mumkin. Quyida o‘zbek tili ta’limiy korpusi uchun ishlab chiqilgan sintaktik teglar tizimi bilan tanishish mumkin.

3-jadval. O‘zbek tili sintaktik analizatori uchun taklif qilingan sintaktik teglar ro‘yxati

POS teg	Belgilanishi	POS teg	Belgilanishi
Ega	**E**	**Hol**	
Kesim		***Ravish holi***	**RH**
Ot-kesim	**OK**	***Payt holi***	**PH**
Fe’l-kesim	**FK**	***O‘rin holi***	**O‘H**
Aniqlovchi		***Sabab holi***	**SH**
Sifatlovchi-aniqlovchi	**SA**	***Maqsad holi***	**MH**
Izohlovchi-aniqlovchi	**IA**	***Miqdor-daraja holi***	**MDH**
Qaratqich-aniqlovchi	**QA**	**Kirish so‘z**	**KS**
To‘ldiruvchi		**Kirish birikma**	**KB**
Vositali-To‘ldiruvchi	**VT**	**Kirish gap**	**KG**
Vositasiz-To‘ldiruvchi	**VST**	**Undalma**	**U**

Agglyutinativ va boy morfologiyaga ega til sifatida o‘zbek tilining parser dasturlarini ishlab chiqishda ba’zi qiyinchiliklarga sabab bo‘ladi. Ma’lumki, har bir tilning gap qurilishi va qoidalari aniq belgilangan. Masalan, ingliz tili va unga o‘xshash tillarda gap qurilishi aniq va so‘zlar tartibi qat’iydir. O‘zbek tili uchun gap qurilishi va so‘z tartibi erkin hisoblanadi. Qoidalarga asoslangan parsing usullarini ishlab chiqish qat’iy grammatik qoidalar va lug‘at ma’lumotlaridan foydalanishga asoslanadi. Ko‘pincha bunda formal grammatika qoidalar to‘plamidan, ya’ni CFG (Context free grammar)dan foydalaniladi. Kontekstsiz grammatikada parse

daraxtidan foydalanadi. Parse daraxtini ko‘rsatish yoki grafik tasvirlanishi parserni tasvirlashning qulay usulidir.

Agglyutinativ tillarda gaplardagi so‘z tartibining erkinligi, o‘ziga xos sintaktik bog‘lanishlar va morfologiyasining murakkabligi sababli qoidalarga asoslangan parser dasturlarini natijadorligini pasaytiradi. **Qoidalarga asoslangan va statistik yondashuvlarni** birlashtirgan holda qoidalarga asoslangan parser va mashinali o‘rganishga asoslangan modellarni kombinatsiya qilish orqali yuqori natijalarni olish mumkin.

Morfologiya va sintaksis NLPdagi eng muhim lingvistik qatlam hisoblanadi. Grammatik morfemalar turkumlarning shaklini o‘zgartirishi bilan birga sintaktik tahlilda ham rol o‘ynaydi. Chunki o‘zbek tilida grammatik shakllar orqali gap bo‘laklarining formal modellarini qurish mumkin. Modellar orqali esa gapning komponentlarini aniqlashga erishiladi. POS teglangan korpus orqali matnlarni sintaktik teglash mumkin. Buni korpusda tahlil qilish mumkin:

Tokzorni[[N]] eslayman[[VB]].
Tokzorni[[T]] eslayman[[K]].

U[[P]] qo‘rqib ketgandi[[VB]].
U[[E]] qo‘rqib ketgandi[[K]].

Bundan[[P]] kimga[[P]] foyda[[N]]?
Bundan[[T]] kimga[[T]] foyda[[K]]?

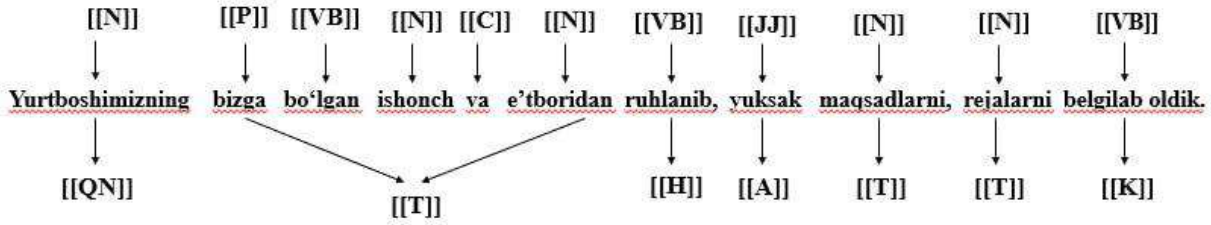
Yana[[RR]] bir[[NUM]] masala[[N]] bor[[MD]].
Yana bir[[A]] masala[[E]] bor[[K]].

Albatta[[MD]], siyosatchilar[[N]] quloq solmadi[[IB]].
Albatta[[KS]], siyosatchilar[[E]] quloq solmadi[[K]].

Yuqoridagi gaplarning POS teg va sintaktik teglashidan kuzatish mumkinki, o‘zbek tilida qoidalarga asoslangan parser qurish orqali gap bo‘laklarini aniqlash mumkin. Bunda har bir so‘z turkumining qanday grammatik morfemalarni olganda, qaysi gap bo‘lagi vazifalarida kelish mumkinligi aniqlanishi, formal modellari ishlab chiqilishi kerak. Murakkab tuzilishli gaplarda to‘g‘ri natijaga ega bo‘lmasligimiz mumkin. Chunki gaplarda sintaktik teglashda bitta lemma bitta bo‘lak bo‘lmasligi yoki ot so‘z turkumidagi so‘z doim ham ega yoki fe’l kesim bo‘lmasligi mumkin. Buni parser qanday aniqlaydi? Masalan:

Yurtboshimizning bizga bo‘lgan ishonch va e’tiboridan ruhlanib, yuksak maqsadlarni, rejalarni belgilab oldik gapini POS teglash va sintaktik teglaganimizda lemmalar soni bir xilmi? Tahlilni quyida ko‘ramiz.

yurtboshimizning – qaratqich aniqlovchi
bizga bo‘lgan ishonch va e’tiboridan – kengaytirilgan to‘ldiruvchi
ruhlanib – hol
yuksak – aniqlovchi
maqsadlarni – to‘ldiruvchi
rejalarni – to‘ldiruvchi
belgilab oldik – fe’l-kesim vazifasida kelmoqda.



Yuqoridagi tahlilda gap POS va sintaktik teglandi. Gapda 12 ta token va 11 ta lemma mavjud. Demak, 11 ta POS teglangan lemma, lekin 7 ta gap bo‘lagi mavjud. Demak, gaplardagi lemma va gap bo‘lagi mos emas. Har bir POS teg bitta gap bo‘lagi bo‘lolmaydi.

Xulosa. Xulosa qilib aytganda, o‘zbek tilida morfologik va sintaktik annotatsiyalangan korpuslarni yaratish NLP sohasidagi eng muhim talablardan biridir. Buning uchun o‘zbek tili matnlarida sintaktik munosabatlarni chuqurroq annotatsiya qilish imkoniyatini o‘rganishimiz, sintaktik teglovchi dasturlar uchun model va algoritmlarni ishlab chiqishimiz zarur.

Turkiy tillarning grammatik xususiyatlari va gap tuzilmalari umumiy jihatdan o‘xshash bo‘lgani uchun kelgusidagi ishlarimizda turkiy tillarda yaratilgan parserlarni tahlil qilishimiz, o‘zbek tilida treebank qurishda tillarda mavjud modellar asosida tahlillarning aniqlik darajasini kuzatishimiz zarur hisoblanadi. Bu kabi tadqiqot va izlanishlar natijasi o‘zbek tilida tabiiy tilga ishlov berish vazifalarini samarali bajarishda dolzarb hisoblanadi.

Foydalanilgan adabiyotlar:

1. Francis W. Nelson, Henry Kucera. Computational Analysis of Present-Day American English. Providence, RI: Brown University Press., 1967.; Greene B.B., Rubin G. M. Automatic grammatical tagging of English. Department of Linguistics, Brown University, Providence, Rhode Island, 1971.

2. Garside R. The CLAWS word-tagging system. In Garside, R., Leech, G., and Sampson, G. (Eds.), The Computational Analysis of English, 30–41. Longman, 1987.

3. Brill Eric. "A simple rule-based part of speech tagger." In Proceedings, Third Conference on Applied Natural Language Processing, ACL, Trento, Italy, 1992.

4. https://universaldependencies.org/

5. Vojtech Kovář. Automatic Syntactic Analysis for Real-World Applications. PhD thesis. Brno, Spring 2014.

6. Slav Petrov, Dipanjan Das, Ryan McDonald. A Universal Part-of-Speech Tagset. // In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 2089–2096, Istanbul, Turkey. European Language Resources Association (ELRA), 2012.

7. Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari. Monografiya. – Toshkent, 2021.

8. Sharipov M., Mattiev J., Sobirov J., Baltayev R. Creating a morphological and syntactic tagged corpus for the Uzbek language / The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing (ALTNLP). - Koper, Slovenia, June 7-8, 2022.

UDK: 811.512.133`42

INCEpTION PLATFORMASI YORDAMIDA O‘ZBEK TILI UCHUN IYERARXIK BOG‘LIK DARAXTI KORPUSI YARATISH

Matlatipov San’atbek G‘ayratovich,
PhD, yetkachi ilmiy xodim
s.matlatipov@nuu.uz
O‘zbekiston Milliy universiteti

Matyoqubova Shaydo Maqsud qizi,
magistratura talabasi
shaydomatyoqubova522@gmail.com
Urganch davlat universiteti

Komilova Madinabonu Baxtiyor qizi,
magistratura talabasi
komilovamadina0920@gmail.com
Urganch davlat universiteti

Annotatsiya. Ushbu maqolada o‘zbek tilidagi matnlardan olinadigan sintaktik iyerarxik bog‘lanishli daraxt korpusini yaratish haqida so‘z boradi. Maqsad – o‘zbek tili uchun so‘zlararo sintaktik munosabatlarni aniq belgilangan daraxt ko‘rinishida tasvirlash korpusini shakllantirish. Korpus materiali sifatida zamonaviy adib Shuhrat Matkarimning “Boljon” to‘plamidan olingan 30 ta sifatli jumla tanlab olindi. Ushbu jumlalar INCEpTION platformasi yordamida ikki nafar annotator tomonidan lemmatizatsiya, morfologik belgilash va sintaktik bog‘lanishlarni qo‘lda belgilandi, so‘ngra kuratsiya (birlashtirish) bosqichi orqali yakuniy kelishuvga erishildi. Natijada, o‘zbek tilidagi hikoyaviy uslub uchun birinchi bog‘lanishli daraxt korpusi yaratildi va undagi sintaktik xususiyatlar tahlil qilindi. Asosiy natijalar shuni ko‘rsatadiki, mazkur korpus kelgusida o‘zbek tili uchun sintaktik tahlil, mashinaviy tarjima hamda ta’limiy lingvistik vositalar uchun asos bo‘la oladi.

Annotation. This article presents the development of a syntactic hierarchical dependency tree corpus based on narrative texts in the Uzbek language. The main objective is to create a corpus that systematically and clearly represents word-to-word syntactic relations in narrative-style texts of Uzbek, using a tree-based structure. As corpus material, 30 high-quality sentences were selected from “Boljon,” a short story collection by the contemporary Uzbek writer Shuhrat Matkarim. These sentences were manually annotated by two annotators using the INCEpTION platform, including lemmatization, morphological tagging, and

syntactic dependency annotation. The final version of the annotations was produced through a curation phase, achieving consensus between annotators. As a result, the first syntactic dependency tree corpus for Uzbek narrative style was created and syntactic features within the corpus were analyzed. The findings demonstrate that this corpus can serve as a valuable foundation for syntactic parsing, machine translation, natural language processing, and educational linguistic tools for the Uzbek language.

Аннотация. В данной статье рассматривается создание синтаксического иерархического корпуса с древовидной структурой на основе повествовательных текстов на узбекском языке. Основная цель исследования — сформировать корпус, в котором синтаксические связи между словами в текстах повествовательного жанра узбекского языка будут чётко и систематично представлены в виде древовидной структуры. В качестве материала корпуса были отобраны 30 качественных предложений из сборника рассказов «Больжон» современного узбекского писателя Шухрата Маткарима. Эти предложения были вручную аннотированы двумя аннотаторами с использованием платформы INCEpTION, включая лемматизацию, морфологическую разметку и синтаксические зависимости. Финальная версия корпуса была сформирована в результате этапа кураторской сверки и согласования аннотаций. В результате был создан первый синтаксический корпус с древовидными зависимостями для повествовательного стиля узбекского языка, в котором были проанализированы синтаксические особенности. Полученные результаты показывают, что данный корпус может стать важной основой для синтаксического анализа, машинного перевода, обработки естественного языка и разработки образовательных лингвистических ресурсов для узбекского языка.

Kalit so‘zlar: *o‘zbek tili, hikoyaviy matnlar, sintaktik bog‘lanish, daraxt korpusi, lemmatizatsiya, morfologik belgilash, Universal Dependencies, INCEpTION platformasi, annotatsiya, kompyuter lingvistikasi, tabiiy tilni qayta ishlash (NLP)*

O‘zbek tili Markaziy Osiyodagi eng yirik tillardan biri bo‘lib, so‘zlashuvchilari soni 30–40 million atrofida baholanadi. Turk tillari orasida turk tilidan keyin ikkinchi eng ko‘p so‘zlashuvchiga ega til hisoblanishiga qaramay, o‘zbek tili uchun kompyuter lingvistikasi (NLP) sohasida resurslar juda cheklangan. Xususan, matnlarni sintaktik daraxt shaklida belgilash uchun treebank (daraxt

korpusi)lar deyarli mavjud emas edi. Bunday korpuslar esa bugungi kunda ko‘plab tillar uchun yaratilgan bo‘lib, ular orqali turli NLP modellari o‘qitiladi va lingvistik tadqiqotlar olib boriladi.

Ayrim so‘nggi tadqiqotlarda o‘zbek tilida turli korpus va modellarning paydo bo‘layotgani kuzatilmoqda: masalan, UZWORDNET leksik-semantik tarmog‘i yaratildi, matnlarni tuyg‘u (sentiment) yorliqlari bilan belgilangan to‘plamlar va matn tasnifi korpuslari taklif qilindi (Kuriyozov va boshq., 2019; Rabbimov va Kobilov, 2020), hatto o‘zbek tili uchun UzBERT hamda BERTbek kabi yirik transformator modellar dastlabki o‘qitish orqali tayyorlandi. Shunga qaramay, hali ham o‘zbek tilida to‘liq qo‘lda belgilangan sintaktik daraxt korpusi deyarli yo‘q edi. 2025-yilda o‘zbek tili uchun ilk Universal Dependencies (UD) daraxt korpusi taqdim etildi, u 500 ta jumlani (taxminan 5850 token) qamrab oladi Biroq, mazkur UD korpusi yangiliklar va badiiy matnlardan avtomatik va qo‘lda belgilash aralash usulida tuzilgan). O‘zbek tilidagi hikoyaviy (badiiy) uslub xususiyatlarini to‘la aks ettiruvchi, mutlaqo qo‘lda, yuqori sifatli sintaktik korpus hali yaratilmagan edi.

Ushbu loyiha ana shu bo‘shliqni to‘ldirishga qaratilgan. Biz o‘zbek tilidagi hikoyaviy matnlardan kichik hajmli bo‘lsa-da, sifatli va murakkab sintaktik bog‘lanishlarni o‘zida jamlagan daraxt korpusini yaratdik. Korpus tuzishda Universal Dependencies tamoyillariga asoslandik, ya’ni so‘z turkumlari va grammatik kategoriyalar xalqaro standart (UD) bo‘yicha belgilandi. Natijada o‘zbek tilini kompyuterda qayta ishlash (NLP) uchun yangi bir resurs – hikoyaviy matnlarga mo‘ljallangan treebank hosil qilindi. Quyida ushbu korpusni shakllantirish metodologiyasi, uni belgilash jarayoni, annotatorlararo kelishuv ko‘rsatkichlari hamda dastlabki tahliliy natijalar yoritiladi.

Adabiyotlar tahlili va metodologiya

O‘zbek tilini kompyuter lingvistikasi yo‘nalishida qayta ishlash bo‘yicha dastlabki ishlar morfologik tahlil va so‘z turkumlarini belgilashdan boshlangan. Masalan, Matlatipov va Vetulani (2009) Prolog muhiti yordamida o‘zbek tilining morfologik tahlilchisini yaratgan, biroq ushbu tizim murakkab so‘z formalarini to‘liq qo‘llab-quvvatlay olmagan. Keyinchalik o‘zbek tilining grammatik xususiyatlarini chuqurroq aks ettiruvchi qo‘shimcha belgilar taklif qilindi (Sharipov va boshq., 2022) va UzbekTagger deb nomlangan qoidaviy qatorga asoslangan so‘z turkumi teglagichi ishlab chiqildi. Ushbu POS-tegger o‘zbek tilida 12 ta asosiy so‘z

turkumini aniqlab, qo‘lda belgilangan kichik korpusda sinovdan o‘tkazilgan (Sharipov va boshq., 2023). Morfologik tahlilda so‘zning negizini (stem) ajratish va lemmalash bo‘yicha ham ishlanmalar mavjud: UzbekStemmer (Sharipov & Yuldashev, 2022) algoritmi va so‘ngra UzMorphAnalyzer tizimi (Salaev, 2023) taqdim etilgan. UzMorphAnalyzer o‘zbek tilida so‘zlarni tarkibiy qismlarga ajratib, ularning lug‘aviy ma’nosini aniqlash (stemmer+lemmatizer+morfologik analiz) imkonini beradi. Shuningdek, UZWORDNET (Agostini va boshq., 2021) singari leksik ma’lumot bazalari ham yaratilgan bo‘lib, o‘zbek tilida sinonimik bog‘lanishlar tarmog‘ini taklif qiladi. Biroq, tilning sintaktik tuzilishini aks ettiruvchi bog‘lanishli daraxt korpusi masalasi e’tibordan chetda qolib kelgan edi. Yaqinda Akhundjanova va Talamo (2025) ilk bor o‘zbek tilining Universal Dependencies treebankini taqdim etdilar. [1] UD_Uzbek-UT deb nomlangan ushbu korpus 500 ta jumla (5850 token)dan iborat bo‘lib, yangiliklar va badiiy adabiyotlardan olingan jumlalar avtomatik analiz va qo‘lda tuzatish kombinatsiyasi orqali belgilangan. Bizning ishimiz esa to‘liq qo‘lda va faqat hikoyaviy uslubga tegishli jumlalarni belgilash orqali o‘zbek tilidagi sintaktik bog‘lanishlarni batafsilroq ochib berishga qaratilgan.[2]

Korpus uchun ma’lumot manbasi sifatida adib Shuhrat Matkarimning “Boljon” nomli hikoyalar to‘plami tanlandi. Ushbu to‘plamdagi matnlardan 30 ta murakkab, grammatik jihatdan boy jumlalar qo‘lda ajratib olindi. Jumlalarning umumiy uzunligi va mazmuniy xilma-xilligi e’tiborga olindi – dialog va monologlar, sodda va qo‘shma gaplar, turli sintaktik konstruktsiyalar kiritildi. Shundan so‘ng har bir jumla tokenlarga ajratildi (so‘z va alomatlarga bo‘linib, maxsus ID raqamlar bilan tartiblandi), har bir token uchun lemma (asosiy lug‘aviy shakl) belgilandi, so‘z turkumi (masalan, ot, fe’l, sifat va hokazo) va uning morfologik xususiyatlari (masalan, son, kelishik, zamon, kishilik) aniqlanib teglar qo‘yildi. So‘ngra jumla tarkibida har bir so‘zning grammatik jihatdan qaysi so‘zga bog‘lanishi (kimga ega, kimni to‘ldiruvchi, qaysi so‘zga aniqlovchi va hokazo) bog‘lanish relatsiyasi ko‘rinishida ko‘rsatildi. Barcha belgilash ishlari INCEpTION dasturiy platformasi yordamida amalga oshirildi. Ikki nafar mustaqil annotator har bir jumlani mustaqil belgiladi. Annotatsiya jarayoni quyidagi bosqichlardan iborat bo‘ldi:

1. Lemmatizatsiya va so‘z turkumini belgilash: Annotatorlar jumladagi har bir so‘z uchun uning boshlang‘ich lemmasini va Universal Dependencies tagiga mos so‘z turkumini belgilandi (masalan, otlar uchun NOUN, fe’llar uchun VERB, ravishlar uchun ADV va hokazo).

2. Morfologik kategoriyalarni belgilash: Har bir tokenning grammatik kategoriyalari aniqlanib yozildi. Masalan, otlarda kelishik va son (Case, Number xususiyatlari), fe’llarda zamon, nisbat, shaxs va son (Tense, Voice, Person, Number) kabi attributlar belgilandi. Bu belgilar UD standartidagi qisqacha taglar yordamida ifodalandi.

3. Sintaktik bog‘lanishlarni aniqlash: Jumladagi har bir so‘zning sintaktik jihatdan qaysi so‘zga va qanday munosabat bilan bog‘langanligi ko‘rsatildi. Bunda har bir bog‘lanish turi UD relatsiyalar inventaridan tanlandi (masalan, ega uchun nsubj, to‘ldiruvchi uchun obj, aniqlovchi uchun amod, hol uchun obl, kesim uchun root va hokazo). Har bir token tegishli “bosh so‘z” (head)ning ID raqami bilan bog‘landi va bog‘lanish turi yozildi. Masalan, kesim fe’l odatda gapning “root”i bo‘lib, boshqa so‘zlar bevosita yoki bilvosita unga bog‘lanadi.

Ikki annotator mustaqil belgilab chiqqach, kuratsiya bosqichi o‘tkazildi. INCEpTION platformasining curation rejimida annotatorlar kiritgan belgilardagi tafovutlar ko‘rishli tarzda taqqoslanib, yakuniy kelishilgan variant shakllantirildi. Bu bosqichda har bir zid holat muhokama qilinib, umumiy to‘g‘ri belgi tanlandi. Kuratsiya natijasida bir xil ID raqamli jumlalar uchun yakuniy “oltin standart” annotatsiya olindi.

Eksperimentlar va natijalar

Yaratilgan korpus kichik hajmli bo‘lsa-da, uning tarkibiy statistikasi va xususiyatlari tahlil qilindi. Korpus quyidagi ko‘rsatkichlarni o‘z ichiga oladi:

- Jumla soni: 30 ta
- Tokenlar soni (umumiy): ~600 ta
- O‘rtacha jumla uzunligi: 20 ta token atrofida
- Turli (unikal) leksemalar: ~400 ta

Yuqoridagi taxminiy taqsimotdan ko‘rinadiki, hikoyaviy matnlarda otlar va fe’llar eng yuqori ulushga ega, bu esa tabiiy holdir – matn asosan voqea-hodisalarni (fe’l orqali) va ishtirokchilarni (ot orqali) ifodalaydi. Korpusdagi har bir jumla sintaktik daraxt ko‘rinishida annotatsiya qilingan. Quyida korpusdan olingan bir misolni CoNLL-U formatida keltiramiz (har bir token uchun ID, so‘zning o‘zi, lemma, asosiy turkum, morfologik grammema, sintaktik bosh ID va bog‘lanish turi ko‘rsatilgan):

1 Boljon Boljon PROPN _ Case=Nom|Number=Sing 5 nsubj _ _

2 kechasi kecha ADV _ _ 5 advmod _ _

3 bog‘da bog‘ NOUN _ Case=Loc|Number=Sing 5 obl _ _

4 sayr sayr NOUN _ Case=Nom|Number=Sing 5 obj _ _

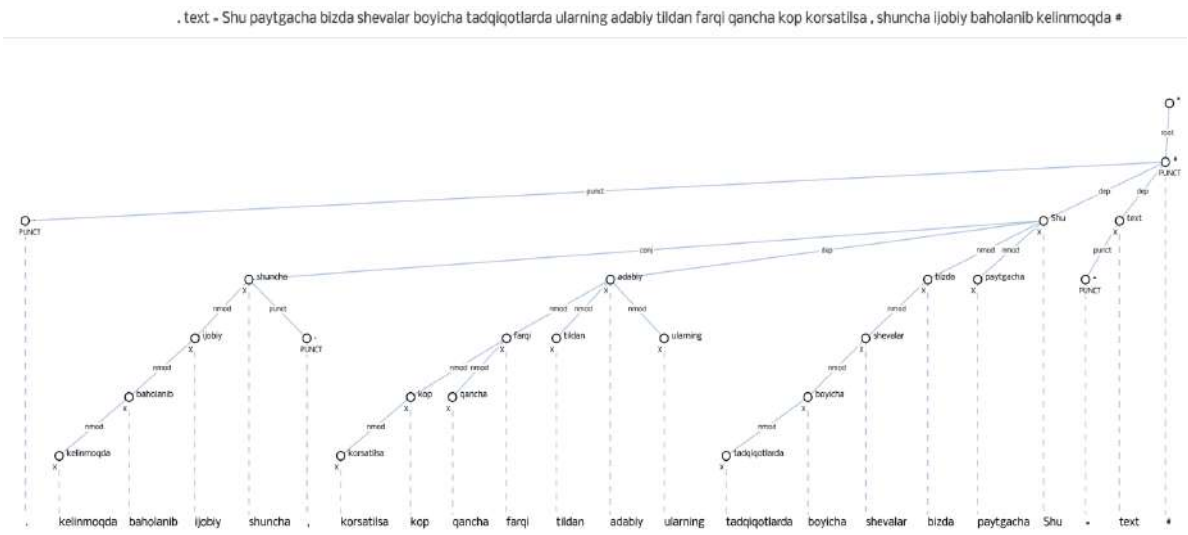
5 qildi qil VERB _ Tense=Past|Person=3|Number=Sing 0 root _ _

6 . . PUNCT _ _ 5 punct _ _

Misol: Boljon kechasi bog‘da uni ko‘rdi. Ushbu jumlada "ko‘rdi" fe’li kesim bo‘lib, root sifatida belgilangan. Boshqa so‘zlar bevosita unga bog‘langan: Boljon – nsubj (ega, kim ko‘rdi?), kechasi – advmod (qachon ko‘rdi?), bog‘da – obl (qayerda ko‘rdi?), uni – obj (kimni ko‘rdi?) munosabatlari bilan bog‘langan. Bu orqali, jumladagi har bir so‘zning grammatik vazifasi va gap tuzilishidagi o‘rni daraxt shaklida ko‘rinadi.[3]

Korpus tahlili jarayonida o‘zbek tilining hikoyaviy uslubiga xos bir qator qiziqarli sintaktik va morfologik holatlar kuzatildi. Avvalo, egasiz gaplar (subyekti tushirilgan, shaxs-son grammatik ko‘rsatkichlari fe’lda mavjud) uchrashi kuzatildi. Masalan, ba’zi jumlalarda gapning egasi yozilmagan bo‘lsa-da, fe‘ldagi shaxs va son qo‘shimchasidan (3-shaxs birlik) uning “u” (3-shaxs) ekani anglashiladi. Bunday holatlar o‘zbek tilining null-subject (pro-drop) xususiyatini tasdiqlaydi va korpusimizda bunday gaplar aniq nsubj:implic yoki shunga o‘xshash usulda belgilandi.

Shuningdek, bog‘langan va ravishdosh vositali qo‘shma gaplar ko‘pligi e‘tiborga molik bo‘ldi. Hikoyaviy jumlalarda bir nechta sodda gaplar va, ham kabi bog‘lovchilar yoki -sa kabi qo‘shimchalar orqali bog‘lanib, murakkab gaplar hosil qilingan. Masalan: “U hovliga chiqib, atrofga nazar tashladi va birpas sukut saqladi.” kabi jumlalar bir vaqtning o‘zida ikkita-yu uchta sodda harakatni ifodalaydi. Korpusimizda bunday hollarda ravishdoshli fe’lning bog‘lanishi advcl (hol ergash gapli qo‘shma gap) yoki conj (bog‘lovchisiz qo‘shma gap) sifatida belgilandi, kontekstga qarab. Natijada, daraxt tuzilishida bir necha rootga ega bo‘lmagan, balki bitta bosh fe’l atrofiga birikkan struktura hosil bo‘ladi (ketma-ket bog‘langan fe’llar).Bu o‘zbek tilining polisentetik sintaksisini ko‘rsatadi.(1-rasm)



1-rasm. Sintaktik bog‘lanishli bog‘lik daraxt

Morfologik jihatdan qaralganda, otlarda kelishik kategoriyasining faol qo‘llanilishi kuzatildi: korpusdagi otlarning ~60% atrofida turli kelishik qo‘shimchalari mavjud (-ni, -ga, -da, -dan va hokazo). Bu hol, o‘zbek tilida predloglarning yo‘qligi va ularning vazifasini asosan qo‘shimchalar bajarishini tasdiqlaydi. Masalan, yuqoridagi misolda bog‘da so‘ziga locative kelishik qo‘shimchasi -da qo‘shilib, ingliz tilidagi “in the garden” ma’nosini bitta so‘z orqali ifodalagan. Shuningdek, ba’zi hollarda tushum kelishigi (akuzativ) qo‘shimchasi tushirib qoldirilgan (vositasiz to‘ldiruvchi) holatlar ham uchradi, bu esa semantik noaniqlikni keltirib chiqarmagan.

Yuqoridagi kuzatishlar bizning korpusimizni rus va ingliz tillaridagi treebanklar bilan solishtirganda qiziqarli farqlarni ko‘rsatadi. Masalan, ingliz tilida ega yoki to‘ldiruvchi tushirilishi deyarli kuzatilmaydi, so‘z tartibi qattiqroq, kelishik degan kategoriya yo‘q. Rus tilida esa kelishiklar mavjud, ammo unda otlarda grammatik rod (jins) kategoriyasi bor, o‘zbek tilida esa jins kategoriyasi yo‘q. Shuningdek, rus tilida ham egani tushirish mumkin (masalan, buyruq gaplarda), lekin o‘zbek tilidagidek harakat fe’l ma’noviy shaxsini qo‘shimchadan bilish odatiy emas (rus tilida fe’l shakli subyektga uncha moslashmaydi 3-shaxsdan tashqari). Turk tiliga solishtirilsa, o‘zbek tilining sintaksisi ko‘p jihatdan o‘xshash: so‘z tartibi erkin SOV, so‘zlar agglutinativ quriladi, artikl yo‘q. Biroq, o‘zbek tilida tarixiy fonetik o‘zgarishlar tufayli turk tilidagidek kuchli tovush moslashuvi (vowel harmony) emas, bir xil qo‘shimchalar qo‘llanadi, bu esa morfologik tahlilni biroz soddalashtirishi mumkin.

INCEpTION platformasida ishlash jarayoni ancha qulay bo‘lib, unda foydalanuvchilar aniq interfeys orqali tokenlarni belgilash, relatsiyalar o‘rnatish va tahrir qilish imkoniyatiga ega. Annotatsiya turlarining ko‘rsatilgan rangi, bog‘lanish yo‘nalishlari va o‘zaro aloqador tokenlar interaktiv ko‘rinishda tasvirlanadi. Shu bilan birga, platforma annotatorlararo kelishuv (inter-annotator agreement)ni baholash uchun avtomatik vositalarni ham taklif qiladi. Bu vositalar yordamida lemmatizatsiya, POS teglash va dependency relatsiyalardagi farqlar hisoblab chiqiladi va kappa koeffitsienti kabi statistik ko‘rsatkichlar orqali umumiy moslik darajasi aniqlanadi.

Xulosa o‘rnida ta’kidlash joizki, mazkur ish natijasida o‘zbek tilining hikoyaviy uslubi uchun dastlabki kichik hajmdagi, ammo boy sintaktik axborotga ega bo‘lgan bog‘lanishli daraxt korpusi yaratildi. Ushbu korpusning asosiy ahamiyati shundaki, u o‘zbek tilida sintaktik parsing (avtomatik tahlil) masalalari uchun qo‘lda belgilangan namunaviy ma’lumotlar to‘plamini taqdim etadi. Korpusdan foydalangan holda, kelgusida o‘zbek tili uchun stochastic parsers (masalan, neyron tarmoq asosidagi DEP parserlar)ni o‘qitish mumkin bo‘ladi. Shuningdek, korpus mashinaviy tarjima sohasida, xususan, o‘zbek tilining sintaktik tuzilishini model qilishda asqotadi – tarjima tizimlari bu korpusdan o‘zbekcha gap tuzilishini o‘rganishi, boshqa tillar daraxt korpusi bilan taqqoslab, mos keluvchi strukturalarni topishi mumkin. Korpusdan o‘rganilgan qoidalar va andozalar ta’limiy dasturlarda, masalan, o‘zbek tili grammatikasini o‘rgatuvchi interaktiv ilovalarda qo‘llanilishi mumkin: o‘quvchilar jumlalarni tahlil qilishni aynan shu korpus misollarida mashq qilishlari mumkin.

Foydalanilgan adabiyotlar:

1. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021). UZWORDNET: O‘zbek tili uchun leksik-semantik ma’lumotlar bazasi. 11-Global WordNet konferensiyasi materiallari, sah. 8–19, Janubiy Afrika (UNISA).

2. Akhundjanova, A., & Talamo, L. (2025). O‘zbek tili uchun Universal Dependencies daraxt korpusi. RESOURCEFUL-2025: Qo‘llab-quvvatlanishi past tillar va domenlar uchun resurslar bo‘yicha 3-Workshop materiallari. ACL..

3. Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodríguez, C. (2019). “Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek.” In *Human Language Technologies as a Challenge for Computer Science and Linguistics: 9th Language and Technology Conference (LTC 2019)*, Poznan, Poland. Revised Selected Papers, pp. 232–243. Berlin: Springer.

4. Kuriyozov, E., Vilares, D., & Gómez-Rodríguez, C. (2024). BERTbek: o‘zbek tili uchun oldindan o‘qitilgan transformator modeli. 3-SIGUL (LREC-COLING 2024) anjumani materiallari, sah. 33–44.

5. Mansurov, B., & Mansurov, A. (2021). UzBERT: o‘zbek tilida BERT modelini dastlabki o‘qitish. *arXiv preprint* arXiv:2108.09814.

6. Matlatipov, G., & Vetulani, Z. (2009). Prolog yordamida o‘zbek tili morfologik tahlili. In *Proceedings of the 7th International Conference on Formal Approaches to South Slavic and Turkic Languages*.

7. Rabbimov, I. M., & Kobilov, S. S. (2020). Multi-class text classification of Uzbek news articles using machine learning. *Journal of Physics: Conference Series*, 1546(1), 012097.

8. Rahmatullayev, Sh. (2006). *Hozirgi adabiy o‘zbek tili* (darslik). Toshkent: Universitet nashri.

9. Salaev, U. (2023). UzMorphAnalyzer: O‘zbek tili uchun morfologik tahlil modeli (so‘zlardagi qo‘shimchalarga asoslangan). *Science and Innovation*, 2(1), 29–34.

10. Sharipov, M., Mattiyev, J., Sobirov, J., & Baltayev, R. (2022). O‘zbek tilining morfologik va sintaktik teglangan korpusi-ni yaratish. ALTNLP 2022: *Agglutinative Language Technologies as a Challenge of NLP* (virtual konferensiya, Koper, Sloveniya), CEUR-WS, vol. 3315, sah. 93–98.

11. Sharipov, M., & Sobirov, O. (2022). O‘zbek tili uchun qoidaviy lemmalash algoritmini ishlab chiqish. ALTNLP 2022 konferensiyasi materiallari, CEUR-WS, vol. 3315, sah. 154–159.

12. Sharipov, M., & Yuldashev, O. (2022). UzbekStemmer: O‘zbek tili uchun qoidaviy stemming algoritmi. ALTNLP 2022 konferensiyasi materiallari, CEUR-WS, vol. 3315, sah. 137–144.

13. Sharipov, M., Kuriyozov, E., Yuldashev, O., & Sobirov, O. (2023). UzbekTagger: O‘zbek tili uchun qoidaviy so‘z turkumlari teglagichi. *arXiv preprint* arXiv:2301.12711.
14. Shuhrat Matkarim.(2017). Kun qanday boshlanadi. – G‘afur G‘ulom nashriyoti.
15. Universal Dependencies Treebank for Uzbek .
16. The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation - ACL Anthology

UDK: 811.512.133`42

O‘ZBEK TILIDA BILIDIRLGAN FIKRLARNI ASPEKTLI SENTIMENT TAHLIL QILISH

Matlatipov San’atbek G‘ayratovich
PhD, yetkachi ilmiy xodim
s.matlatipov@nuu.uz
O‘zbekiston Milliy universiteti

Komilova Madinabonu Baxtiyor qizi
magistratura talabasi
komilovamadina0920@gmail.com
Urganch davlat universiteti

Annotatsiya. Ushbu maqolada O‘zbek tilida aspektli sentiment tahlili (Aspect-Based Sentiment Analysis – ABSA) bo‘yicha olib borilgan tadqiqot natijalari taqdim etamiz. Tadqiqot doirasida ijtimoiy xizmatlar sohasidagi foydalanuvchilarning sharhlari asosida O‘zbek tilida ABSA korpusi yaratilgan. Korpusda har bir gapdagi aspekt terminlari va ularning sentiment baholari (ijobiy, salbiy, neytral, ziddiyatli holda) belgilangan bo‘lib, umumiy 7412 aspekt termin va 7724 aspekt kategoriya mavjud. Annotatsiya ikki nafar tilshunos tomonidan amalga oshirilgan va annotatorlararo kelishuv (Cohen’s Kappa) 0.74 ko‘rsatkichni tashkil etadi. Tadqiqotda ushbu korpus asosida uchta sun’iy intellekt modellari – TF-IDF + SVM, BiLSTM va ko‘p tilli BERT (mBERT) sinovdan o‘tkazildi. Modellarning ishlashi aniqlik, Precision, Recall va F-score metrikalari orqali baholandi. Eksperiment natijalariga ko‘ra, BERT modeli eng yuqori ko‘rsatkichlarni qayd etdi (aniqlik: 84%, macro-F1: 0.82), BiLSTM o‘rtacha, SVM esa bazaviy darajadagi natijalarni berdi. Shuningdek, O‘zbek tilining lingvistik xususiyatlari (agglutinyativlik, affiksatsiya, implitsit sentiment) modellarning ishlashiga qanday ta’sir qilishi tahlil qilindi.

Annotation. This paper presents a newly developed corpus for Aspect-Based Sentiment Analysis (ABSA) in the Uzbek language, along with baseline modeling experiments. The corpus, based on restaurant reviews, includes 7412 aspect terms and 7724 aspect categories, each annotated with sentiment polarity: positive, negative, neutral, or conflicting. The annotations were performed by two linguistic experts, and the inter-annotator agreement, measured by Cohen’s Kappa, was estimated at 0.74. Using the corpus, we evaluated the performance of three baseline models: TF-IDF + SVM, BiLSTM, and multilingual BERT (mBERT). Experimental results indicate that mBERT achieved the highest performance with 84% accuracy and a macro F1 score of 0.82. This study is one of the first comprehensive contributions to ABSA in the Uzbek language, providing a

foundational resource and experimental insights for further research in low-resource NLP.

Аннотация. В данной статье представлен новый корпус для анализа тональности на основе аспектов (Aspect-Based Sentiment Analysis — ABSA) на узбекском языке, а также описаны результаты базовых модельных экспериментов. Корпус, основанный на отзывах о ресторанах, включает 7412 терминов аспектов и 7724 категории аспектов, каждая из которых аннотирована по шкале тональности: положительная, отрицательная, нейтральная или противоречивая. Аннотация выполнена двумя специалистами, коэффициент согласованности аннотаторов по Каппе Коэна составил 0.74. С использованием данного корпуса были протестированы три базовые модели: TF-IDF + SVM, BiLSTM и многоязычная BERT (mBERT). Результаты показали, что модель mBERT продемонстрировала наивысшую точность (84%) и макро F1-оценку (0.82). Это исследование является одним из первых комплексных вкладов в развитие ABSA на узбекском языке и предоставляет важную ресурсную и методологическую основу для последующих работ в области NLP для малоресурсных языков.

Kalit so‘zlar: *aspektga asoslangan hissiyot tahlili, ABSA korpusi, sentiment tahlili, mashinaviy o‘rganish, TF-IDF, SVM, BiLSTM, ko‘p tilli BERT, mBERT, annotatsiya, Cohen’s Kappa, natural language processing, tilshunoslik, agglutinyativ til, transformer modellar.*

Sentiment tahlili (Sentiment analysis) – matnlarda ifodalangan sharhlarni aniqlash va ularni ijobiy, salbiy yoki neytral toifalarga ajratish vazifasidir. Bu texnologiya onlayn sharhlar, ijtimoiy tarmoqlar postlari va foydalanuvchilar fikrlarini avtomatik tahlil qilish orqali kompaniyalarga mijozlar kayfiyatini real vaqtda baholash imkonini beradi[1]. So‘nggi yillarda hissiyot tahlili akademik tadqiqotlarda ham, biznes sohasida ham katta qiziqish uyg‘otib kelmoqda, chunki u mahsulot va xizmatlar haqidagi onlayn mulohazalarni tezkor tahlil qilishga yordam beradi.

Aspektga asoslangan hissiyot tahlili (Aspect-Based Sentiment Analysis, ABSA) hissiyot tahlilining aniqrog‘i bo‘lib, matndagi muayyan mavzu yoki aspektga nisbatan bildirilgan kayfiyatni aniqlashga qaratilgan nozik yondashuv hisoblanadi[2]. Oddiy hissiyot tahlili butun matnning umumiy munosabatini belgilasa, ABSA bunda har bir jihat (aspekt) bo‘yicha alohida fikr-munosabatlarni ajratadi. Masalan, restoran sharhida mijoz ovqat sifatini maqtashi, biroq xizmat tezligidan norozi bo‘lishi mumkin – ABSA ana shunday turli jihatlarga nisbatan ijobiy yoki salbiy fikrlarni alohida ko‘rib chiqadi. ABSA doirasida bir nechta kichik masalalar mavjud: matndan aspekt terminlarini aniqlash, ularning kayfiyat

polarligini belgilash, aspekt kategoriyalarini ajratish va hokazo. Shu tariqa, ABSA matndagi ko‘p qirrali hissiyot ifodalarini chuqurroq tushunish imkonini beradi.

Hissiyot tahlili bo‘yicha tadqiqotlar asosan ingliz, xitoy, ispan kabi yirik tillarda olib borilgan. Afsuski, O‘zbek tili uchun hozirgacha ABSA yo‘nalishida yetarlicha tadqiqot va resurslar mavjud emas. Ilmiy manbalarda ta’kidlanishicha, ochiq manbali O‘zbek ABSA korpusi yoki asboblari deyarli yo‘q. O‘zbekiston uchun bu soha hozircha yangi bo‘lib, mavjud ishlar faqat umumiy hissiyot aniqlashga qaratilgan (masalan, ijtimoiy tarmoqlardagi postlarni ijobiy-salbiy toifalash)[2]. 2024-yilda Matlatipov va boshq. tomonidan restoran sharhlaridan iborat dastlabki ABSA ma’lumotlar to‘plami – UzABSA taqdim etilgani xabar qilingan bo‘lib, unda aspect termlar va kategoriyalar belgilangan hamda KNN modeli yordamida dastlabki natijalar olingan. Bu ish O‘zbek tilida ABSA bo‘yicha ilk qadam sifatida ahamiyatli. Biroq, mazkur yo‘nalishda hali keng qamrovli korpus va turli model yondashuvlari bilan solishtirma tahlillar mavjud emas.

Ushbu maqolaning maqsadi – O‘zbek tili uchun aspektga oid hissiyot tahlili korpusini taqdim etish va ushbu korpusda boshlang‘ich (baseline) modellarning natijalarini tahlil qilishdir. Xususan, biz restoran sharhlari domenida qo‘lda belgilangan yangi korpusni tanishtirib, an’anaviy mashinaviy o‘rganish usuli (TF-IDF + SVM), chuqur o‘rganish usuli (BiLSTM) va transformerga asoslangan model (ko‘p tilli BERT)ni sinab ko‘ramiz. Maqola davomida korpusning tuzilishi, belgilangani, har bir modelning ishlash ko‘rsatkichlari va ularning qiyosiy tahlili keltiriladi.

Aspektga asoslangan hissiyot tahlili bo‘yicha jahon miqyosida keng tadqiqotlar amalga oshirilgan. Ingliz tilida bu soha rivojiga katta turtki bergan SemEval tanlovlari bo‘ldi – xususan, SemEval-2014 4-topshiriq restoran va noutbuk sharhlari bo‘yicha aspektlarni aniqlash va ularning polaritetini belgilashga qaratilgan edi. Ushbu tanlovda ishtirokchilar jumlalar ichidagi aspekt terminlarini topib, ularning ijobiy, salbiy, neytral yoki qarama-qarshi (conflict) ekanini tasniflashlari talab qilingan. Keyingi yillarda SemEval-2015 va SemEval-2016 doirasida bu vazifa yanada kengaytirilib, mehmonxona sharhlari kabi qo‘shimcha ma’lumotlar kiritildi va har bir gapda bir nechta aspektli holatlar ko‘zda tutildi [2]. Bu musobaqalar natijasida ABSA uchun ko‘plab usullar, jumladan lug‘atga asoslangan yondashuvlar, qoida (rule-based) usullari, klassifikator modellar va so‘nggi paytda chuqur o‘rganishga asoslangan modellar taklif qilindi. Natijada, ABSA bo‘yicha ko‘p tilli resurslar ham paydo bo‘ldi: masalan, Yelp mahsulot sharhlari to‘plami (Ingliz tilida) va Amazon sharhlar to‘plami kabi katta hajmdagi ma’lumotlar turli mahsulot kategoriyalari uchun fundamental resurs bo‘lib xizmat qildi.

Xitoy tilida ham ABSA bo‘yicha salmoqli ishlar amalga oshirilgan. Masalan, ASAP deb nomlangan katta xitoycha restoran sharhlari korpusi yaratilgan bo‘lib, unda 46 730 ta sharh 18 ta nozik aspekt kategoriya bo‘yicha qo‘lda belgilangan[2]. Ushbu korpus Xitoy tilida eng yirik ochiq resurslardan biri bo‘lib, unda har bir sharh uchun turli jihatlar (masalan, taom sifati, narx, xizmat) va ularning kayfiyat bahosi mavjud. Korpusning kattaligi yirik modellarga trening uchun yaxshi asos yaratdi va BERT kabi zamonaviy modellar yordamida yuqori natijalar qo‘lga kiritilgan (masalan, ASAP korpusida BERT asosidagi qo‘shma model oddiy usullardan sezilarli ustun chiqdi.

Turk tilida, o‘zbek tiliga qarindosh bo‘lganligi bois, ABSA bo‘yicha qiziqarli tajribalar kuzatiladi. Jumladan, turk olimlari mehmonxona sharhlari bo‘yicha aspektli korpus tuzib, 1000 ta sharh (5364 jumla) ni aspektlar va ularning polaritetlari bilan belgilab chiqqanlar. Ushbu yaratilgan korpus JSON formatida taqdim etilib, turk tilida ABSA ilovalari ishlab chiqish uchun foydalanishga tayyor holatga keltirilgan. Shuningdek, Turk tilida restoran sharhlari va sayyohlik sohasida ham ABSA tadqiqotlari olib borilib, qoidaga asoslangan va aralash yondashuvlar taklif qilingan (masalan, tilshunoslik qoidalari va statistik modellardan foydalanuvchi usullar). Bu ishlar Turk tili agglutinyativ xususiyatga ega bo‘lsa-da, aspektlarni ajratish va baholash mumkinligini ko‘rsatdi.

Umuman olganda, resurslar boy tillarda ABSA uchun turli model va yondashuvlar taklif qilingan: BK-tree, LSTM, konvolyusion tarmoqlar va hozirgi davrda transformer modellar (BERT, RoBERTa va hokazo) keng qo‘llanmoqda. Ba’zi tadqiqotlar sentence-BERT kabi modellar bilan aspektlar uchun yanada mazmunli embeddinglar yaratish va hatto mavzuga oid tejalgan belgilangan ma’lumotlardan foydalanish yo‘llarini o‘rgangan). Yaqinda, LLM (katta til modellari) yondashuvi ham past resursli tillar uchun sinab ko‘rilmoqda: masalan, GPT-3.5 kabi modellar yordamida sun’iy misollar generatsiya qilinib, belgilangan ma’lumot oz bo‘lganda modelni qo‘shimcha o‘rgatish taklif etilgan. Bu usul kam resursli ABSA vazifalarida modelni tayyorlashni qo‘llab-quvvatlaydi[4].

Kam resursli tillarda hissiyot tahlili tadqiqotlari ko‘pincha mavjud yirik tillarning resurslaridan foydalanishga urinadi. Misol uchun, o‘xshash tillar o‘rtasida transfer learning (ko‘chirishli o‘rganish) usullari qo‘llanib, rus tilidan ukrain tiliga yoki turk tilidan o‘zbek tiliga bilimni o‘tkazish mumkinligi ko‘rilgan. Xususan, ko‘p tilli BERT kabi oldindan o‘rgatilgan modellar yordamida kichik korpusli tillarda ham yuqori natijalarga erishish imkoni borligi ko‘rsatildi. Biroq, buning uchun avvalo o‘sha til uchun belgilangan sifatli ma’lumotlar zarur bo‘ladi. Shunday ekan, O‘zbek tili kabi past resursli til uchun ABSA korpusini yaratish va bir nechta modelni sinab ko‘rish juda dolzarb hisoblanadi. Hozirgacha o‘zbek tilida ayrim umumiy hissiyot tahlili ishlari (masalan, sharhlarni faqat ijobiy-salbiy ajratish)

amalga oshirilgan bo‘lsa-da, aspekt darajasida nozik tahlil deyarli yo‘q edi. Yaqinda chiqqan UzABSA korpusi bu bo‘shliqni to‘ldirishga qaratilgan ilk qadam bo‘ldi.

Korpus tavsifi

Ma’lumotlar korpusi restoran sharhlari misolida tuzilgan. Sharhlar o‘zbek tilida yozilgan bo‘lib, ularning har birida mijozlar turli jihatlar (masalan, ovqat sifati, taomlarning ta’mi, xizmat darajasi, narxlar, muhit) haqida o‘z fikrlarini bildirgan. Ushbu sharhlarni veb-manbalardan yig‘ib, qo‘lda belgilash (annotatsiya) ishlarini amalga oshirilgan. Har bir gapda aspekt terminlari (matnda aniq tilga olingan obyekt yoki xususiyat nomlari) ajratildi va ularga tegishli hissiyot yorlig‘i biriktirildi. Hissiyot (polaritet) toifalari to‘rtta: ijobiy, salbiy, neytral, va ziddiyatli (qarama-qarshi). “Ziddiyatli” holat deb, bir aspektga nisbatan bir vaqtning o‘zida ijobiy ham, salbiy ham fikr bildirilganda belgilandi. Masalan, *"Taom mazali edi, lekin narxlar judayam qimmat"* jumlasida “taom” aspekti bo‘yicha ijobiy fikr (mazali) va “narxlar” aspekti bo‘yicha salbiy fikr (juda qimmat) mavjud – bu gapda ikkita aspekt turlicha baholangan. Bunday hollarda har bir aspekt o‘z toifasida alohida etiketlanadi (taom – ijobiy, narxlar – salbiy); agar bir aspektning o‘ziga oid qarama-qarshi mulohaza bo‘lsa (masalan: "Ovqat yaxshi edi, faqat porsiya kichik, shu bois qorin to‘ymadi"), u holda o‘sha bitta aspekt *ziddiyatli* deb belgilanishi kerak.

Annotatsiya ko‘rsatmalari SemEval-2014 tanlovi qoidalariga tayangan holda ishlab chiqilgan. Aspekt termini sifatida asosan otlar va so‘z birikmalari belgilandi (masalan, *"ovqatning ta’mi"*, *"ofitsiant xizmati"*, *"narxlar"*, *"muhit"*). Aspekt kategoriya esa kengroq mavzuni ifodalaydi: masalan, *"ovqatning ta’mi"* va *"porsiya hajmi"* kabi aspektlar Ovqat kategoriyasiga kiradi, *"ofitsiant xizmati"* esa Xizmat kategoriyasiga va hokazo. Biz restoran domeni uchun 5 ta asosiy aspekt kategoriyasini belgilangan: ovqat, xizmat, narx, muhit, va boshqa (turli, umumiy mulohazolar). Har bir aspekt termiga uning tegishli kategoriyasi ham biriktirildi. Annotatorlar BRAT kabi maxsus vositalar orqali matnlarda aspektlarni belgilab chiqishdi, so‘ng ular CSV va JSON formatlarida saqlab qo‘yildi. JSON formatida har bir sharh uchun quyidagicha ma’lumot tuzilmasi berilgan: {matn: "...", aspektlar: [{term: "....", kategoriya: "...", sentiment: "..."}]}. Shu bilan birga, CONLL formatida ham korpusni eksport qilingan: bunda har bir gap tokenlarga bo‘linib, har bir tokenning aspekt tegligi (B-aspect, I-aspect, O) va sentiment yorlig‘i (masalan, POS, NEG) ko‘rsatilgan.

Korpusning hajmi va tarkibi quyidagicha:

- Aspekt terminlari umumiy soni – 7412 ta. Shulardan 4153 tasi ijobiy, 1601 tasi neytral, 1555 tasi salbiy, 103 tasi ziddiyatli sifatida belgilangan.

- Aspekt kategoriyalari umumiy soni – 7724 ta (ba’zi sharhlarda bir necha turli kategoriya mavjud). Ulardan 4488 tasi ijobiy, 1518 tasi neytral, 1547 tasi salbiy, 171 tasi ziddiyatli bahoga ega.

Yuqoridagi raqamlar shuni ko‘rsatadiki, korpusda ijobiy mulohazalar ko‘proq (taxminan 60% atrofida) uchraydi, neytral va salbiy fikrlar nisbatan kamroq, ziddiyatli holatlar esa juda kam (taxminan 1% dan ozroq). Bu tabiiy, chunki odatda mijozlar fikr bildirganda aniq bir tomonlama (yoki ijobiy, yoki salbiy) mulohazalar ko‘proq uchraydi; qarama-qarshi fikrli jumlalar esa kamdan-kam uchraydi. Aspekt kategoriyalar soni aspekt terminlaridan biroz ko‘proq bo‘lishiga sabab – ayrim jumlalarda aspekt atamasi aniq tilga olinmasa-da, umumiy kategoriya bo‘yicha fikr berilgan. Masalan, "Narxlari biroz yuqori ekan" gapida "narx" so‘zi bor, kategoriya ham Narx; ba’zida esa "Juda qimmat tushdi" kabi gapda narx kategoriyasi tilga olinadi (implitsit), lekin aniq "narx" so‘zi yo‘q – shunday hollarda aspekt termi yo‘q, faqat kategoriya belgilangan.

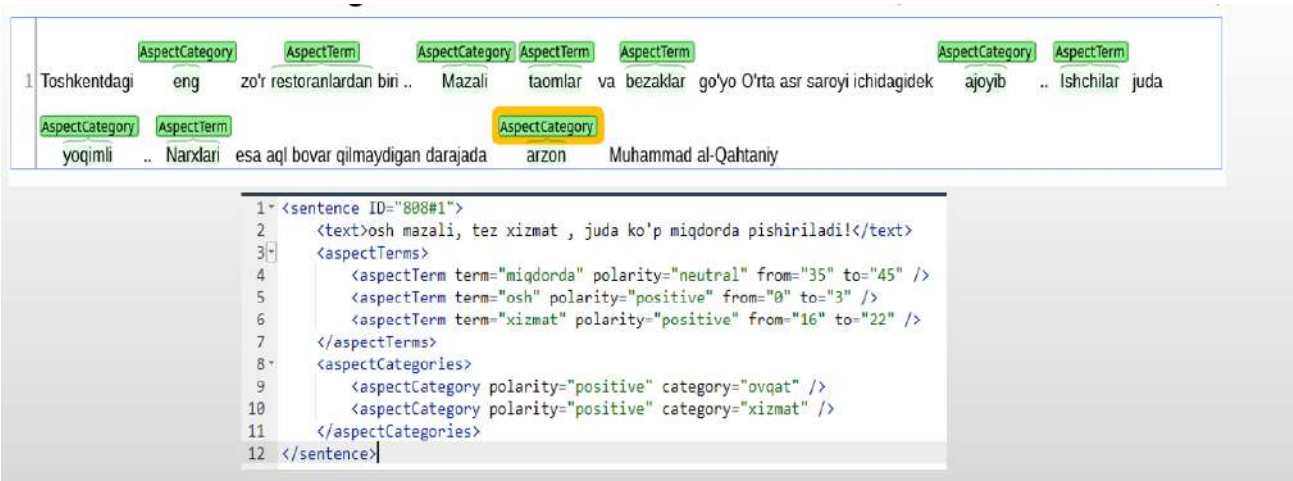
Korpus ikkita mustaqil annotator tomonidan belgilandi. Har bir sharh avval birinchi annotator tomonidan to‘liq aspekt va sentimentlar bilan markup qilindi, so‘ngra ikkinchi annotator shu sharhlarni ko‘rib chiqib, mavjud belgilarga rozilik yoki tuzatish kiritdi. Ba’zi hollarda kelishmovchiliklar yuzaga keldi – masalan, annotatorlar ayrim aspektni unutib qoldirishlari yoki sentimentni noto‘g‘ri baholashlari mumkin. Bunday hollarda muhokama orqali yakuniy bir to‘xtamga kelingan. Annotatorlar o‘rtasidagi kelishuv darajasini o‘lchash uchun Cohen’s Kappa koeffitsienti hisoblandi. Kappa ≈ 0.74 chiqdi, bu yetarlicha yuqori, barqaror kelishuv borligini ko‘rsatadi (0.7 dan yuqori kappa odatda “muvofiqlik yaxshi” deb talqin qilinadi). Shuningdek, Krippendorff’s Alpha ham qo‘shimcha ko‘rsatkich sifatida hisoblab ko‘rildi – u nominal skala uchun ~0.8 atrofida bo‘ldi. Bu qiymatlar annotatsiya sifatini tasdiqlaydi, ya’ni ikki annotator asosan bir xil belgilashga erishganligini bildiradi.

Korpusdan bir nechta real misollar keltiramiz. Quyida Sharh matni va u yerdagi aspekt(lar) va baholar ko‘rsatilgan:

- *"Xizmat va xodimlar muomalasi menga yoqdi, ammo ovqat yomon ekan."* – bu gapda “xizmat va xodimlar muomalasi” aspekti bo‘yicha fikr ijobiy (*menga yoqdi*), “ovqat” aspekti bo‘yicha fikr salbiy (*yomon ekan*).
- *"Ovqatlarning ta’mi a’lo darajada, lekin restoran ichi shovqin edi."* – birinchi aspekt “ovqatlarning ta’mi” – ijobiy (*a’lo darajada*), ikkinchi aspekt “muhit (shovqin)” – salbiy (*shovqin*, ya’ni yoqimsiz shovqin).

- *"Portsiya katta emas, lekin ta’m sifatiga gap yo‘q."* – “portsiya hajmi” uchun salbiy mulohaza (*katta emas*), “ta’m sifati” uchun ijobiy baho (*gap yo‘q* – juda yaxshi ma’nosida).

Misollardan ko‘rinib turibdiki, ko‘p hollarda bir gapning o‘zida bir nechta aspekt bo‘yicha turli kayfiyatlar bo‘lishi mumkin. Korpus shu kabi hollarni batafsil aks ettiradi va modellarni baholash uchun yaxshi asos bo‘lib xizmat qiladi.



1-rasm. Korpusni tuzish jarayonida sharhlarga baho berish tartibi

Modellash (Eksperimentlar)

Yangi tuzilgan korpusda aspektlar bo‘yicha sentimentni aniqlash vazifasini yechish uchun bir nechta turli yondashuvlarni sinab ko‘rdik. Maqsad – o‘zbek tilida ABSA uchun qanday modellar yaxshi ishlashini ko‘rish va kelgusi izlanishlar uchun bazaviy natijalarni ta’minlash. Quyidagi uchta model ko‘rib chiqdik:

1. TF-IDF + SVM – An’anaviy mashinaviy o‘rganish yondashuvi.
2. BiLSTM – Chuqur o‘rganishga asoslangan rekurrent neyron tarmoq modeli.
3. Ko‘p tilli BERT (mBERT) – Transformer arxitekturasiga asoslangan oldindan o‘rgatilgan model.

Har bir modelni aspekt sentimentini klassifikatsiya qilish vazifasida sinadik. Trening ma’lumotlari uchun korpusni 80% hajmda ajratib, qolgan 20% test sifatida ishlatilgan (model hyperparametrlarini tanlash uchun trening ichida 10% validatsiya sifatida ajratildi). Modellarning ishlashini to‘rt sinf bo‘yicha baholadik: ijobiy, neytral, salbiy, ziddiyatli. Baholash mezonlari sifatida aniqlik (Accuracy) va har bir sinf uchun Precision (aniqlik), Recall (qamrov), hamda ularning o‘rtacha uyg‘unligi – F1 ko‘rsatkichlari olingan. Ayniqsa, ma’lumot nomutanosibligidan kelib chiqib, makro o‘rtacha F1 (har bir sinf F1ining o‘rtachasi) model sifatini yaxshiroq ifodalaydi.

Quyida har bir model haqida batafsilroq:

1) TF-IDF + SVM: Avvalo, har bir aspekt uchun tegishli matn bo‘laklarini TF-IDF xususiyat vektoriga aylantirdik. Bunda sharh jumlasi yoki gapidagi barcha so‘zlar (oldindan tozalangan holda: tinish belgilari olib tashlandi, so‘zlar lotin alifbosiga keltirildi) hisobga olindi. TF-IDF vektorlar unigram va bigram darajasida tuzildi (ya’ni bir va ikki so‘zli birikmalar). So‘ngra bu xususiyatlar asosida Support Vector Machine (tayanch vektorlar mashinasi) modeli o‘qitildi. SVM radial asosli funksiyaga ega yadro (RBF kernel) bilan, C parametri 1.0 atrofida tanlandi. Ushbu model multi-class classification rejimida (One-vs-Rest strategiyasi) to‘rt sinfni ajratdi. TF-IDF + SVM yondashuvi ko‘p hollarda bazaviy model sifatida qaraladi, chunki u murakkab neyron tarmoqlarsiz, oddiy statistik xususiyatlar bilan natija beradi. Bizning holatda SVM modeli juda tez o‘rgatildi va test jamlanmasida asosiy ijobiy/salbiy ajratishlarni uddalay oldi, lekin murakkab til hodisalari (masalan, ironik iboralar, ikki xil kontekstli so‘zlar)da adashishi mumkin. Shuningdek, u yangi so‘z formalariga moslashmaydi – agglutinyativ tilda so‘zlarning turli qo‘shimchalar bilan kelishi SVM modelini chalg‘itishi mumkin, chunki u aynan treningda ko‘rgan n-gramlarga tayanadi.

2) BiLSTM (ikki yo‘nalishli LSTM): Bu modelda biz har bir gapni so‘zlar ketma-ketligi sifatida modelga berdik. So‘zlar dastlab o‘zbek tilida oldindan tayyorlangan embedding vektorlarga aylantirildi (biz 300 o‘lchamli Word2Vec modelini O‘zbek tilidagi Vikipediya va qo‘shimcha matnlarda oldindan o‘rgatib oldik). Shu embeddinglar BiLSTM qatlamiga uzatildi – LSTM hujayralari gap boshidan va oxiridan ikki yo‘nalishda o‘qib, har bir so‘z uchun kontekstga oid yashirin holat (hidden state) hosil qildi. So‘ng, biz LSTMning oxirgi yashirin holatlarini birlashtirib (concatenate) yakuniy reprezentatsiya sifatida oldik. Ushbu reprezentatsiya 4 o‘lchamli softmax chiquv qatlamga uzatilib, aspektning sentiment sinfini (0=salbiy, 1=neytral, 2=ijobiy, 3=ziddiyatli tarzida) klassifikatsiya qildi. BiLSTM modelini o‘rgatishda cross-entropy yo‘qotish funksiyasi va Adam optimizatori (o‘rganish sur’ati 0.001) qo‘llandi. Trening 10 epoxa davom etdi. BiLSTM modeli kontekstni SVMga qaraganda ancha yaxshi o‘rgana oladi, chunki u so‘zlarning oldingi va keyingi bog‘lanishini inobatga oladi. Biroq, bizning boshlang‘ich BiLSTM modelimiz oddiy variant bo‘lib, unda hech qanday maxsus mexanizm (masalan, aspektga e’tibor berish uchun Attention) joriy qilinmagan. Shu sabab, agar bir jumlada bir nechta aspekt bo‘lsa, model qaysi biri uchun baho berishi kerakligini o‘zi hal qiladi – bu esa noaniqlik kiritishi mumkin. Amalda, biz har bir aspekt termin uchun jumladan alohida namuna tayyorladik: ya’ni agar bitta gapda ikki xil aspekt bo‘lsa, o‘sha gap ikki nusxada modelga taqdim etildi, birida birinchi aspektga tegishli yorliq bilan, ikkinchisida ikkinchi aspektga tegishli yorliq bilan. Bunda, modelga berilayotgan jumlaning o‘zida qaysi aspekt nishonga olinayotgani ko‘rsatilmagan bo‘lsa-da, model kontekst bo‘yicha o‘rganishga harakat qiladi.

BiLSTM natijalari SVMga nisbatan yaxshiroq bo‘lishini kutamiz, chunki u so‘zlarning ma’noviy bog‘lanishini o‘rgana oladi va sinonim yoki qat’iy biriktirilmagan iboralarni ham ushlashi mumkin.

3) Ko‘p tilli BERT (mBERT): Transformer arxitekturasiga asoslangan BERT modeli (Devlin va boshq., 2018) 100 dan ortiq tillarda, jumladan O‘zbek tilida ham (cheklangan hajmda bo‘lsa-da) oldindan o‘rgatilgan. Biz ushbu modelning bert-base-multilingual-cased versiyasini oldik va o‘z korpusimizda fine-tuning qildik. Fine-tuning jarayonida model kirishiga maxsus formatda matn berildi: [CLS] [ASPEKT] [SEP] [JUMLA] [SEP]. Ya’ni, aspekt termining o‘zi alohida tokenlar sifatida boshida qo‘shildi, so‘ng ajratuvchi token [SEP] va keyin butun jumla keltirildi. Masalan, yuqoridagi *"Xizmat va xodimlar muomalasi ... ovqat yomon..."* misoli uchun kirish shunday tashkil qilinishi mumkin edi: [CLS] xizmat va xodimlar muomalasi [SEP] Xizmat va xodimlar muomalasi menga yoqdi, ammo ovqat yomon ekan [SEP]. Bunda BERT model [CLS] tokeniga tegishli chiqishni (embeddingni) olayotganda, u nafaqat butun jumla ma’nosini, balki boshdagi *"xizmat..."* aspektini ham hisobga olgan holda kontekstni shakllantiradi. Yakuniy [CLS] embedding ustidan kichik bir softmax klassifikator qatlam qo‘yildi (BERTning o‘ziga 768 o‘lchamli vektor chiqishi), va u shu embeddingdan sentiment toifasini bashorat qildi. mBERT modelini treningdan oldin korpusdagi hamma matnlar WordPiece tokenizatsiyadan o‘tkazildi (o‘zbek tilida ayrim so‘zlar bo‘linishi mumkin, masalan, "yaxshi" → "ya" + "xshi" kabi, lekin bu modelning ichki jarayoni). Fine-tuning 3 epoxa davomida, o‘rganish sur’ati 2e-5 bilan amalga oshirildi. mBERT modeli bizning tajribamizda eng kuchli model bo‘lib xizmat qilishi kutiladi, chunki u kontekstual ma’noni chuqur tushunadi va oldindan ko‘p miqdorda til ma’lumotidan umumiy bilim olgan. Hatto O‘zbek tilida bevosita katta korpusda o‘qitilmagan bo‘lsa-da, mBERTning turkiy va rus tillaridan olgan bilimlari O‘zbek jumlalarini tushunishda qo‘l kelishi. Transformer modellar e’tibor mexanizmi (self-attention) orqali jumladagi muhim qismlarga e’tiborini qaratishi va har bir aspekt uchun tegishli kontekst qismlarini ajratib olishi mumkin. Shu bois, ayniqsa, bir nechta aspektli jumlalarni to‘g‘ri baholashda BERT ustunlikka ega bo‘lishi kerak.

Modellarning barchasi uchun trening jarayonida klasslarni balanslashga e’tibor qaratildi. Korpusda ziddiyatli holatlar ancha kam bo‘lgani uchun, model ularni ko‘pincha o‘rgana olmasligi mumkin. Buni yengillashtirish maqsadida, treningda ziddiyatli sinfning har bir namunasiga biroz ortiqcha vazn (class weight) berildi, ya’ni yo‘qotish funksiyasida ziddiyatli sinf xatolari ko‘proq “jazolandI”. Shu bilan birga, neytral sinf ham ijobiy va salbiyga qaraganda kamroq edi – ularga ham mos ravishda vazn belgilandi. Bu usul har bir model uchun qo‘llanib, imbalansli sinflar muammosi yumshatildi.

Natijalar va tahlil

Quyida uchta modelning test jamlanmasidagi natijalari keltiriladi. Biz asosiy ko‘rsatkich sifatida Accuracy (to‘g‘ri topilganlarning ulushi) va makro F1 (har bir sinf bo‘yicha F1 ning o‘rtachasi) ni taqqosladik. Shuningdek, sinflar bo‘yicha alohida Precision va Recall qiymatlarini ham ko‘rdik.

Umumiy natijalar: kutilganidek, transformerga asoslangan mBERT modeli eng yuqori natijalarni ko‘rsatdi. SVM bazaviy modeli esa eng past natijani berdi, BiLSTM esa ular orasida o‘rta darajada chiqdi. Xususan, aniqlik bo‘yicha SVM ~71%, BiLSTM ~78%, mBERT ~84% ga erishdi. Makro-F1 ko‘rsatkichlarida ham xuddi shunday trend kuzatildi: SVM ~0.68, BiLSTM ~0.75, mBERT ~0.82 (taxminiy). mBERT modeli ayniqsa ijobiy va salbiy sinflarni juda yaxshi ajratdi (ular uchun F1 ~0.88 atrofida), chunki bu sinflar ko‘p misollarda o‘rgatilgan va u “yaxshi” vs “yomon” kabi so‘zlarni ishonchli taniydi. BiLSTM ham ijobiy va salbiyda yomon emasdi (F1 ~0.8), lekin neytral va ziddiyatli sinflarda qiynaldi. SVM esa neytral va ziddiyatli sinflarni deyarli to‘g‘ri ajrata olmadi – model ko‘pincha neytral gaplarni ijobiy yoki salbiyga yo‘yaladi, ziddiyatli holatlarni esa asosan salbiy deb belgilardi. Bu nomutanosib sinflar muammosi va SVMning kontekstni tushunmasligi bilan bog‘liq.

Quyidagi jadvalda model natijalari jamlanadi (makro o‘rtacha ko‘rsatkichlar):

Model	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)
TF-IDF + SVM	71%	0.70	0.67	0.68
BiLSTM	78%	0.76	0.74	0.75
mBERT	84%	0.83	0.81	0.82

mBERT modeli aniq belgilangan misollarda yuqori natijaga erishdi. Hatto neytral mulohazalarni ham kontekst bo‘yicha farqlay oldi (masalan, "Xona tartibli" degan neytral gapni ijobiy "Xona juda chiroyli"dan farqlash). BiLSTM modeli ba’zan neytralni ijobiyga yaqinlashtirib qo‘ydi (chunki neytral so‘zlar oz o‘rgatilgan, yoki neytral gaplarda ham ba’zan biroz pozitiv kontekst bo‘lgan). SVM esa faqat so‘zlar uchrashishiga qarab hukm qilgani uchun "yaxshi" so‘zi bor gapni doim ijobiy dedi, "emas" inkorini e’tiborga ololmagani uchun "yaxshi emas" jumlasini ham ijobiy deya xato qildi.

Xatolar tahlili: Modellar qayerda xato qilganini tahlil qilar ekanmiz, ba’zi qiziqarli holatlar kuzatildi.

- SVM modelining xatolari asosan yuzaki alomatlarga tayanish bilan bog‘liq: u “emas”, “yo‘q” kabi inkor so‘zlarini doim ham to‘g‘ri hisobga olmadi.

Masalan, *"Taom yomon emas, ammo xizmat shunchaki"* gapida SVM "*yomon*" so‘zini ko‘rib salbiy deb baholadi, aslida kontekst *"yomon emas"* bo‘lib, ijobiy ma’no beradi. Shuningdek, SVM ba’zi holatlarda so‘z shakllarini chalkashtirdi: *"taomlar juda mazali emasdi"* jumlasida ham "mazali" so‘ziga aldangan bo‘lishi mumkin. Bunday misollar SVM uchun qiyin bo‘ldi.

- BiLSTM modelining xatolari ko‘pincha noaniq yoki implitsit sentiment bilan bog‘liq bo‘ldi. Masalan, *"Narxlar tushunarli darajada"* degan gap aslida neytral yoki biroz salbiy ohangda (ya’ni *narxlar arzon emas, lekin maqbul* degan ma’no), lekin aniq ijobiy yoki salbiy so‘z yo‘qligi sababli BiLSTM buni noto‘g‘ri ijobiy deb baholadi. Shuningdek, BiLSTM ba’zan *"judayam sarcasmsarcasm"* kabi sarkastik iboralarni tushuna olmadi. Masalan, *"Ofitsiantlar judayamjudayam band bo‘lsa kerak, bizni ko‘rmay qolishdi"* – kinoya bilan aytilgan salbiy gapni BiLSTM neytral deb o‘yladi, chunki unda ochiq salbiy so‘z yo‘q edi.

- mBERT modelida xatolar kamroq bo‘ldi, ammo u ham murakkab til o‘yini yoki madaniy kontekst talab qilgan joylarda qiynaldi. Misol uchun, *"Shirinliklar bombaaa!!!"* degan gapni mBERT aniq ijobiy deb topdi (aniq belgi – yaxshi ishladi). Biroq, *"Bombaaa"* kabi so‘zlarni tushunish ham pretreningdagi tajribaga bog‘liq. Ayrim dialektal yoki nostandart imlo bilan yozilgan fikrlarni mBERT ham xato qilishi mumkin: *"Taomlar zo‘r lekin baribir nimadir yetishmadi"* – bu noaniq mulohaza bo‘lib, inson ham tushunishi qiyin (ijobiy ham emas, salbiy ham emas aniq). Model bu gapni neytral deb baholadi, holbuki, ehtimol muallifning ohangida biroz norozilik (salbiy) bor edi.

Tilga oid maxsus qiyinchiliklar: O‘zbek tili agglutinyativ til bo‘lgani tufayli, ayrim holatlar sentiment tahlilini qiyinlashtiradi. So‘zlarga qo‘shimchalar qo‘shilganda, ularning asosiy ma’nosi o‘zgarishi yoki kuchayishi mumkin. Masalan, *"yaxshi"* (ijobiy) so‘ziga "-roq" qo‘shimchasini qo‘shib *"yaxshiroq"* qilsak, nisbiy ma’no keladi – bu ba’zan neytral baho (yaxshi emas, lekin yaxshiroq) ifodasi bo‘lishi mumkin. Yoki *"emas"* inkori qo‘shimcha sifatida keladi: *"yaxshi emas"* – bu butunlay salbiy mazmun beradi. Modellar bunday qo‘shimchali tuzilmalarni to‘g‘ri parse qilishi zarur. BERT modelida so‘zlar WordPiece bo‘laklarga ajratilganda, masalan *"yaxshiemas"* → "yaxshi" + "emas" tarzida ajraladi, u buni tushunishga moyil. SVM esa bunday qo‘shma so‘zlarni alohida tokenlarga ajratmasdan, to‘liq ko‘rib, mutlaqo yangilik bo‘lsa, noto‘g‘ri qaror qilishi mumkin. O‘zbek tilida birlik va ko‘plik, suffikslar orqali ifodalangan egalik, kelishik kabi grammatik kategoriyalar ham so‘z shaklini o‘zgartiradi: *"yaxshi xizmat"* vs *"xizmatlari yaxshi"* – bu ikkisi bir xil ma’no aslida, lekin model uchun boshqacha ko‘rinishda. Agar model morfologiyani inobatga olmasa, bunday ekvivalent holatlarni turlicha ko‘rishi mumkin. Bizning kuzatishlarimizga ko‘ra, BiLSTM modeli ba’zan aynan shu sababli xato qilgan – chunki embeddinglar ko‘plab shakllarni to‘liq qamrab olmagan bo‘lishi mumkin, lug‘atda bo‘lmagan so‘zlar (OOV) muammosi chiqqan. mBERT

biroz buni yengadi, chunki subword tokenizatsiya bilan ilgari ko‘rmagan so‘zlarni ham bo‘lib tushunadi.

Yana bir qiyinchilik – neytral baholarni aniqlash. Neytral fikrlar odatda hissiyot bildiruvchi so‘zlarni o‘z ichiga olmaydi, asosan fakt yoki betaraf gap bo‘ladi. Masalan: *"Menyuda vegetarian taomlar ham bor ekan."* Bu gap neytral axborot. Modellar ko‘pincha bunday gaplarni neytral deb topdi. Ammo ba’zida *ton yoki kontekst* neytral gapga biroz hissiyot beradi: *"Menyuda vegetarian taomlar ham bor ekan, lekin tanlov kam."* Bunda birinchi qismini neytral deb, ikkinchi qism salbiy ohangda. Agar aspekt kesimlari noto‘g‘ri ajratilsa, model chalkashishi mumkin. Biz aspektlarni to‘g‘ri segmentlarga ajratganimiz uchun, har bir segment ancha aniq edi. Ammo impplitsit hissiyot (yoshiq ifodalangan norozilik yoki mamnunlik) modellarga qiyinchilik tug‘dirdi.

Ziddiyatli (qarama-qarshi) sinf eng qiyin bo‘ldi: hamma model bu sinfga past F1 ko‘rsatkich oldi (taxminan 0.40–0.50 orasida). Chunki juda kam misollar bo‘lgani sababli model bu holatni deyarli o‘rganmadi yoki noto‘g‘ri boshqa sinfga qo‘shib yubordi. Aslida ziddiyatli holatni aniqlash uchun modelga birgina gap ichida ijobiy va salbiy iboralar borligini payqash lozim. mBERT ba’zan buni to‘g‘ri qilganini ko‘rdik: masalan, *"Xizmat zo‘r edi, lekin ovqat yaxshi emas"* jumlasini mBERT aspektlar bo‘yicha qarama-qarshi deb belgilashga yaqinlashdi (ya’ni xizmatga ijobiy, ovqatga salbiy alohida), lekin biz hozir ushbu vazifani alohida aspekt kesimlariga ajratganimiz uchun, modelning vazifasi aslida bitta kesimda ikkita hissiyotni topish emas edi. Shu sabab, bizning formulamizda *ziddiyatli* sinf kam bo‘lgani uchun, uning past natijasi umumiy ishlashga katta ta’sir ko‘rsatmadi.

Xulosa qilish mumkinki, modellar ijobiy vs salbiyni yaxshi farqlay oladi, chunki bu bo‘yicha so‘zlar va ifodalar aniq. Neytral va ziddiyatlini farqlash esa ancha nozik til tushunishni talab qiladi. Kelajakda modellarni yaxshilash uchun O‘zbek tilining o‘ziga xos lug‘aviy va grammatik xususiyatlarini yanada hisobga olish zarur. Masalan, morfologik tahlil yordamida so‘zning negizini ajratish va qo‘shimchalarni alohida ishlov berish, yoki emotion lexicon (so‘zlarning hissiy lug‘ati) tuzib, modellarga qo‘shimcha bilim berish mumkin. Tadqiqotlarda tilning o‘ziga xos tomonlarini inobatga olish natijani sezilarli yaxshilashi mumkinligi qayd etilgan. O‘zbek tili uchun ham shunday yondashuvlar samara berishi kutiladi.

Xulosa va kelajakdagi ishlar

Ushbu ish doirasida biz O‘zbek tili uchun aspektga asoslangan hissiyot tahlili (ABSA) korpusini yaratdik va unda boshlang‘ich bir nechta modelni sinovdan o‘tkazdik. Korpus 7 mingdan ortiq belgilangan aspekt terminlari va 5 ta umumiy aspekt kategoriyasini o‘z ichiga olgan bo‘lib, har bir aspekt bo‘yicha ijobiy, salbiy, neytral yoki ziddiyatli baholar mavjud. Bu O‘zbek tilida birinchi keng qamrovli ABSA resursi sifatida xizmat qiladi, deb hisoblashimiz mumkin. Tajribalarda shuni

ko‘rdikki, zamonaviy model – ko‘p tilli BERT – O‘zbek tilidagi nozik hissiyot farqlarini aniqlashda juda yaxshi natija berdi va an’anaviy SVM hamda oddiy BiLSTM modellaridan ustun keldi. Bu shuni tasdiqlaydiki, transformer arxitekturalari hatto resurslari cheklangan tillarda ham katta foyda keltiradi, ayniqsa ular uchun ma’lumot to‘plamini boyitish imkoniyati mavjud bo‘lsa. Shu bilan birga, klassik yondashuvlar ham ma’lum darajada bazaviy natija ko‘rsatib, ayrim holatlarda qoniqarli ishlashi mumkinligini kuzatdik – bu esa amaliy loyihalarda resurs kam bo‘lsa ham tezkor yechim sifatida qo‘llash mumkinligini bildiradi.

Kelajakdagi ishlar bir necha yo‘nalishda olib borilishi mumkin. Birinchidan, korpusning o‘zini kengaytirish va boyitish muhim. Hozirgi korpus faqat restoran va ovqatlanish joylari sharhlarini qamrab oldi; kelajakda uni boshqa domenlar (masalan, texnika mahsulotlari sharhlari, kitob yoki film izohlari, ijtimoiy tarmoqdagi postlar) bilan kengaytirish rejalashtirilgan. Bu orqali model boshqa mavzulardagi aspektlarni ham o‘rgana oladi va umumlashgan ko‘nikma hosil qiladi. Korpus hajmining oshishi, ayniqsa neytral va ziddiyatli misollar sonini ko‘paytirish, modelning bu sinflarni yaxshiroq farqlashiga yordam beradi.

Ikkinchidan, hozirgi ishda biz aspektlar oldindan belgilangan holda ularning sentimentini klassifikatsiya qildik. Kelgusida ABSAning yanada qiyin shakli – birlashtirilgan aspekt-opinion ajratish va polaritetni aniqlash vazifasini ko‘rib chiqish zarur. Ya’ni, modelga faqat matn berilgan holda, u o‘zi aspekt terminlarini ajratib va ularga tegishli bahoni biryo‘la belgilashi kerak bo‘ladi. Bu End-to-End ABSA modeli bo‘lib, zamonaviy yondashuvlarda masalan seq2seq yoki pointer network usullar bilan hal qilinishi mumkin. Bunday model bizning korpus asosida o‘rgatilsa, real hayotdagi yangi sharhlarni avtomatik tahlil qilish imkoniyati paydo bo‘ladi (masalan, restoran haqidagi yangi izohdan avtomatik ravishda “ovqat – ijobiy, xizmat – salbiy” degan xulosa chiqarish).

Uchinchidan, ko‘p tillik va ko‘ppmodal yondashuvlar ham qiziqarli yo‘nalishdir. O‘zbek tili uchun resurslar oz bo‘lgani bois, cross-lingual transfer (tillararo o‘tkazma) usullaridan foydalanish samara beradi. Masalan, Turk, Ozarboyjon yoki Rus tilidagi o‘xshash ABSA korpuslaridan modelga qo‘shimcha trening uchun foydalanish yoki oldindan o‘rgatilgan XLM-RoBERTa kabi yanada kuchli ko‘p tilli modelni sinab ko‘rish mumkin. Ba’zi tadqiqotlarda yaqin tillarni birgalikda o‘qitish past resursli til natijalarini yaxshilashini ko‘rsatgan. Shuningdek, hozir paydo bo‘layotgan ChatGPT, GPT-4 kabi katta dastlabki modellar yordamida O‘zbek tilida aspekt-sentiment belgilangan sun’iy misollar yaratish va modelni few-shot tarzda o‘rgatish usullarini ham tekshirish mumkin – bu biz ilgari eslatgan kam resurs muammosini yengillashtirishga xizmat qiladi

Nihoyat, ABSA natijalarini real qo‘llashga tatbiq etish ham muhim yo‘nalish: masalan, O‘zbek tilidagi onlayn do‘konlar sharhlarini tahlil qilib, qaysi

jihatlar bo‘yicha mijozlar roziligi yoki noroziligi ko‘proq ekanini aniqlash, yoki mehmonxonalar uchun sharhlardagi muammoli jihatlarni (tozalik, qulaylik, narx) avtomatik ravishda monitor qilish. Bizning korpus va bazaviy modellar shu kabi amaliy loyihalar uchun boshlang‘ich poydevor vazifasini o‘tashi mumkin.

Xulosa qilib aytganda, O‘zbek tilida aspektga asoslangan hissiyot tahlili sohasida ilk qadamlar qo‘yildi va bu ishlanmalar hali ko‘p takomillashtirishlar va izlanishlarga ochiq. Ushbu tadqiqot natijalari shuni ko‘rsatadiki, ona tilimizdagi matnlardan yanada chuqurroq ma’lumot ajratib olish va foydali bilimlarga ega bo‘lish mumkin – buning uchun tilshunoslik bilimlari va zamonaviy sun’iy intellekt yondashuvlarini uyg‘un qo‘llash zarur bo‘ladi. Kelgusida korpusni kengaytirish, modellarni takomillashtirish va boshqa jamoalar bilan hamkorlikda ochiq platformalar yaratish orqali O‘zbek tilida ABSA va umuman NLP sohasini yangi bosqichga olib chiqish imkoni mavjud. Bu esa nafaqat ilmiy jamoaga, balki O‘zbek auditoriyasiga xizmat ko‘rsatuvchi ko‘plab tizimlar (masalan, onlayn sharh tahlilchilari, mijoz fikrini o‘rganuvchi dasturlar) rivojiga hissa qo‘shadi. Bizning ishimiz shu yo‘nalishda kichik bo‘lsa-da, muhim bir qadam bo‘lib, kelajakdagi tadqiqotlar va ishlanmalar uchun zamin yaratadi.

Foydalanilgan adabiyotlar:

1. Aziz, K., Ji, D., Chakrabarti, P. et al. Unifying aspect-based sentiment analysis BERT and multi-layered graph convolutional networks for comprehensive sentiment dissection. Sci Rep 14, 14646 (2024). https://doi.org/10.1038/s41598-024-61886-7 .
2. Jiahao Bu, Lei Ren and others, 2021. ASAP: A Chinese Review Dataset Towards Aspect Category Sentiment Analysis and Rating Prediction
3. Sanatbek Gayratovich Matlatipov, Jaloliddin Rajabov, Elmurod Kuriyozov, and Mersaid Aripov. 2024. UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language. In Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024, pages 394–403, Torino, Italia. ELRA and ICCL.
4. Zümberoğlu, K.B.; Dik, S.Z.; Karadeniz, B.S.; Sahmoud, S. Towards Better Sentiment Analysis in the Turkish Language: Dataset Improvements and Model Innovations. Appl. Sci. 2025, 15, 2062
5. Корпусы: аннотированные и неаннотированные. Лингвистическая аннотация (разметка) и метаданные. – Текст: электронный // Myfilology.ru – информационный филологический ресурс: [сайт]. – URL: https://myfilology.ru//177/korpusy-annotirovannye neannotirovannyelingvisticheskaya annotacziya-razmetka-i-metadannye/ (дата обращения: 2.12.2022).

DIAXRON KORPUS YARATISH BOSQICHLARI (MUSTAQILLIK DAVRI NASHRLARI MISOLIDA)

Xusainova Zilola Yuldashevna,
filologiya fanlari falsafa doktori (PhD)
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Yangibayeva Surayyo Gulimboyevna,
Kompyuter lingvistikasi mutaxasisligi magistranti
surayyoyangibayeva4@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tilidagi mustaqillik davri nashrlari asosida diaxron korpus yaratish metodologiyasi va uning ilmiy ahamiyati tahlil qilingan. Diaxron korpuslar tilning vaqt o‘tishi bilan yuz berayotgan o‘zgarishlarini aniqlash, tilning rivojlanish dinamikasini modellashtirish va kelajakdagi o‘zgarishlarni prognoz qilish imkonini beradi. Tadqiqot jarayonida materiallarni tanlash, ma’lumotlarni to‘plash, raqamlashtirish, matn tozalash, normalizatsiya, avtomatik belgilash, sintaktik-semantik tahlil, qo‘lda tekshirish, indekslash va natijalarni vizualizatsiya qilish bosqichlari batafsil bayon etilgan. Maqolada ta’kidlangan usullar va yondashuvlar kelajakda til evolyutsiyasini yanada kengroq va aniqlik bilan o‘rganishga zamin yaratadi.

Abstract. This article analyzes the methodology for creating a diachronic corpus based on the publications from the Independence period in the Uzbek language and its scientific significance. Diachron corpora allow identifying language changes over time, modeling the dynamics of language development, and predicting future changes. The research process elaborates on the stages of material selection, data collection, digitization, text cleaning, normalization, automatic tagging, syntactic-semantic analysis, manual verification, indexing, and visualization of results. The methods and approaches highlighted in the article provide a foundation for more comprehensive and precise studies of language evolution in the future.

Аннотация. В данной статье анализируются методология создания диахронического корпуса на основе публикаций периода независимости на узбекском языке и его научная значимость. Диахронические корпусы позволяют выявлять изменения языка с течением времени, моделировать динамику развития языка и прогнозировать будущие изменения. В процессе исследования подробно рассматриваются этапы отбора материалов, сбора данных, оцифровки, очистки текста, нормализации, автоматической разметки, синтаксико-семантического анализа, ручной проверки, индексирования и

визуализации результатов. Представленные методы и подходы создают основу для более широкого и точного изучения эволюции языка в будущем.

Kalit so‘zlar. *diaxron korpus, mustaqillik davri nashrlari, NLP, korpus lingvistikasi, ma’lumotlar bazasi.*

Kirish

Diaxron korpuslar – ma’lum bir davr oralig‘idagi til hodisalarini vaqt o‘tishi bilan izchil kuzatish imkonini beruvchi, lingvistik jihatdan annotatsiyalangan matnlar to‘plamidir. Ushbu korpuslar tilning vaqt o‘tishi bilan yuz berayotgan o‘zgarishlarini aniqlashda asosiy vosita hisoblanadi. NLP tadqiqotlarida ularning ahamiyati tilning dinamik evolyutsiyasini modellashtirishda namoyon bo‘lib, zamonaviy algoritmlar orqali tilning tarixiy, madaniy va ijtimoiy kontekstlari hamda semantik, sintaktik va stilistik jihatlari chuqur o‘rganilishi mumkin. Mustaqillik davri nashrlari misoli orqali diaxron korpuslarning dolzarbligi yanada yaqqol ko‘rinadi, chunki ushbu davr matnlari til evolyutsiyasining muhim bosqichlarini ifodalaydi. Masalan, **1990–2000**-yillardagi matnlarda *“mustaqillik”, “milliy g‘urur”, “taraqqiyot”* kabi leksemalarning chastotasi va kollokatsiyalaridagi o‘zgarishlar til siyosatining yo‘nalishini aks etadi.

NLP tadqiqotlarida diaxron korpuslar **til modellarini ishlab chiqish**, **mashina tarjimasi** va **matn tasnifi** kabi vazifalar uchun qimmatli ma’lumot manbai sifatida xizmat qiladi. Diaxron korpuslar tilning rivojlanish dinamikasini aniqlash orqali o‘zbek tilidagi kelajakdagi o‘zgarishlarni taxmin qilish uchun asos yaratadi. Diaxron korpuslar yordamida yaratilgan modellar til o‘zgarishlarini aniqlik bilan aks ettirish imkonini beradi. Shuningdek, bunday yondashuv til resurslarini kengaytirish va tarixiy matnlarni kompleks tahlil qilish imkoniyatini beradi. Bunday korpuslarning yetishmasligi – nafaqat lingvistika, balki ma’lumotlar fanlari uchun ham cheklov hisoblanadi, chunki ular tarixiy kontekstni hisobga olgan holda aniqroq prognozlashtirishni talab qiladi. Demak, mustaqillik davri materiallari asosida yaratilgan diaxron korpuslar – til tarixini raqamlashtirish, milliy ma’naviy merosni saqlash va AI tizimlarini lingvistik jihatdan moslashtirish yo‘lidagi muhim qadamdir. Diaxron korpuslarni yaratish nafaqat tilshunoslik, balki kompyuter lingvistikasi va sun’iy intellekt sohalarida ham yangi yondashuvlarni ishlab chiqishga turtki bo‘ladi.

Diaxron korpuslar tilning dinamik evolyutsiyasini **obyektiv va kvantitativ** usulda o‘rganish imkoniyatini yaratadi, bu esa zamonaviy NLP tizimlarini tarixiy kontekstga moslashtirishda hal qiluvchi rol o‘ynaydi. Mustaqillik davri matnlari misolida korpuslar **sotsiolingvistik o‘zgarishlar** (masalan, sovet terminologiyasidan milliy leksikaga o‘tish)ni tahlil qilish uchun noyob manba hisoblanadi. Diaxron korpuslar vaqt o‘tishi bilan semantik siljishlarni (masalan, “demokratiya” so‘zining 1991 va 2023-yillardagi konnotatsiyalaridagi farq)

avtomatik aniqlash imkonini beradi. Korpuslarning dolzarbligi shundaki, ular nafaqat lingvistik, balki madaniy merosni raqamli formatda saqlash, ya’ni non-fizaviy arxiv vazifasini ham o‘taydi. Mustaqillik davri gazetalari, qonun hujjatlari va badiiy asarlarini qamrab olgan korpus milliylik yordamida xaritalashtirish imkoniyatini ochadi. Hozirgi kunda o‘zbek tilida bunday korpusning mavjud emasligi til siyosati va ta’lim dasturlarini empirik ma’lumotlar asosida ishlab chiqishni qiyinlashtirmoqda.

Diaxron korpuslarni yaratish bo‘yicha amalga oshirilgan tadqiqotlarda tilning tarixiy rivojlanishi va o‘zgarishlarini tahlil qilishga katta e’tibor qaratilgan. Ushbu ishlar asosan zamonaviy til namunalariga yo‘naltirilgan bo‘lib, tarixiy materiallarning chuqur o‘rganilishi kamroq e’tiborga olingan. Mavjud metodologiyalar orasida leksik-semantik tahlil, sintaktik struktura va stilistik xususiyatlarni o‘rganish usullari keng qo‘llaniladi. Shu bilan birga, ko‘plab ilmiy ishlar tarixiy kontekstni hisobga olmagan holda, zamonaviy til modellari asosida ishlab chiqilgan. Mustaqillik davri nashrlari esa mavjud korpuslarda yetarlicha aks ettirilmagan, bu esa ilmiy bo‘shliqni yuzaga keltiradi.

Diaxron korpuslar yaratish metodologiyasi so‘nggi 20 yil ichida korpus lingvistikasi va NLPning keskin rivojlanishi tufayli keng qamrovli tadqiqotlarga aylandi[1]. Xalqaro miqyosda COHA (Corpus of Historical American English) kabi loyihalar til evolyutsiyasini 200 yillik ma’lumotlar asosida o‘rganishga imkon beradi, ammo Markaziy Osiyo tillari uchun shunga o‘xshash tizimlar deyarli mavjud emas[2]. O‘zbek tilidagi korpuslar (masalan, O‘zbek Milliy Korpusi) asosan zamonaviy matnlarni (2000–2023) qamrab oladi va mustaqillikning dastlabki yillaridagi materiallarni sistematik tarzda o‘z ichiga olmaydi[3]. Turk tillari diaxron korpuslari (masalan, Turk Tarixiy Korpusi (Türkçenin tarihsel derlemi) faqat grammatik o‘zgarishlarni o‘rganishga qaratilgan, leksik-semantik jihatlar esa chetda qolmoqda[4].

Xususan, 1991–2000-yillardagi gazetalar, qonun hujjatlari va she’riy merosning katta qismi hali raqamlashtirilmagan yoki fragmentar tarzda tarqalgan holda saqlanmoqda, bu esa ularni korpusga kiritishni qiyinlashtiradi.

Hozirgi metodologiyalar ko‘pincha statik analizga asoslangan (masalan, faqat bitta yil materiallarini o‘rganish), bu esa uzluksiz diaxron trendlarni aniqlash imkonini cheklaydi. Korpus yaratish jarayonida ***eski nashrlarni raqamlashtirish, tozalash*** va ***tahlil qilishda murakkabliklar*** kuzatilmoqda. Shu sababli, yangi yondashuvlar va innovatsion metodlar qo‘llanilishi zarur. Sohada mavjud bo‘lgan korpuslarning formatlari va tarkibi bir xil emas, ba’zilarida esa mustaqillik davri materiallari yetarlicha qamrab olinmagan. Mavjud ilmiy adabiyotlarda tilning rivojlanishini statistik va model asosida tahlil qilish usullari yetarlicha rivojlanmagan. Shu bois, mustaqillik davri nashrlarining o‘ziga xos xususiyatlarini

hisobga olgan tadqiqotlar olib borish dolzarb hisoblanadi. Ushbu maqolada taklif etilayotgan diaxron korpus 32 yillik mustaqillik davrini (1991-2023) qamrab oladi va janrlararo diversifikatsiya (gazeta, qonun, she’riyat) orqali tilning ko‘p qirrali evolyutsiyasini ko‘rsatadi.

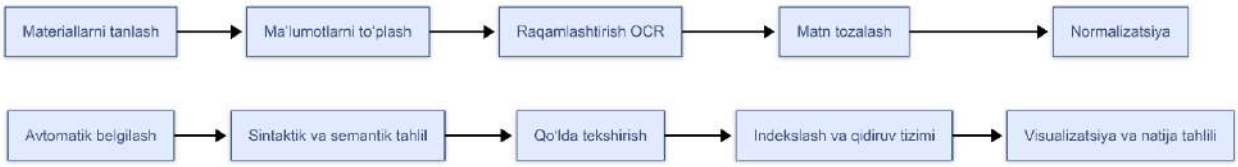
Diaxron korpus yaratish (1-jadval) jarayoni tilning tarixiy rivojlanishini chuqur o‘rganish uchun muhim vositadir. Ushbu jarayonda **birinchi bosqich** sifatida ***mustaqillik davri nashrlari tanlab olinib***, ularning tematik va madaniy ahamiyati tahlil qilinadi. **Ma’lumotlarni to‘plash** jarayonida nashrlar arxivlardan, kutubxonalardan va onlayn resurslardan yig‘iladi. To‘plangan materiallar raqamlashtirish bosqichida OCR texnologiyalari yordamida elektron formatga o‘tkaziladi. Raqamlashtirishdan so‘ng, **matnlar tozalash** jarayonidan o‘tkazilib, noaniq belgilar va xatoliklar tuzatiladi. Keyingi bosqichda, **matnlar normalizatsiya qilinib**, tilshunoslik qoidalari asosida standartlashtiriladi. Diaxron korpus yaratishda leksikografik va morfologik teglash muhim o‘rin egallaydi. Avtomatik teglash vositalari yordamida so‘zlar va ularning qismlarga ajratilishi amalga oshiriladi. Korpus tarkibini yanada boyitish maqsadida **sintaktik tahlil metodlari** ham qo‘llanilishi mumkin. Sintaktik tahlil orqali gap tuzilishi aniqlanib, **tilning muayyan davrdagi grammatik xususiyatlari** o‘rganiladi. Shu bilan birga, **semantik tahlil metodlari** yordamida matnlarning ma’nosi va kontekstual o‘zgarishlari aniqlanadi. Korpusning sifatini ta’minlash uchun qo‘lda tekshirish va qayta ishlash jarayonlari ham muhim hisoblanadi. Yaratilgan ma’lumotlar bazasi doirasida **qidiruv va indekslash tizimlari** ishlab chiqilib, ma’lumotlarga tezkor kirish imkoniyati yaratiladi. Indekslash jarayoni orqali foydalanuvchilar kerakli ma’lumotlarni oson topish imkoniga ega bo‘ladi.

Diaxron tahlil bosqichida vaqt o‘tishi bilan tilning o‘zgarish tendensiyalari batafsil o‘rganiladi. Ushbu tahlillar til evolyutsiyasini, leksik boylik va sintaktik strukturalarning o‘zgarishini ochib beradi. Korpus yaratishning so‘nggi bosqichida **natijalarni vizualizatsiya qilish orqali** tadqiqotchilarga qulay interfeys taqdim etiladi. Natijada, mustaqillik davri nashrlari asosida yaratilgan diaxron korpus tilshunoslik va NLP tadqiqotlari uchun yangi imkoniyatlar ochib beradi.

1-jadval: Diaxron korpus yaratish bosqichlari

Bosqich	Tavsif	Asosiy vazifalar
1. Materiallarni tanlash	Mustaqillik davri nashrlari namunalarini aniqlash va tanlash	Tematik va madaniy ahamiyatni aniqlash
2. Ma’lumotlarni to‘plash	Tanlangan ma’lumotlarni arxivlar, kutubxonalar va onlayn resurslardan yig‘ish	Ma’lumot bazasini shakllantirish
3. Raqamlashtirish	Matnlarni raqamli formatga o‘tkazish	Matnni elektron ko‘rinishga keltirish

4. Matnni tozalash	Raqamlashtirilgan matnlardagi xatoliklar va noaniqliklarni bartaraf etish	Xatoliklarni tuzatish, formatlashni yaxshilash
5. Normalizatsiya	Matnlarni standartlashtirish, tilshunoslik qoidalariga moslashtirish	Terminologiya va yozuv qoidalarini bir xil holga keltirish
6. Avtomatik teglash	Matn bo‘ylab leksik, morfologik va grammatik teglashni amalga oshirish	Teglash vositalari yordamida avtomatik tahlil
7. Sintaktik va semantik tahlil	Matnning grammatik tuzilishi va ma’nosini aniqlash	Sintaksis va semantika asosida tahlil qilish
8. Qo‘lda tekshirish	Avtomatik tahlil natijalarini qo‘lda nazorat qilish	Aniqlik va ishonchlilikni oshirish
9. Indekslash va qidiruv tizimi	Matnlarni indekslash va tizimlashtirish orqali tezkor qidiruv imkoniyatini yaratish	Foydalanuvchi uchun qulay kirish interfeysini ta’minlash
10. Visualizatsiya va natija tahlili	Natijalarni grafiklar va diagrammalar shaklida vizualizatsiya qilish hamda tahlil qilish	Til evolyutsiyasini osonroq anglash va taqdim etish



Xulosa

Ushbu maqolada mustaqillik davri nashrlari asosida diaxron korpus yaratishning asosiy bosqichlari, ya’ni materiallarni tanlash, ma’lumotlarni to‘plash, raqamlashtirish, matn tozalash, normalizatsiya, avtomatik belgilash, sintaktik-semantik tahlil, qo‘lda tekshirish, indekslash va natijalarni vizualizatsiya qilish bosqichlari batafsil ko‘rib chiqildi. Tadqiqot davomida har bir bosqichning o‘ziga xos vazifalari va qo‘llaniladigan metodologiyalar aniq aks etdi. Korpus yaratish jarayonining har bir bosqichi tilning tarixiy rivojlanishini yanada chuqurroq o‘rganishga xizmat qilgani ta’kidlandi. Shu bilan birga, tadqiqot natijalari asosida korpusning sifatini oshirish va yanada boyitish istiqbollari ko‘rib chiqildi. Kelgusida korpusga audio-video materiallarni qo‘shish imkoniyatlari ham o‘rganilishi, shuningdek, ko‘p modalli ma’lumotlar asosida yangi tahlil usullarini joriy etish tavsiya etiladi. Ushbu yondashuv tilshunoslik va NLP sohalarida yangi ilmiy qarashlar va modellashtirish metodlarini ishlab chiqishga asos bo‘lib xizmat qiladi.

Diaxron modellashtirish bo‘yicha kelajakdagi tadqiqotlar esa til evolyutsiyasining murakkab jarayonlarini yanada aniqroq ko‘rsatish imkonini beradi. Tadqiqot davomida aniqlangan cheklovlar va xatoliklar asosida metodologik takomillashtirishga ehtiyoj sezildi. Natijada, yaratilgan korpus nafaqat tarixiy til tahlilida, balki hozirgi va kelajak til modellarini ishlab chiqishda muhim resurs sifatida namoyon bo‘ldi. Ushbu ilmiy izlanish kelgusidagi tadqiqotlar uchun yangi yo‘nalishlar ochib, til evolyutsiyasini yanada keng qamrovda o‘rganishga zamin yaratadi.

Foydalanilgan adabiyotlar:

1. McEnery T. and Baker P. (2016). Corpus Linguistics and 19th-Century Historical Texts. Edinburgh University Press.
2. Davies M. (2020). The Corpus of Historical American English (COHA): 200 years of historical data for linguistic analysis. Brigham Young University Press.
3. Karimov X. va boshq. (2021). O‘zbek Milliy Korpusi va uning zamonaviy til tadqiqotlaridagi o‘rni. O‘zbek tilshunosligi ilmiy tadqiqotlari jurnali, 8(3), 45–60.
4. Aissen J. (2003). Differential Object Marking: Iconicity vs. Economy. Natural Language and Linguistic Theory, 21(3), 435–483.
5. Ataboyev N.B. Mediamatnlar diaxron korpusida til rivojining empirik tahlil tamoyillari. Monografiya. Buxoro. (2024).
6. Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari. (2023).
7. Meyer C.F. English Corpus Linguistics: An Introduction. Cambridge University Press. (2002).
8. Xusainova Z., Yangibayeva S., Diaxron korpus va uning arxitekturasi. (2025).

O‘ZBEK TILI KORPUSI: TARIXIY RIVOJLANISHI, MAQSAD VA VAZIFALARI

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Primova Mastura Hakim qizi,
O‘qituvchi
primovamastura@navoiy-uni.uz
ToshDO‘TAU

Amirkulov Marufjon Alikulovich,
tayanch doktorant
amirkulov.edu01@gmail.com
ToshDO‘TAU

Annotatsiya. Korpus lingvistikasi zamonaviy tilshunoslikning muhim yo‘nalishlaridan biri bo‘lib, tilning haqiqiy foydalanishini aks ettiruvchi katta hajmdagi matnlar to‘plamini tahlil qilishga asoslanadi. Korpus tilning turli kontekstlarda qo‘llanilishini namoyon etuvchi, maxsus tanlangan va tizimlashtirilgan matnlar bazasiga aytiladi. Bu soha tilni o‘rganishda empirik ma’lumotlarga tayanish imkonini beradi va tilshunoslik nazariyalarini sinovdan o‘tkazishda muhim vosita sifatida xizmat qiladi. O‘zbek tili milliy korpusi esa ushbu umumiy yondashuvning mahalliy ilovasi sifatida o‘zbek tilining o‘ziga xos xususiyatlarini saqlash, hujjatlashtirish va zamonaviy texnologiyalar bilan integratsiya qilishda markaziy o‘rin tutadi. O‘zbek tili korpusi nafaqat tilshunoslik tadqiqotlari uchun muhim resurs, balki sun’iy intellekt (SI) va tabiiy tillarni qayta ishlash (NLP) texnologiyalari uchun ham asosiy ma’lumotlar bazasi hisoblanadi. Ushbu maqolada o‘zbek tili korpusining tarixiy rivojlanishi, uning maqsadlari va vazifalari batafsil tahlil qilinadi.

Kalit so‘zlar: *Til modellari, n-gram til modeli, unigram, modelni baholash, Laplas silliqlash, mashinali o‘qitish.*

Kirish

Korpus lingvistikasi – bu tabiiy tillarning yirik elektron ma’lumotlar bazasini (korpus) yaratish, tahlil qilish va ularni lingvistik tadqiqotlar uchun qo‘llashga qaratilgan fan sohasi. O‘zbek tili korpusi bu sohaning asosiy obyekti bo‘lib, u o‘zbek tilining leksik, grammatik, semantik va pragmatik xususiyatlarini o‘rganish uchun asos bo‘lib xizmat qiladi. **Til korpusi** – strukturallashtirilgan, annotatsiyalangan (teglangan) va statistik jihatdan ishonchli matnlar to‘plami. O‘zbek tili korpusidan turli dialektlar, tarixiy matnlar va zamonaviy matnlar

(matbuot, adabiyot, ilmiy asarlar)ni o‘z ichiga olgan holda tilning to‘liq spektrini aks ettirishda foydalaniladi.

Til korpusi tushunchasi jahon tilshunosligida 1960-yillardan shakllana boshlagan bo‘lib, dastlab ingliz tilida Braun korpusi (Brown Corpus) yaratilgan (1961). Keyinchalik yirik milliy korpuslar, masalan, Britaniya Milliy Korpusi (BNC, ~100 million so‘z) va boshqalar tuzilib, korpus lingvistikasi fan sifatida rivoj topdi. Korpuslar tilshunoslik tadqiqotlarida inqilob yasadi: ilgari olimlar misollarni qo‘lda yig‘ib, ko‘p vaqt sarflagan bo‘lsa, endilikda raqamli korpuslar yordamida soniyalar ichida yuzlab misollarni topish va tahlil qilish imkoni tug‘ildi. Shu tariqa, zamonaviy lingvistikada til korpuslari **empirik tadqiqotlar** uchun asosiy manba sifatida e’tirof etila boshlandi.

O‘zbek tilida korpus lingvistikasi bo‘yicha dastlabki tadqiqotlar 20-asrning ikkinchi yarmida tilshunoslik tadqiqotlari doirasida shakllana boshlagan. Bu davrda O‘zbek tilining leksikasi, grammatikasi va dialektlarini o‘rganishda kichik hajmdagi matnlar to‘plamlari qo‘lda tuzilgan edi. Biroq, zamonaviy ma’noda korpus yaratishga oid loyihalar 21-asr boshlarida, kompyuter texnologiyalarining rivojlanishi bilan birga boshlandi. Dastlabki loyihalar asosan ma’lum bir janrga (masalan, adabiy matnlar yoki rasmiy hujjatlar) yo‘naltirilgan bo‘lib, ularning hajmi va funksionalligi cheklangan edi.

O‘zbek tili korpusining shakllanishi quyidagi bosqichlarni o‘z ichiga oladi:

1. **Raqamlashtirish davri (2000-yillar):** Kompyuter texnologiyalarining kirib kelishi bilan matnlarning elektron formatda saqlanishi. Toshkent davlat universiteti va O‘zbekiston FA tilshunoslik instituti tomonidan dastlabki elektron korpus loyihalari.
2. **Tizimlashtirish va kengaytirish (2010-yillar):** Ushbu bosqichda matnlar to‘plami nafaqat hajman o‘sdi, balki turli janrlar (ilmiy, ommaviy axborot vositalari, og‘zaki nutq transkriptlari) bilan boyitildi. Shu bilan birga, matnlarni annotatsiya qilish (masalan, so‘z turlari bo‘yicha belgilash) bo‘yicha ilk tajribalar amalga oshirildi.
3. **Milliy korpusning shakllanishi (2010-yillardan boshlab):** O‘zbekiston Respublikasi Vazirlar Mahkamasining qarorlari asosida “O‘zbek tili milliy korpusi” loyihasining ishga tushirilishi. O‘zbek tili milliy korpusi loyihasi rasman boshlanib, davlat dasturlari va ilmiy grantlar qo‘llab-quvvatlovi bilan rivojlandi. Bu davrda korpusning hajmi millionlab so‘zlarga yetdi va undan turli ilmiy va texnologik maqsadlarda foydalanish imkoniyati kengaydi.
4. **BigData va AI integratsiyasi (2020-yillar):** Hadoop, Spark kabi BigData platformalarida korpus ma’lumotlarini qayta ishlash. GPT-3/4, BERT kabi neyron tarmoqlarning o‘zbek tili uchun moslashtirilishi.

Hozirgi kunda O‘zbek tili milliy korpusi turli manbalardan (kitoblar, gazetalar, veb-saytlar, ijtimoiy tarmoqlar) olingan matnlarning katta to‘plamini o‘z ichiga oladi. Korpusning umumiy hajmi millionlab so‘zlardan iborat bo‘lib, u turli davrlar (tarixiy matnlar, sovet davri, mustaqillik davri) va janrlarni (ilmiy, adabiy, publitsistik) qamrab oladi. Shu bilan birga, korpus foydalanuvchilar uchun qulay interfeys va qidiruv tizimi bilan ta’minlangan bo‘lib, bu uni tadqiqotchilar va texnologlar uchun qulay vositaga aylantiradi.
O‘zbek tili korpusining asosiy maqsadlari quyidagilardan iborat:

1) *O‘zbek tilining tarixiy va zamonaviy variantlarini arxivlash.*
2) *Yo‘qolib borayotgan dialektlarni (Qashqadaryo, Surxondaryo) hujjatlashtirish.*
3) *Lingvistik qonuniyatlarni statistik usullar bilan tekshirish (masalan, so‘z chastotasi, kolokatsiyalar).*
4) *Adabiyotshunoslik, tarix va sotsiolingvistika sohalarida qo‘llash.*
5) *Mashina tarjimasi, ovozli assistentlar, avtomatik tekshirish tizimlari uchun ma’lumotlar bazasi.*
6) *NLP (Tabiiy tilni qayta Ishlash) algoritmlarini o‘zbek tiliga moslashtirish.*
7) *Sun’iy intellekt va NLP texnologiyalarini rivojlantirish.*
8) *O‘quv qo‘llanmalari, lug‘atlar va onlayn platformalar yaratishda korpusdan foydalanish.*

Adabiyotlar sharhi

Tony McEnery va Andrew Hardie tomonidan yozilgan “Corpus Linguistics: Method, Theory and Practice”[1] kitobi korpus lingvistikasi sohasidagi metodologiya, nazariy asoslar va amaliy qo‘llanilishni har tomonlama yoritgan muhim ilmiy manba hisoblanadi. Mualliflar korpusga asoslangan tadqiqotlarning statistik va kvalitativ usullarini batafsil ko‘rib chiqib, lingvistik nazariyalarni empirik ma’lumotlar bilan sinab ko‘rishning ahamiyatini ta’kidlaydi (masalan, kolokatsiyalar, so‘z chastotasi). Ular monolingval, multilingval va ixtisoslashgan korpuslar kabi turli korpus turlarini, shuningdek, konkordans, kollokatsiya va kalit so‘z tahlili kabi analitik usullarni batafsil o‘rganib, tadqiqotchilarga o‘z ishlarini loyihalash va amalga oshirishda yo‘l-yo‘riq beradi. Bundan tashqari, korpus lingvistikasidagi cheklovlar, xususan, reprezentativlik va umumlashtirish muammolari tanqidiy baholanib, soha bo‘yicha chuqur va o‘ylangan yondashuvni rag‘batlantiradi. Umuman olganda, ushbu kitob yangi boshlayotgan va tajribali tadqiqotchilar uchun metodologik qat’iylik va refleksivlik bilan ilmiy tadqiqot olib borishda qimmatli qo‘llanma vazifasini o‘taydi.

O‘zbek tilshunosligida ham til korpusi g‘oyasi nisbatan kechroq shakllandi. Ilk marotaba 2010-yillar atrofida mahalliy olimlar korpus lingvistikasi masalalariga

e’tibor qaratib, ushbu yo‘nalishda tadqiqotlarni boshladilar. Jumladan, O‘zbekiston Milliy Universiteti professori N.Z. Abdurahmonova o‘zbek tili elektron korpusini yaratish g‘oyasini ilgari surib [2], uning konseptual va ilmiy asoslarini ishlab chiqdi. Abdurahmonova rahbarligidagi ilmiy guruh o‘zbek tili uchun ilk elektron korpus loyihasini boshladi va bu boradagi natijalar keyinchalik 2021-yilda himoya qilingan doktorlik dissertatsiyasida yoritildi. Shu bilan birga, ToshDO‘TAU olimasi Sh. Hamroyeva mustaqil tadqiqotchi sifatida o‘zbek tilida **mualliflik korpusi** yaratish masalasini o‘rganib, bu bo‘yicha ilk ilmiy izlanishlarni amalga oshirdi. Hamroyevaning tadqiqoti o‘zbek tilida mualliflik korpusini tuzish mezonlarini belgilab beruvchi muhim manba bo‘lib, milliy korpus g‘oyasining shakllanishida dastlabki qadamlardan biri sifatida e’tirof etildi.

Ilk ilmiy qarashlarda Abdurahmonova va Hamroyeva o‘zbek milliy korpusini yaratishning dolzarbligini asoslab berdilar. Ular o‘zbek tilini raqamlashtirish, katta hajmdagi matnlarni jamlab, zamonaviy til holatini aks ettiruvchi elektron baza yaratish nafaqat lingvistik tadqiqotlar, balki tilning taraqqiyoti va xalqaro maydonda tutgan o‘rnini mustahkamlash uchun zarur, degan xulosalarga kelishgan. Ayniqsa, Sh. Hamroyeva 2018-yilda chop etgan maqolasida milliy korpusni *“til ma’naviy merosni saqlash va boyitishning asosiy vositasi”* deb ta’riflab, **o‘zbek tilini jahon tillari qatoriga olib chiqishning yo‘li – uni raqamlashtirish va milliy korpusini yaratish** ekanini alohida ta’kidlagan. Shunday qilib, o‘tgan asr oxiri va XXI asr boshlariga kelib o‘zbek tilida milliy korpus yaratish g‘oyasi ilmiy adabiyotlarda shakllandi va dastlabki amaliy qadamlar tashlandi.

O‘zbek tilining Milliy Korpusini amalda shakllantirish va uni dasturiy jihatdan hayotga tatbiq etishda Toshkent davlat o‘zbek tili va adabiyoti universiteti olimlari, xususan, B. Elov va uning hamkorlari katta hissa qo‘shdilar. Abdurahmonova boshchiligidagi loyiha natijasi o‘laroq 2020-yilda internetda “O‘zbek tilining milliy korpusi” (https://uzschoolcorpara.uz/) ishga tushirildi. Mazkur korpus dasturiy jihatdan to‘liq funksionallikka ega bo‘lib, maxsus korpus menejeri yordamida matnlarni saqlash, izlash va tahlil qilish imkoniyatlarini beradi. Korpus kontenti juda boy: **u o‘zbek tilidagi lug‘atlar, internet sahifalari, o‘quv qo‘llanmalar, ilmiy, rasmiy va badiiy matnlar majmuasini** o‘z ichiga oladi[3]. Matnlar zamonaviy o‘zbek tilining xilma-xil uslub va janrlarini qamrab olishi uchun keng ko‘lamda to‘plangan. Natijada elektron korpus o‘zbek tilining hozirgi holatini aks ettiruvchi katta hajmdagi yozma ma’lumotlar bazasini yaratishga muvaffaq bo‘lindi.

B. Elov va jamoasi korpusning texnik arxitekturasini ishlab chiqishda zamonaviy korpus menejment tizimlariga tayandilar. Korpusning veb-interfeysi orqali foydalanuvchilar turli usullarda qidiruv o‘tkazishi mumkin: xususan, **so‘zning asl ko‘rinishi (token) bo‘yicha, so‘zning lug‘aviy shakli (lemma)** bo‘yicha, **so‘z birikmalar va konkordans bo‘yicha qidirish** imkoniyati mavjud.

Bunday qidiruv turlari korpusdan foydalanuvchilarga kerakli til birligini tez va aniq topish hamda keng kontekstda ko‘rish imkonini beradi. Korpus tizimi foydalanuvchiga natijalarni qulay shaklda taqdim etadi: bunda kontekstda topilgan so‘zlar ajratib ko‘rsatiladi, zarur hollarda statistik ma’lumotlar (so‘z chastotasi, kombinatsiyalanuvchi so‘zlar va h.k.) ham birgalikda ko‘rsatiladi. Korpus platformasi tarkibida maxsus **morfologik tahlilchi** va **stemming dasturlari ham integratsiya qilingan** bo‘lib, foydalanuvchi xohlagan so‘zning tuzilishini, uning boshi va qo‘shimchalarini ajratib tahlil qilishi mumkin. Bu esa o‘zbek tilidek boy morfologiyali til korpusida samarali navigatsiya qilish va chuqurroq lingvistik tahlil o‘tkazishga xizmat qiladi.

Milliy korpusda lingvistik annoatsiya va teglash ishlari bosqichma-bosqich amalga oshirilmoqda. Ma’lumki, korpusning qimmati uning ichidagi matnlarning ma’lum darajada teglangan, ya’ni lingvistik belgi qo‘yilgan bo‘lishida namoyon bo‘ladi. B. Elov va Sh. Hamroyeva ilmiy hamkorlikda o‘zbek tilining nutq qismlarini belgilash (PoS tagging) uchun maxsus teglar tizimini ishlab chiqdilar [4]. Ular 2022-yilgi maqolalarida o‘zbek tilidagi so‘z turkumlari uchun batafsil belgi to‘plamini (tegsetni) taklif qilib, **korpus matnlarini avtomatik teglash** masalasini ko‘tardilar [5]. Bu ish o‘zbek korpus lingvistikasi uchun muhim ahamiyatga ega, chunki agglutinativ xususiyatga ega o‘zbek tilida so‘zlarni morfologik jihatdan tahlil qilgan holda to‘g‘ri teglash murakkab vazifa hisoblanadi. Hozirgi kunda korpus matnlarining asosiy qismi **morfologik teglashdan o‘tkazilgan** va foydalanuvchi so‘zlarni **so‘z turkumi, grammatik kategoriyalari** bo‘yicha filtrlab qidirishi mumkin. Korpus jamoasi bundan tashqari **leksik-statistik ko‘rsatkichlarni** ham hisoblash ustida ishlamoqda. Misol uchun, TF-IDF kabi statistik usullar o‘zbek korpusi matnlarida muayyan so‘zning nisbiy ahamiyatini aniqlash uchun qo‘llanilgan bo‘lib, bunday tadqiqotlar natijalari ham e’lon qilindi[6]. Korpusda so‘zlarning chastotasi, eng ko‘p qo‘llanuvchi birliklar ro‘yxati, kollokatsiyalar (qo‘shma yasalmalarning chiqish tezligi) kabi ko‘rsatkichlar avtomatik tarzda olinadi va ilmiy izlanishlar uchun tayyor material sifatida xizmat qiladi.

O‘zbek milliy korpusining hajmi va qamrovi doimiy kengayib bormoqda. Dastlabki versiyada korpus hajmi bir necha million so‘zni tashkil etgan bo‘lsa, qisqa fursatda matnlar bazasi sezilarli ravishda oshirildi (hozirgi kunda o‘n millionlab so‘zlar). Matnlar **uslubiy jihatdan turli bo‘lgani** uchun korpus tarkibini statistik tahlil qilish qiziqarli natijalar bermoqda. Masalan, ilmiy uslubdagi matnlarda eng yuqori chastotali so‘zlar rasmiy va badiiy uslub matnlaridagidan farq qilishi kuzatilgan (bu boradagi aniq ma’lumotlar korpus asosida tuzilgan tezislarda keltirilmoqda). Bundan tashqari, korpusdagi matnlardan foydalanib **N-grammalar, so‘zlarning o‘rtacha uzunligi, gap qurilishidagi o‘rtacha elementlar soni** kabi statistik ko‘rsatkichlar ham avtomatik chiqarilishi yo‘lga qo‘yilgan. B. Elov va hamkorlarining izlanishlari natijasida kelgusida **korpusni intellektual tahlil qilish**

(masalan, matnlardan mavzuga oid axborotni avtomatik ajratish, ma’lumotlarni klasterlash) kabi funksiyalarni joriy etish rejalashtirilmoqda.

Qarshiyev, Tursunov va Maxmidov (2022)ning ilmiy tadqiqoritda o‘zbek tili milliy korpusini loyihalashning strukturaviy tamoyillari, lingvistik va texnologik jihatlari ko‘rib chiqilgan [5]. Mualliflar korpusni multidisiplinar yondashuv asosida shakllantirish zarurligini ta’kidlab, uning tarkibiga dialektal matnlar, tarixiy manbalar va zamonaviy matbuot materiallarini kiritishning ahamiyatini alohida urg‘ulaydi. Tadqiqotda o‘zbek tilining morfologik murakkabligi (masalan, so‘z yasalishi, izofa konstruksiyalari) va annotatsiya jarayonidagi qiyinchiliklar (tokenizatsiya, lemmatizatsiya) algoritmik yechimlar bilan tavsiflangan.

M. B. Kasimova (2024) [7] o‘zbek tilining milliy korpusida darajanish hodisasini ilmiy tahlil qilib, ushbu jarayonning til ichidagi ifodalanishini o‘rganadi. Muallif o‘zbek tilidagi nisbiy va orttirma darajalar (masalan, “ko‘proq”, “eng yaxshi”), shuningdek, so‘z birikmalari orqali intensivlikni ifodalash vositalarini korpus ma’lumotlari asosida tizimli tahlil qilgan. Tadqiqotda korpus annotatsiyasining lingvistik aniqligi (POS teglar, sintaktik bog‘lanishlar) va ma’lumotlarning reprezentativligi urg‘ulanib, darajalanishning dialektal va adabiy variantlari solishtirilgan. Bu ish o‘zbek tilining dinamik xususiyatlarini korpus lingvistikasi nuqtai nazaridan tushunishga muhim hissa qo‘shadi, shu bilan birga NLP tizimlarini milliy tilga moslashtirish uchun amaliy asos yaratadi.

Xudayberganov, N. (2024) [8] tomonidan yozilgan “O‘zbek tili korpusiga morfologik ishlov berish” maqolasi o‘zbek tili milliy korpusida morfologik tahlilning dolzarb masalalarini ilmiy jihatdan o‘rganadi. Muallif o‘zbek tilining murakkab morfologik tuzilishiga mos keladigan usullar va algoritmlarni tahlil qilib, so‘z shakllarining xilma-xilligi hamda affikslar tizimini aniqlashdagi muammolarga e’tibor qaratadi. Tadqiqotda korpus lingvistikasi va kompyuter lingvistikasi sohasidagi joriy qiyinchiliklar yoritilib, morfologik ishlov berishning til resurslari sifatini oshirishdagi ahamiyati ta’kidlanadi. Ushbu ish o‘zbek tilini qayta ishlashda yangi imkoniyatlar ochishi bilan birga, kelajakdagi lingvistik tadqiqotlar uchun muhim poydevor vazifasini o‘taydi.

Umuman, B. Elov boshchiligidagi jamoa tomonidan yaratilgan o‘zbek tili milliy korpusi *texnik jihatdan zamonaviy, lingvistik jihatdan boy annotatsiyalangan va statistik jihatdan foydali ma’lumotlar manbai* bo‘lib shakllandi [9]. Bu korpus O‘zbekistonda korpus lingvistikasi rivojini amalda yangi bosqichga olib chiqdi va endilikda turli ilmiy-amaliy maqsadlar uchun tayanch platforma bo‘lib xizmat qilmoqda.

Korpus tuzish jarayonidagi asosiy bosqichlar va yondashuvlar

Har qanday til korpusini yaratish bir necha muhim **bosqichlar** asosida amalga oshiriladi. O‘zbek tili korpusini tuzish tajribasi ham shuni ko‘rsatadiki, quyidagi bosqichlar izchil bajarilganda sifatli korpus shakllantirish mumkin:

1. **Maqsad va qamrovni belgilash:** Dastlab korpusning maqsadi va chegaralari aniqlanadi. Ya’ni, korpus *qanday til materialini* o‘z ichiga oladi, uning *janrlar va uslublar bo‘yicha tarkibi* qanday bo‘lishi lozim – shu kabi mezonlar belgilanadi. O‘zbek milliy korpusini tuzishda ham avvaliga zamonaviy adabiy tilni maksimal darajada qamrab oluvchi korpus yaratish maqsad qilindi. Bunda *badiiy adabiyot, ommaviy axborot vositalari, ilmiy adabiyot, rasmiy uslub* kabi barcha asosiy sohalardan matnlar jalb etilishi rejalashtirildi.
2. **Matnlar to‘plash va raqamlashtirish:** Korpus uchun zarur materiallarni yig‘ish bosqichi ko`p mehnat talab qilinadi. O‘zbek tili uchun matnlar turli manbalardan olindi: *elektron hujjatlar, internet saytlari, elektron kitoblar, gazetalar arxivi, raqamlashtirilgan adabiyotlar* va hokazo. Ba’zi kerakli manbalar faqat qog‘ozda mavjud bo‘lsa, ularni **skanerlash va OCR yordamida raqamli ko‘rinishga keltirish** ishlari qilindi. Shuningdek, o‘zbek tilida **lotin va kirill** yozuvlaridan ikkalasi ham qo‘llangani bois, matnlarni bir xil formatga keltirish masalasi hal etildi: kirillda bo‘lgan matnlar lotinga o‘girildi yoki aksincha, foydalanuvchi uchun qidiruvda qulay bo‘lishi uchun dasturiy yechimlar joriy qilindi. Bu bosqichda matnlarning mualliflik huquqi masalalariga ham e’tibor qaratilib, ochiq manbalar yoki ruxsat etilgan materiallar ishlatildi.
3. **Matnlarni tozalash va teglash:** Toplangan ma’lumotlar bazasi shakllantirilgach, ularni korpusga kiritishdan avval **tozalash (cleaning)** lozim bo‘ladi. Ya’ni, matnlardagi ortiqcha texnik axborot, reklama, kod parchalari kabi til uchun ahamiyatsiz qismlar chiqarib tashlanadi. O‘zbek korpusini tuzishda ham web-sahifalardan olingan matnlar filtrlanib, faqat lingvistik jihatdan **sof matn** qoldirildi. Keyin esa matnlarni korpusga yuklash jarayonida tokenlash, ya’ni matnni alohida so‘zlarga ajratish amalga oshirildi. Keyingi muhim qadam – annotatsiya (teg qo‘yish): bu yerda har bir so‘zga uning lug‘aviy ma’nosi (lemma) va grammatik xususiyatlarini belgilovchi tegar teglar biriktirildi. Annotatsiya darajasiga qarab korpus **teglanmagan korpus** yoki **teglangan korpus** deb ataladi. O‘zbek milliy korpusining ilk versiyasi asosan tokenizatsiya va lemmatizatsiya bilan chegaralangan bo‘lsa, hozirgi kunda morfologik teglash ham qo‘shilmoqda. Ushbu bosqichlarda lingvist-mutaxassislar va dasturchilar birgalikda ishlashi talab etiladi, chunki masalan, **yopiq qismlar** (“va”, “bu” kabi)ni to‘g‘ri aniqlash yoki birikmalarni yaxlit birlik sifatida belgilash kabi masalalar tilshunoslik bilimini ham, dasturlashni ham talab etadi.

4. **Ma’lumotlar bazasini shakllantirish:** Tozalangan va teglangan matnlar korpusning ichki bazasiga yuklanadi. Bunda maxsus korpus platformasi ishlab chiqildi. O‘zbek milliy korpusi uchun maxsus veb-platforma yaratildi va matnlar **janr, sana, muallif, manba kabi metama’lumotlar bilan birga bazaga kiritildi**. Bu kelgusida foydalanuvchilarga, masalan, faqat 2000-yillardan keyingi gazetalar tilini qidirish, yoki faqat badiiy adabiyotdan misollar topish kabi filtrlash imkonini beradi.

5. **Test sinov va tuzatishlar:** Korpus dastlabki shakliga keltirilgach, uni sinovdan o‘tkazish muhim. Bu bosqichda korpusdan turli namuna so‘rovlar qilinib, tizimning to‘g‘ri ishlashi, qidiruv natijalarining relevanti tekshiriladi. Test jarayonida aniqlangan xatolar – masalan, noto‘g‘ri tokenlarga bo‘linib ketgan so‘zlar, xato teglangan shakllar – tuzatiladi. O‘zbek korpusi yaratishda dasturchi va filologlardan iborat ishchi guruh bir necha bor sinov rejimini o‘tkazib, foydalanuvchi nuqtayi nazaridan noqulayliklar bo‘lsa, ularni bartaraf etdi.

6. **Yangilash va boyitib borish:** Korpus yaratish bir martalik jarayon emas, **uzluksiz jarayondir**. Tayyor korpus muntazam yangilanib va kengayib boradi. Masalan, har yili yangi gazeta maqolalari, yangi kitob matnlari qo‘shilishi mumkin, yoki korpusga ilgari kiritilmagan janrlar qo‘shilishi rejalashtiriladi (masalan, internet forumlaridagi yozishmalar, suhbat chatlaridan olingan matnlar va hokazo). O‘zbek tilining milliy korpusini rivojlantirish strategiyasida ham har yili uning hajmini oshirish, yangi **subkorpuslar** qo‘shish (masalan, dialektal nutq korpusi, she’riy matnlar korpusi va hokazo) ko‘zda tutilgan. Shu bilan birga, korpus foydalanuvchilaridan kelib tushayotgan fikr-mulohazalar asosida tizim funktsiyalari takomillashtirilib boriladi.

Korpus tuzish jarayonida ikki xil yondashuv mavjud: **statik korpus** va **dinamik (monitor) korpus** yondashuvlari. Statik korpus belgilangan davr va manbalardan tuzilib, tugallangan ma’lumotlar to‘plamini anglatadi (masalan, ma’lum yillargacha bo‘lgan adabiyotlar bazasi). Dinamik korpus esa doimiy to‘ldirilib boriladi, tilning o‘zgarishi va yangi birliklar paydo bo‘lishini doimiy kuzatish imkonini beradi. O‘zbek milliy korpusi dastlab statik modelda ishga tushgan bo‘lsa-da, hozirda u dinamik tarzda boyitilmoqda. Zero, til uzluksiz rivojlanishda davom etar ekan, korpus ham ushbu jarayonni monitor sifatida kuzatib borishi va aks ettirishi lozim. Xususan, yildan-yilga tilga kirib kelayotgan yangi so‘zlar (neologizmlar) yoki yangi matn janrlari (masalan, internet memlari tili) ham korpusga qo‘shib borilishi bilan uning ilmiy qimmati oshadi.

O‘zbek tilidagi korpusning asosiy maqsadi va vazifalari

O‘zbek tili milliy korpusi bir qator muhim maqsad va vazifalarni amalga oshirish uchun yaratilgan integral platformadir. Uning asosiy maqsadlari quyidagicha izohlanadi:

1. **Tilshunoslik tadqiqotlari:** Korpus, eng avvalo, nazariy va amaliy tilshunoslikda empirik bazani ta’minlashni maqsad qilgan. Ya’ni, o‘zbek tilining lug‘aviy boyligi, grammatik qurilishi, uslubiy xususiyatlari bo‘yicha **dalillarga tayangan tadqiqotlar** qilish uchun korpus asosiy manba bo‘ladi. Misol uchun, olimlar ma’lum bir so‘zning yangi ma’no kasb etgan-etmaganini aniqlash uchun korpusdan foydalanib, turli davrlardagi qo‘llan malarni solishtirishi mumkin. Shu ma’noda korpus til tizimidagi o‘zgarishlarni va bugungi holatni aniqlash vazifasini bajaradi. Korpus yordamida tilshunoslar uchun ilgari imkonsiz bo‘lib kelgan ko‘plab masalalarni (masalan, so‘z birikmalarning statistik xususiyatlarini) o‘rganish imkoni tug‘ildi. Korpusda to‘plangan ma’lumot til jarayonlarining real manzarasini taqdim etadi, bu esa har qanday lingvistik nazariyani ishonchli asoslarga tayantirish imkonini beradi.

2. **Kompyuter lingvistikasi va NLP (Tabiiy tilda ishlov berish):** Bugungi kunda NLP sohasida katta hajmdagi til ma’lumotlarisiz muvaffaqiyatga erishib bo‘lmaydi. O‘zbek tili korpusi shu sohada ham asosiy infratuzilma vazifasini o‘taydi. Misol uchun, mashina tarjimasi tizimlarini yaratishda parallel korpuslar talab etiladi – o‘zbek milliy korpusida maxsus parallel korpus bo‘limi ham yaratilmoqda, unda o‘zbek va xorijiy tillardagi tarjima juftliklari jamlanmoqda. Shuningdek, **matnlarni avtomatik tasniflash, so‘z tanib olish (speech recognition), matnni tahlil qilish (text analytics)** kabi ko‘plab NLP masalalarida korpusdan olinadigan til modellari muhim ahamiyatga ega. Milliy korpus negizida o‘zbek tilining n-gramma modellari, statistik til modellari ishlab chiqilmoqda. Bu modellardan kompyuter lingvistikasi amaliyotida – ovozli yordamchilar, intellektual qidiruv tizimlari, matnni avtomatik tuzatish vositalari kabi ko‘plab yo‘nalishlarda foydalanish mumkin. Korpusning yana bir vazifasi – **agglutinativ tillarga xos NLP muammolarini hal etish** uchun ilmiy maydon yaratishdir: masalan, o‘zbek tilida so‘zlarning turli shakllarini tanib olish (lemmatizatsiya) va ularni bir ma’noli formaga keltirish (disambiguation) bo‘yicha tadqiqotlar aynan korpus orqali sinovdan o‘tkaziladi. Shunday qilib, o‘zbek tili korpusi nafaqat lingvistika, balki sun’iy intellekt va dasturiy yechimlar sohasida ham asosiy resurs rolini o‘ynaydi.

3. **Raqamli lug‘atchilik va leksikografiya:** Korpusning muhim maqsadlaridan yana biri – zamonaviy **elektron lug‘atlar** uchun poydevor yaratish. An’anaviy lug‘atlar odatda tuzuvchilarning tajribasi va cheklangan materiallarga tayanib tuzilardi. Korpus esa lug‘at tuzishda boy va ishonchli material manbasi bo‘lib xizmat qiladi. Masalan, hozirgi kunda tayyorlanayotgan yangi avlod o‘zbek tilining izohli lug‘ati korpusdan olingan misollar bilan boyitilmoqda. Korpusdan **so‘zning eng tipik qo‘llanish misollari, ko‘p ma’noli so‘zlarning har bir ma’nosiga mos kontekstlar, so‘zning kollokatsion (qo‘shma) bog‘lanishlari** kabi ma’lumotlar avtomatik tarzda olinib, lug‘at moddalariga kiritilishi mumkin. Bu esa lug‘atlarni ancha hayotiy va foydali qiladi, foydalanuvchi lug‘atdan so‘zning

haqiqiy ishlatilishi haqida tasavvur oladi. Shuningdek, korpus tez-tez yangilanib boruvchi resurs bo‘lgani uchun, lug‘atlardagi ma’lumotlarni ham osonlik bilan yangilash imkoni mavjud – masalan, oxirgi o‘n yilda paydo bo‘lgan yangi so‘zlar korpusga qo‘shilgan bo‘lsa, lug‘atning navbatdagi nashrida ular aks ettiriladi. **Terminologik lug‘atlar** tuzishda ham korpusdan foydalanilmoqda: muayyan soha matnlari filter orqali ajratib olinib, o‘sha sohaga xos atamalar korpus ichida aniqlanadi va lug‘atga kiritiladi. Demak, korpusning lug‘atchilikdagi vazifasi – **milliy leksikografik bazani boyitish va dalillash**.

4. **Ta’lim va til o‘rgatish:** O‘zbek tili korpusining yana bir ustuvor maqsadi – til ta’limi sohasiga xizmat qilishdir. Korpus materiallari ona tilini yoki chet til sifatida o‘zbek tilini o‘rgatishda innovatsion resurs bo‘la oladi. Xususan, 2021 yilda ishga tushirilgan **“O‘zbek tilining ta’limiy korpusi” (uzschoolcorpara.uz)** ushbu maqsadga yo‘naltirilgan maxsus subkorpusdir. Ta’limiy korpus maktab darsliklari, o‘quv qo‘llanmalar va o‘quv lug‘atlaridan tuzilgan bo‘lib, o‘quvchilar hamda o‘qituvchilar uchun moslashtirilgan. Uning yordamida o‘qituvchilar darsda ishlatish uchun tezda misollar topishi, mashqlar tuzishi mumkin – korpus bir necha soniyada kerakli mavzu bo‘yicha yuzlab real misollarni taqdim eta oladi. Misol uchun, o‘quvchi *“sifat”* mavzusini o‘tayotganda, korpusdan *“yaxshi”, “chiroyli”* kabi sifatlarning haqiqiy matnlardagi qo‘llanishini topib, ular ishtirok etgan gaplarni ko‘rishi mumkin. Bu usul kontekst orqali o‘rgatish tamoyiliga juda mos tushadi. Tilni chet tili sifatida o‘rganuvchilar ham korpusdan foydalansa, ular sun’iy misollardangina emas, balki jonli til materialidan ta’lim oladilar. Korpus asosidagi topshiriqlar, masalan, biror so‘zning kollokatsiyalarini topish, ma’lum grammatik strukturaning qancha va qanday variantlarda qo‘llanishini aniqlash kabi mashqlar talabalar uchun qiziqarli va foydali bo‘ladi. Shuningdek, korpus lingvopedagogik tadqiqotlar uchun ham zamin yaratadi – masalan, xorijiy auditoriya uchun o‘zbek tilidagi eng qiyin so‘zlar yoki tuzilmalarni korpus yordamida statistik tahlil qilish va shu asosda o‘quv materiallarini takomillashtirish mumkin. Demak, milliy korpusning ta’limiy vazifasi – **til o‘qitishda innovatsion elektron vosita** bo‘lish, o‘quv jarayonini boyitish va yengillashtirishdan iborat.

Yuqorida sanab o‘tilgan maqsadlar shuni ko‘rsatadiki, o‘zbek tilining milliy korpusi ko‘p qirrali loyiha bo‘lib, u **ilm-fan, ta’lim, texnologiya va madaniyat sohalarini** bog‘lovchi infratuzilma vazifasini o‘tamoqda. Bir tarafdan, u sof akademik tadqiqotlar uchun qo‘llanilsa, ikkinchi tarafdan jamiyatdagi turli foydalanuvchilar – jurnalistlar, tarjimonlar, yozuvchilar, talabalar – uchun ham foydali axborot manbaidir. Shuning uchun korpusni yaratishda uning ko‘p maqsadli va ko‘p foydalanuvchili tizim ekani inobatga olinadi.

Korpusning asosiy funksiyalari va foydalanuvchilari

Til korpusining funksiyalari deganda, undan foydalangan holda bajarilishi mumkin bo‘lgan amaliy ishlar tushuniladi. O‘zbek tili milliy korpusi quyidagi asosiy funksional imkoniyatlarni taklif qiladi:

1. **Konkordans qidiruvi:** Korpusning eng asosiy funksiyasi – kerakli so‘z yoki iborani barcha kontekstlari (atrofidagi so‘zlari bilan)da topib berishdir. Bu **konkordans qidiruv** deb ataladi. Masalan, foydalanuvchi korpusga *“tarix”* so‘zini kiritib qidirsa, korpus bir necha soniya ichida ushbu so‘z uchragan barcha jumlalarni ro‘yxatlab beradi, bunda *“tarix”* so‘zi har bir misolda ajratib ko‘rsatiladi. Bu usul tadqiqotchiga so‘zning qo‘llanish xilma-xilligini ko‘rish, ma’nosini kontekstda tushunish, hatto **frazeologik birikmalarni aniqlash** imkonini beradi. Konkordans funksiyasi orqali biror adib tilini o‘rganayotgan adabiyotshunos o‘sha adib eng ko‘p ishlatgan so‘zlar ro‘yxatini tuzishi yoki til xususiyatlarini tahlil qilishi mumkin. Shuningdek, til o‘rganuvchilar uchun ham konkordans – boy misollar xazinasi.

2. **So‘zlar statistikasini chiqarish:** Korpusdan nafaqat alohida misollar, balki **umumiy statistik ma’lumotlarni olish** ham mumkin. Masalan, *“mustaqillik”* so‘zi korpusda necha marta uchragan? U qaysi janrda ko‘proq qo‘llanadi – gazeta uslubidami yoki badiiy adabiyotdami? Korpus bunday savollarga javob berish uchun tegishli funksiyani o‘z ichiga oladi. Maxsus hisoblash rejimi orqali berilgan so‘z yoki ibora korpusda *uchrash tezligi (per million ko‘rinishida),* u bilan *eng ko‘p birga keluvchi so‘zlar (kollokatsiyalar)*, hatto vaqt o‘tishi bilan *dinamika (yillarga ko‘ra chastota o‘zgarishi)* kabi statistik ma’lumotlar olinadi. Bu funksiyalar tilshunoslar uchun ayniqsa qimmatlidir, chunki so‘zning trendsini, taraqqiyotini ko‘rish, lug‘at boyligi miqyosini baholash imkonini beradi.

3. **Filtrlash va tanlangan korpuslar (subkorpuslar):** O‘zbek milliy korpusida foydalanuvchilar uchun turli **subkorpuslar** ham ajratilgan. Misol tariqasida, faqat badiiy adabiyot matnlarini yoki faqat ommaviy axborot vositalari matnlarini o‘z ichiga olgan alohida tanlamalarni ko‘rish mumkin. Bu orqali tadqiqotchi o‘ziga kerak yo‘nalishdagi matnlarnigina tahlil qilishi osonlashadi. Misol uchun, lingvist olim rasmiy uslubdagi hujjatlarda yangi paydo bo‘lgan so‘zlarni o‘rganmoqchi bo‘lsa, korpusdan rasmiy hujjatlar subkorpusini tanlab, u yerdan qidiruv va tahlillarni amalga oshiradi. Yana bir muhim filtr – **xronologik filtr**. Bu yordamida ma’lum bir davr (masalan, 1990–2000 yillar) matnlarini ajratib olish va ularda tahlil o‘tkazish mumkin. Natijada, korpus **sinxron (bir davr ichida)** va **diaxron (turli davrlar kesimida)** tadqiqotlar uchun moslashadi. Foydalanuvchilar korpus interfeysida oddiy menyu orqali bunday filtrlarni tanlay oladilar, bu esa murakkab so‘rovlarni sodda usulda shakllantirish imkonini beradi.

4. **Lug‘aviy va grammatik axborot chiqarish:** Korpus tarkibidagi matnlar lingvistik jihatdan teglangani bois, undan grammatik ma’lumotlarni ham izlab topish imkoni bor. Masalan, foydalanuvchi korpusdan faqat fe’l turkumiga oid so‘zlarni

yoki faqat ko‘plik shakldagi otlarni qidirishi mumkin. Bu jarayonda korpusning ichki morphologik teglar bazasi ishlaydi. O‘zbek tilining milliy korpusida, masalan, faqat ko‘plikdagi otlarning ro‘yxatini chiqarish orqali ko‘plik shakllantirish qoidalarining amaliy ko‘rinishini kuzatish mumkin – eng ko‘p *“-lar”* qo‘shimchasi bilan kelgan otlar yoki noodatiy ko‘plik shakllarini topish kabi. Shuningdek, korpusdan lingvistik qurilishlarni izlash imkoniyati ham qo‘shilgan: masalan, *“A dan ko‘ra B”* konstruksiyasini qidirib, uning turli variatsiyalari (masalan, “undan ko‘ra bunday…”) korpus bo‘yicha qanday namunalarda uchrashini ko‘rish mumkin. Bu funksiyalar tilshunoslik izlanishlari uchun ayni muddaodir – ayniqsa, o‘zbek tilidagi **bog‘lanish vositalari, frazeologizmlar** kabi murakkab hodisalarni o‘rganishda korpus noyob imkoniyatlar yaratadi.

Yuqoridagi funksiyalardan turli foydalanuvchilar manfaatdor bo‘lishi mumkin. **Akademik tilshunoslar** korpusdan o‘z izlanishlari uchun asosiy material sifatida foydalanadilar – ular uchun korpus faktlar xazinasi. **Leksikograflar (lug‘at tuzuvchilar)** korpusdan so‘z ma’nolarini tasdiqlovchi misollar, yangi so‘zlarni topish, sinonim va antonimlarni kontekstda aniqlash uchun foydalanadilar. **Adabiyotshunoslar va matnshunoslar** muayyan adib yoki davr tilining xususiyatlarini o‘rganish uchun korpusga murojaat etadilar – masalan, bir yozuvchining uslubini boshqa yozuvchilar uslubi bilan solishtirish kabi. **Tarjimonlar** korpusdan murakkab iboralarning tarjima muqobilini topish yoki ikki til korpusidagi parallellarni ko‘rish orqali o‘z ishlarida yordam oladilar. **O‘qituvchilar va talabalar** korpusdan ilmiy ishlar yozishda, til o‘rganishda interaktiv vosita sifatida foydalanmoqdalar – masalan, talabalar insho yozishda biror atamaning to‘g‘ri qo‘llanishiga ishonch hosil qilish uchun korpusga murojaat qilishi mumkin. **Jurnalistlar va blogerlar** ham korpusdan foyda olishi mumkin: ma’lum bir so‘z ommaviy axborot vositalarida qanday ishlatilayotganini ko‘rib, uslubiy to‘g‘ri qo‘llashga oid xulosa chiqaradilar. Hatto, huquq sohasida, xorij tajribasida bo‘lgani kabi, **til ekspertizasi** uchun ham korpusdan foydalanish imkoniyatlari mavjud – masalan, sud lingvistik tahlillarida matnlarni qiyosiy o‘rganish uchun.

Shunday qilib, o‘zbek milliy korpusining funksional imkoniyatlari keng, foydalanuvchilar doirasi esa turli qatlamlarni qamrab oladi. Bu korpus tilshunos olimdan boshlab oddiy til o‘rganuvchigacha bo‘lgan har bir kishi uchun **foydali elektron manbadir**. Korpus orqali har kim o‘zbek tilining real qo‘llanilishi bilan tanishishi, zarur ma’lumotni olishi va tahlil qilishi mumkin.

O‘zbek tilidagi mavjud korpuslar: baho, ustunlik va kamchiliklar

Hozirgi kunga kelib, o‘zbek tilida bir nechta korpus loyihalari mavjud yoki rivojlantirilmoqda. Ularning har biri o‘ziga xos ustunlik va ayrim cheklovlarga ega. Quyida mavjud korpuslarning qisqacha tavsifi va ularga berilayotgan umumiy baho keltiriladi:

1. **O‘zbek tilining milliy korpusi (uzbekcorpus.uz):** Bu – yuqorida batafsil tavsifi berilgan asosiy milliy korpus. Uning asosiy ustunligi – qamrovi keng va turli, ya’ni o‘zbek adabiy tilining deyarli barcha sohalarini o‘z ichiga olgan. Shuningdek, uning integratsiyalashgan subkorpuslar tizimi mavjud: masalan, parallel korpus, ta’limiy korpus, mualliflik korpusi kabi bo‘limlar milliy korpus tarkibida ajratilgan. Bu esa uni ko‘p funksiyali kompleks tizimga aylantiradi. Milliy korpusning yana bir afzalligi – uning onlayn ochiqligi va bepul foydalanish imkoniyatidir: istalgan foydalanuvchi internet orqali ro‘yxatdan o‘tib, korpusdan foydalanishi mumkin. Shu bilan birga, kamchilik tomonlari ham yo‘q emas. Avvalo, korpus hajmi hali dunyoning yirik korpuslari bilan qiyoslaganda kichik – taxminan o‘n millionlab so‘zlar (aniq raqamlar doimiy o‘sib bormoqda). Mutaxassislar milliy korpus hajmini kelgusida bir necha yuz million so‘zga yetkazishni rejalashtirmoqdalar. Yana bir cheklov – og‘zaki nutq materiallarining kamligi. Hozircha korpus asosan yozma matnlardan iborat bo‘lib, jonli og‘zaki nutq (masalan, audio yozuvlardan transkripsiyalar) bazasi endi-endi shakllanmoqda. Bu esa tilning og‘zaki ko‘rinishini tadqiq etishda noqulaylik tug‘diradi. Milliy korpusning yana bir muammosi – morfologik teglashning to‘liqligi masalasi: hali barcha matnlar to‘liq avtomatik teglangan emas, teglash sifati ham ba’zan oqsaydi (ayniqsa, murakkab qo‘shma so‘zlar yoki dialektal birliklarda). Biroq bu kamchiliklarni bartaraf etish ustida ishlar davom etmoqda – tez orada korpusning yanada takomillashgan yangi versiyasi (“yangi versiya” havolasi mavjud) ishga tushirilishi rejalangan. Umuman, milliy korpus o‘zbek tilini raqamlashtirishdagi tarixiy qadam bo‘lib, uning yaratilishi tilimiz taraqqiyotida muhim madaniy voqelik sifatida baholanmoqda.

2. **Ta’limiy korpus (uzschoolcorpara.uz):** Ushbu korpus milliy korpusning tarkibiy qismi bo‘lsa-da, alohida loyiha sifatida ham e’tirofga loyiq. Ta’limiy korpusning ustun tomoni – maqsadli auditoriyaga mos tanlanganligidir. Maktab darsliklari va o‘quv matnlari asosida tuzilgan bu korpusdan bevosita ta’lim jarayonida foydalanish mumkin. Foydalanuvchi uchun sodda interfeys va didaktik funksiyalar (masalan, sinf darajasi bo‘yicha tekstlarni tanlash, mashq uchun jumlalarni tasodifiy tanlab berish va h.k.) joriy qilingan. Ta’limiy korpus kichikroq hajmga ega (bir necha yuz mingdan bir necha million so‘zgacha), bu esa uni tezkor va yengil qiladi. Kamchilik tomoni – u milliy tilning to‘liq manzarasini bermaydi, chunki faqat darslik materiallari bilan cheklangan. Biroq uning maqsadi boshqa – o‘quvchilar uchun til materiallarini yetkazish. Shuningdek, ayrim matnlarning eskirgan bo‘lishi (eski darsliklardan kiritilgan bo‘lishi) mumkin, bu holatda zamonaviy til holatini aks ettirishda cheklov tug‘iladi. Ta’limiy korpus jamoasi bu muammoni yechish uchun doimiy ravishda yangi darsliklar bilan bazani yangilab bormoqda.

3. **Mualliflik korpusi:** Bu korpus hali rasmiy e’lon qilingan to‘liq mahsulot emas, balki ilmiy tadqiqotlar doirasida yaratilgan kichik korpusdir. Sh.

Hamroyevaning tashabbusi bilan boshlangan ushbu korpusni yaratishdan maqsad – muayyan mualliflar asarlari tilini chuqur o‘rganish, avtorial uslubni statistik tahlil qilishdir. Masalan, dastlabki tajriba sifatida O‘zbekiston xalq shoiri Usmon Azim dramatik asarlarining mualliflik korpusi tuzilgan. Mualliflik korpusining ustunligi shundaki, u adabiyotshunoslik va stilistika uchun noyob material beradi: bitta muallifning barcha asarlari yagona bazada bo‘lgani holda, ularning lug‘at tarkibi, uslubiy qurilishi oson taqqoslanadi. Bu plagiatni aniqlash, mualliflikni aniqlash kabi masalalarda ham qo‘l kelishi mumkin. Kamchiligi esa – qamrov tor, umumtil xususiyatlarini tadqiq etib bo‘lmaydi. Har bir muallif uchun alohida kichik korpus tuzish talab etiladi, bu esa ko‘p mehnat va vaqt oladi. Hozircha mualliflik korpusi loyihasi eksperimental bosqichda, lekin uning nazariy asoslari ishlab chiqilgan (Hamroyevaning PhD tadqiqoti) va kelgusida bu yo‘nalish boyitilishi mumkin.

4. **Parallel korpus:** O‘zbek tilida parallel korpuslar tuzish borasida ham ilk qadamlar tashlangan. Milliy korpus tarkibida parallel bo‘lim mavjud bo‘lsa-da, hozircha unda oz miqdorda material bor (ayniqsa, o‘zbek–ingliz va o‘zbek–rus tarjimalari). Ustunlik shuki, ushbu bazalar mashina tarjimasi va tarjimashunoslik uchun asqatadi – bir jumlaning ikkita tildagi variantini qiyosiy tahlil qilish, adekvat va noadeekvat tarjima namunalarini topish mumkin. Kamchilik tomoni esa – parallel korpus yig‘ish nihoyatda murakkab: sifatli tarjima topish, matnlarni bir-biriga moslashtirib bog‘lash (alignment) ko‘p manbalarni talab qiladi. Shunga qaramay, 2022–2023 yillarda ayrim nodavlat tashabbuslar bilan o‘zbek tilidan ingliz va rus tiliga tarjimalar korpusi yuzaga keldi (masalan, Tatoeba bazasi orqali jamoaviy yig‘ilgan jumla juftliklari). Ayni paytda bu korpuslar kichik hajmda bo‘lsa-da, istiqbolda ularni kengaytirish rejalari bor.

5. **Jahon korpuslari doirasidagi o‘zbek tili korpuslari:** Qizig‘i shundaki, o‘zbek tiliga oid bir necha korpuslar xorijiy ilmiy markazlar tomonidan ham yaratilgan. Masalan, Germaniyadagi Leyptsig universiteti lingvistika instituti Leipzig Corpora Collection dasturi doirasida 2017-yilda “Uzbek web corpus” tuzilgan bo‘lib, unga internetdan avtomatik yig‘ilgan ~9 million so‘zdan iborat matnlar kiritilgan. Bu korpus asosan veb-sahifalardagi jamoat matnlarini o‘z ichiga oladi va alohida qidiruv interfeysi orqali taqdim etilgan. Uning ustunligi – ma’lumot avtomatik to‘plangani sabab tezkor yaratilgan va internet tilini aks ettiruvchi namunalarni beradi. Kamchiligi – tozaligi nisbatan past: chunki avtomatik yig‘ishda reklama, takroriy axborot kabi keraksiz qismlar ham tushgan bo‘lishi mumkin. Shuningdek, Leyptsig korpusida lingvistik annotatsiya chuqur emas – faqat oddiy konkordans qidiruv imkoni bor, xolos. Yana bir misol, Rossiya Fanlar akademiyasi Tilshunoslik institutida “Corpus Turcicum” loyihasi doirasida turkiy tillarning solishtirma korpusi yaratilgan, ularda ham o‘zbekcha matnlar qatnashadi (asosan matnlarning ruscha tarjimalari bilan parallel holatda). Bunday xalqaro loyihalar

o‘zbek tilini jahon korpus maydoniga kiritayotgani bilan ahamiyatli, lekin ular milliy korpusning o‘rnini bosolmaydi – chunki maqsad va mezonlar boshqacha.

Yuqoridagi tahlillar asosida umumiy xulosa qilish mumkinki, **o‘zbek tilining korpuslari** endigina shakllanish bosqichida bo‘lsa-da, qisqa muddatda bir necha turdagi korpuslar paydo bo‘ldi va ular bir-birini to‘ldiruvchi xususiyatga ega. Milliy korpus umumo‘zbektillik manzarani bersa, ixtisoslashgan korpuslar maxsus ehtiyojlarni qondiradi. Ustun jihatlar – tilimiz nihoyat korpusli tillar safiga qo‘shilgani, tadqiqotchilar va mutaxassislar qo‘lida qudratli vosita paydo bo‘lgani. Kamchilik jihatlar – hali hajm va qamrovda cheklilik, annotatsiya darajasining to‘liq emasligi, ayrim korpuslar orasida integratsiya yo‘qligi (masalan, parallel va milliy korpuslar bazasini birlashtirish masalasi). Biroq bu kamchiliklar vaqt o‘tishi bilan bartaraf etilishi tabiiy, zero korpuslarning o‘zi ham doimiy rivojlanishda bo‘ladigan tuzilmalardir.

Jahon tajribasi bilan qisqa qiyosiy tahlil

O‘zbek tilining korpuslari rivoji jahon tajribasi bilan solishtirilganda endigina boshlang‘ich bosqichda ekani ko‘zga tashlanadi. Ingliz, rus, fransuz kabi yirik tillarda korpus lingvistikasining **o‘nlab yillik tarixi mavjud bo‘lib**, ular yaratgan resurslar bugungi kunda billionlab so‘zlarni qamrab oladi. Masalan, Britaniya Milliy Korpusi 1990-yillarda 100 million so‘z bilan tuzilgan bo‘lsa, hozirda **Bank of English** va **Google Ngrams** kabi loyihalar milliardlab so‘zli korpuslarni taqdim etmoqda. Rus tilining milliy korpusi **(RNC)** 300 milliondan ortiq so‘z birligini o‘z ichiga oladi va unda nafaqat grammatik, balki semantik va pragmatik teglash ham joriy etilgan. Bu tillar korpuslarida matnlar soni ko‘pligi, xilma-xilligi, hatto sheva va lahjalar, tarixiy davrlar bo‘yicha alohida bo‘linmalari borligi bilan ajralib turadi.

Shu bilan birga, jahon tajribasi shuni ko‘rsatadiki, **korpus tuzish hech qachon yakunlangan loyiha bo‘lmaydi** – doimo rivojlanadi. Ingliz tili uchun ham keyingi yillarda ixtisoslashgan korpuslar (masalan, Twitter korpusi, huquqiy matnlar korpusi, tibbiy matnlar korpusi) paydo bo‘lgan. O‘zbek tilida ham xuddi shunday yo‘l kuzatilmoqda: umumiy milliy korpus yonida turli sohaviy korpuslar shakllanmoqda (ta’limiy, parallel, mualliflik va hokazo). Bu jihatdan o‘zbek tilining korpus loyihasi jahon andozalarini asta-sekin ergashmoqda, deyish mumkin.

Korpuslarning sifat ko‘rsatkichlari jihatidan qaraganda, yirik tillar korpuslari allaqachon **qo‘shma annotatsiyalangan (morfologik, sintaktik, semantik)** ko‘rinishga o‘tgan. Masalan, Amerika Milliy Korpusi matnlari sintaktik daraxt tuzilmalarigacha belgilangan, shuningdek, korpusdan bevosita mashina tarjima modellarini o‘rgatish, intonatsion tahlillar qilish kabi murakkab funksiyalarni qo‘llab-quvvatlaydi. O‘zbek tilida bunday chuqurlikdagi korpus ishlanmalari hali reja bosqichida, biroq ilmiy izlanishlar boshlangan. B. Elov va hamkorlarining NER (Named Entity Recognition), lemmatizatsiya va sintaktik tahlil bo‘yicha qilayotgan

ishlari buning asosini yaratmoqda. Ya’ni, yaqin yillarda o‘zbek korpusida ham shaxs nomlari, joy nomlari alohida belgilanishi, gaplarning sintaksis daraxti avtomatik chiqarilishi kutilmoqda – bular jahon tajribasida sinovdan o‘tgan va biz uchun yangi bo‘lgan jihatlar.

Qiyosiy tahlilda e’tiborga molik yana bir omil – **resurslar va hamjamiyat** masalasi. Dunyoning yirik korpuslari ortida ko‘pincha katta ilmiy jamoalar, davlat dasturlari yoki universitet laboratoriyalari turadi. O‘zbekistonda korpus yaratish ishlariga so‘nggi yillarda davlat va ilmiy jamiyat e’tibor qaratib, Innovatsion rivojlanish vazirligi grantlari ajratila boshlandi. Masalan, 2020–2022 yillarda “O‘zbek tilining milliy korpusi” bo‘yicha maxsus innovatsion loyiha moliyalashtirildi va bu ishlarni yangi bosqichga olib chiqdi. Bu holat jahon tajribasiga mos – ko‘plab davlatlar milliy korpuslarni **milliy loyihalar darajasida** qo‘llab-quvvatlaydilar (Angliyaning BNC loyihasi Britaniya Kengashi homiyligida bo‘lgani kabi). Bundan tashqari, jahon korpus hamjamiyatlarida (masalan, Corpus Linguistics konferensiyalarida) endi o‘zbek tilining korpusi haqida ham ma’lumotlar paydo bo‘lib, xalqaro mutaxassislar bizning tajribamizni o‘rganishga qiziqish bildirmoqdalar.

Jahon tajribasi bilan qiyoslanganda, o‘zbek korpusining yana bir farqi – **bu tilning xususiyatlari bilan bog‘liq**. Agglutinativ va o‘z kelib chiqishi turkiy bo‘lgan tillar korpuslarini yaratishda turli muammolar bo‘ladi; masalan, turk tilida ham milliy korpus yaratish davomida so‘zlarning juda ko‘p shakllanmasi muammo tug‘dirgan. O‘zbek tilida ham shunday – ingliz yoki rus tillariga qaraganda **morfolgiya ko‘lami keng**, shu bois avtomatik teglash, lemmatizatsiya yechimlari murakkabroq. Qiyosiy o‘rganishlardan ma’lumki, turk tilining milliy korpusini tuzish jarayonida grammatikaviy omillar chuqurroq hisobga olingan, natijada hozir turk milliy korpusida har bir so‘zning butun paradigmatik shakllari bog‘langan holda ko‘rsatiladi (masalan, bir fe’lning barcha zamon va shaxs formalarini bir joyga jamlab ko‘rish mumkin). O‘zbek korpusida hozircha bu jihat cheklangan – lekin aynan shu turdagi funksiyalar ham joriy etish rejalari mavjud.

Xulosa. O‘zbek tili milliy korpusi tilshunoslik va texnologiya sohasida muhim yutuq bo‘lib, uning kelajakdagi rivojlanishi yanada katta imkoniyatlarni ochadi. Korpusning hajmini kengaytirish, annotatsiya sifatini oshirish va zamonaviy bigdata texnologiyalari bilan integratsiya qilish uning samaradorligini oshiradi. Bigdata yordamida korpusni qayta ishlash va tahlil qilish imkoniyatlari tilni o‘rganishda yangi yondashuvlarni keltirib chiqaradi va o‘zbek tilini global miqyosda raqobatbardosh qilishga xizmat qiladi.

O‘zbek tili korpusi tilshunoslik, ta’lim va texnologiya sohalarini integratsiyalashgan holda rivojlantirishning muhim vositasi hisoblanadi. Tarixiy jihatdan u an’anaviy usullardan zamonaviy BigData yechimlariga o‘tgan bo‘lsa-da,

hali ham resurslar etishmovchiligi va metodologik muammolar mavjud. Kelajakda korpusning samaradorligini oshirish uchun xalqaro hamkorlik va AIga asoslangan yondashuvlar asosiy yo‘nalish bo‘lishi kerak. Jahon tajribasi o‘zbek tilining korpusini rivojlantirish uchun o‘ziga xos namuna va motivasiya vazifasini o‘tayapti. O‘zbek korpusi hozircha hajm va chuqurlikda yetakchi tillar korpuslaridan orqada bo‘lsa-da, ularni namunali model sifatida olib, tez sur’atlarda rivojlanmoqda. Bir necha yil ichida o‘zbek tilining milliy korpusi Markaziy Osiyo tillari orasida eng ilg‘or elektron til resursiga aylanishi kutilmoqda. Eng muhimi, milliy korpusimiz jahon korpuslar oilasiga kirib, xalqaro hamkorlik uchun ham ochiq bo‘lmoqda – masalan, ko‘plab xorijiy tadqiqotchilar endi uzbekcorpus.uz ma’lumotlaridan foydalana olishadi. Bu esa tilimizning xalqaro maydondagi nufuzini oshirishga xizmat qilishi shubhasiz. Ushbu paragrafda Milliy korpus loyihasi bo‘yicha ma’lumotlar O‘zbekistonda chop etilgan ilmiy maqolalar va konferensiya materiallariga tayangan holda yoritildi, xususan B. Elov, Sh. Hamroyeva, N.Z. Abdurahmonova va ularning hamkorlarining ishlaridan olindi. Mazkur bo‘limdagi dalillar va tahlillar ushbu manbalardagi faktlarga asoslangan bo‘lib, ilmiy izchillik tamoyiliga amal qilingan. Bu yondashuv o‘zbek tili korpusining tarixiy shakllanishini, maqsad va vazifalarini chuqur anglash hamda uni jahon tajribasi bilan bog‘lash imkonini beradi.

Foydalanilgan adabiyotlar:

1. McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.
2. Abdurahmonova N.Z. O‘zbek tili elektron korpusining kompyuter modellari: Filol. fan. d-ri. (DSc) diss. avtoreferati. – Toshkent, 2021.
3. https://uzschoolcorpara.uz/uz/Search
4. Elov, B., & Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).
5. Hamroyeva, S., Abdullayeva, O., & Uzoqova, M. (2022). O‘zbek tilida pos tegging masalasi: muammo va takliflar. Uzbekistan language and culture, 5(2), P. 51-68.
6. https://scholar.google.com/citations?user=wYCKlsoAAAAJ&hl=ru
7. Qarshiyev, A., Tursunov, M., & Maxmidov, S. (2022). O‘zbek tili milliy korpusini loyihalash. Computer linguistics: problems, solutions, prospects, *1*(1).
8. Kasimova, M. B. (2024). O‘zbek tili milliy korpusida darajalanish hodisasining berilishi. Miasto Przyszłości, *44*, 641-645.
9. Xudayberganov, N. (2024). O‘zbek tili korpusiga morfologik ishlov berish. Computer linguistics: problems, solutions, prospects, *1*(1).
10. Elov, B., Hamroyeva, S., Alayev, R., Xusainova, Z., & Yodgorov, U. (2023). O‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and artificial intelligence, *1*(3), 117-129.

PARALLEL KORPUSLARDA SENTIMENT TAHLIL QILISHDA “ALIGNER” VOSITASINING O‘RNI

Matyakubova Noila Shakirjanovna,
tayanch doktaranti
matyakubovanoila@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. Parallel korpuslarda sentiment tahlilini amalga oshirishda “Aligner” vositalari muhim rol o‘ynaydi. Ular matnlarni to‘g‘ri moslashtirish orqali SLda berilgan gapning hissiy holatini saqlashda, tarjimaning sifatini oshirishda va semantik xatolarni kamaytirishda yordam beradi. “Aligner” vositalarining samarali ishlatilishi parallel korpuslardan to‘liq foydalanishga, shuningdek, turli tillarda kayfiyatni aniq aks ettirishga imkon yaratadi. Bu jarayonni mukammal bajarish, parallel korpuslar asosida yaratilgan tarjima tizimlarining sifatini oshirishga yordam beradi va umumiy tilni qayta ishlash sohasiga katta hissa qo‘shadi.

Annotation. Aligner software tools play a crucial role in conducting sentiment analysis on parallel corpora. By accurately aligning texts, they help preserve the emotional tone of a sentence in the source language (SL), improve translation quality, and reduce semantic errors. The effective use of aligner tools enables full utilization of parallel corpora and allows for accurate representation of sentiment across different languages. Proper implementation of this process contributes to enhancing the quality of translation systems built on parallel corpora and makes a significant contribution to the broader field of natural language processing.

Аннотация. Программные инструменты Aligner играют ключевую роль в проведении сентимент-анализа на параллельных корпусах. За счёт точного выравнивания текстов они помогают сохранить эмоциональную окраску предложений, заданных на языке-источнике (SL), повысить качество перевода и уменьшить количество семантических ошибок. Эффективное использование инструментов выравнивания позволяет полноценно использовать параллельные корпуса, а также точно передавать настроение на разных языках. Корректное выполнение данного процесса способствует повышению качества систем перевода, основанных на параллельных корпусах, и вносит значительный вклад в развитие области обработки естественного языка.

Kalit so‘zlar. *Parallel korpus, Aligner vositasi, asliyat til (SL), tarjima tili (TL), sentiment tahlil, ijobiy, salbiy, neytral.*

Kirish. Parallel korpuslar – bu ikki yoki undan ortiq tillarda o‘xshash mazmundagi matnlar to‘plamidir. Tabiiy tilni qayta ishlash (NLP) sohasida parallel korpuslar tarjima tizimlarini o‘rgatishda, til o‘rtasidagi so‘z va frazalarni moslashtirishda keng qo‘llaniladi. Biroq, parallel korpuslardan samarali foydalanish uchun ularni to‘g‘ri moslashtirish zarur. “Aligner” vositalari parallel matnlar orasidagi to‘g‘ri so‘z yoki ibora moslashuvlarini topish uchun ishlatiladi. Sentiment tahlili, matndagi kayfiyatni aniqlashda yordam beruvchi jarayon sifatida, parallel korpuslar orqali samarali o‘rganilishi mumkin.

Parallel korpuslarni moslashtirishda “Aligner” vositalarining ishlatilishi, til o‘rtasidagi sintaktik va semantik farqlarni kamaytirishga yordam beradi, ya’ni bir tildagi so‘z yoki iborani boshqa tilga to‘g‘ri moslashtirishda ishlatiladi. Bu vositalar yordamida sentimentni to‘g‘ri aks ettiruvchi so‘zlar va iboralar aniqlanadi, salbiy yoki ijobiy kayfiyatini aks ettiruvchi so‘zlari, boshqa tilda ham to‘g‘ri tarjima qilinganligini aniqlashga yordam beradi. “Aligner” vositalari parallel korpuslarning sifatini oshiradi va asliyat til (SL) matnidagi kayfiyatni tarjima til (TL) matnida to‘g‘ri saqlashga yordam beradi.

Sentiment tahlili va “Aligner” vositalarining birgalikda ishlatilishi, parallel korpuslarda kayfiyatni aniq aks ettirishga va tarjima tizimlarining sifatini oshirishga imkon yaratadi. Shu bilan birga, bu yondashuv mashina tarjimasi tizimlarining samarasini oshiradi va ko‘p tilli NLP modellarini kuchaytiradi. Maqola parallel korpuslarni to‘g‘ri moslashtirish orqali sentiment tahlilining aniqligini ta’minlashning ilmiy va amaliy ahamiyatini ko‘rsatadi.

Adabiyotlar tahlili.

So‘nggi yillarda sentiment tahlili (sentiment analysis/opinion mining) yo‘nalishlarida xalqaro va milliy miqyosda ko‘plab ilmiy tadqiqotlar olib borilmoqda. Sentiment tahlili ijtimoiy tarmoqlar, mahsulot sharhlari, yangiliklar, fikr va munosabatni ifodalovchi matnlarda keng qo‘llanilmoqda [1].

Ko‘p tilli sentiment tahlili (Multilingual Sentiment Analysis, MSA) doirasida, turli tillar o‘rtasida semantik moslikni aniqlash muammosi dolzarb hisoblanadi[2]. Ayniqsa, kam resursli tillar (low-resource languages), jumladan, o‘zbek tili uchun bu masala yana ham keskinroq bo‘lib, parallel korpuslarga bo‘lgan ehtiyoj ortib bormoqda [3]. O‘zbek tilida sentimentga oid belgilangan korpuslar mavjud emasligi sababli, parallel matnlar asosida tahlil yuritish strategiyasi afzallik kasb etadi.

Bu borada “Aligner” vositalari ko‘p tilli sentiment tahlilida asosiy vositalardan biri sifatida e’tirof etiladi. IBM Model 1–5, GIZA++, hunAlign kabi an’anaviy moslashtirish vositalari uzoq yillar davomida parallel korpuslar uchun ishlatilib kelinmoqda [4/5]. So‘nggi yillarda esa neyron tarmoqlariga asoslangan “Aligner”lar (semantik moslikni chuqur kontekstual model yordamida aniqlash

imkonini yaratdi [6]. Ayniqsa, BERT, mBERT, XLM-R kabi pre-trained til modellari yordamida kontekstual semantika chuqurroq tahlil qilinmoqda.

O‘zbekistonlik olimlar ham so‘nggi yillarda matnlar semantikasini tahlil qilish, mashinaviy tarjima va korpus lingvistikasi yo‘nalishlarida tadqiqotlar olib bormoqda. Jumladan, N.Abdurahmonova, Sh.Xamroyeva, N.Matyakubova, S.Avezov, R.Karimov, A.Axmedova kabi tadqiqotchilar tomonidan o‘zbek tilida avtomatik tarjima, parallel korpuslar tuzish va lingvistik birliklarni moslashtirish usullari bo‘yicha tadqiqotlar e’lon qilingan[7/8]. Ushbu ishlar o‘zbek tili uchun til juftligi bo‘yicha parallel korpuslar yaratish va tahlil qilishning dastlabki bosqichlarini shakllantirgan.

Bundan tashqari, E.Kuriyozov, S.Allanazarova, U.Salaev, S.Matlatipov, N.Raximov, va U.Bekmuradov tomonidan o‘zbek tilidagi matnlar uchun NLP modellari, klassifikatsiya usullari, shuningdek, elektron korpuslar asosida sentiment tahlil uslublarini ishlab chiqish bo‘yicha amaliy tadqiqotlar olib borilgan[9/10]. Ularning ishlarida asosiy e’tibor matn segmentatsiyasi, belgilash va sentiment tahlil qilish imkoniyatlarini yaratishga qaratilgan.

Parallel korpuslar va sentiment tahlili

Parallel korpuslarda SL va TL matnlari orasidagi semantik va sintaktik o‘xshashliklarni o‘rganish, tilni qayta ishlashning muhim jihatidir. Sentiment tahlili – matnda ijobiy, salbiy yoki neytral kayfiyatni aniqlash jarayonidir. Parallel korpuslar yordamida, bir tilning matni boshqa tilga tarjima qilinayotganda, SL matn ifodalayotgan kayfiyat saqlanishi yoki o‘zgarishi mumkin. Shuning uchun parallel korpuslarda sentiment tahlilini amalga oshirishda asosiy e’tibor, kayfiyatni to‘g‘ri aks ettirishga qaratilishi lozim.

Sentiment tahlilining parallel korpuslarda o‘rganilishining ahamiyati shundaki, bu metod bir tilning hissiy holatini boshqa tildagi mos keluvchi elementlar bilan to‘g‘ri tarjima qilishga yordam beradi. Bunda, qayta tarjima jarayonida ijobiy yoki salbiy kayfiyatni o‘zgartirishga yo‘l qo‘ymaslik uchun sentiment tahlilidan foydalanish zarur.

Aligner vositalari va parallel korpuslarni moslashtirish

“Aligner” vositalari parallel korpuslarda SL va TL orasidagi moslikni aniqlashga hizmat qiladi. Ular matnlarni moslashtirish, ya’ni bir tildagi so‘zlarni yoki iboralarni boshqa tilga to‘g‘ri tarjima qilish uchun ishlatiladi. Aligner vositalari yordamida parallel matnlarning har bir elementini o‘zaro bog‘lash (align) mumkin. Bu vositalar so‘zlar va gaplar o‘rtasidagi o‘zaro bog‘lanishlarni aniqlashga imkon beradi, shuningdek, tarjimadagi noaniqliklarni va kayfiyatdagi farqlarni kamaytirishga yordam beradi.

Bugungi kunda asosan statistik va neyron tarmoqlariga asoslangan “Aligner”lar NLPning ko‘plab sohalarida keng qo‘llanilmoqda. Statistik “Aligner”lar matnlar orasidagi moslik statistik metodlar asosida aniqlaydi. Giza++, FastAlign kabi moslashtiruvchi vositalar bunga misol bo‘ladi. Neyron tarmoqlariga asoslangan “Aligner”lar so‘ngi avlod moslashtiruvchi vositalardan bo‘lib, matnlarni moslashtirishda yuqori aniqlik darajasidagi natijalar taqdim etadi va mashina o‘rganish modellari yordamida matnlarni samarali tarzda moslashtiradi.

“Aligner” vositalarining sentiment tahlilidagi o‘rni

Sentiment tahlilining muvaffaqiyatli amalga oshirilishi uchun, parallel korpuslarda hissiy holatni to‘g‘ri aks ettiruvchi so‘z va iboralarning to‘g‘ri moslashuvi zarur. Aligning (moslashtirish) jarayoni, so‘zlarning yoki iboralar orasidagi bog‘lanishni aniq belgilashga yordam beradi, bu esa sentiment tahlilining to‘g‘ri bajarilishi uchun zarur.

“Aligner” vositalari yordamida parallel korpuslarda ijobiy yoki salbiy kayfiyatga ega bo‘lgan so‘zlar to‘g‘ri moslashtiriladi. Masalan, ingliz tilidagi “*brilliant*” so‘zi o‘zbek tilida “*a’lo*” deb tarjima qilinadi, va ijobiy hissiy holatni saqlab qoladi. Agar bu so‘z noto‘g‘ri moslashtirilsa, hissiy holatni o‘zgarishiga olib kelishi mumkin. “Aligner” vositalari bu jarayonni aniq va to‘g‘ri amalga oshirishga yordam beradi.

Ba’zan til o‘rtasidagi semantik farqlar natijasida sentiment tahlili xatolari yuzaga kelishi mumkin. Misol uchun, ba’zi tillarda salbiy hissiy holatni ifodalovchi so‘zlar boshqa tildagi so‘zlarga mos kelmasligi mumkin. “Aligner” vositalari bu kabi xatolarni aniqlashda yordam beradi, chunki ular matndagi ijobiy yoki salbiy elementlarni to‘g‘ri bog‘lashda aniq va sifatli moslashuvni ta’minlaydi.

Sentiment tahlilining parallel korpuslarda samarali amalga oshirilishi uchun, sentimentga mos tarjima modellari ishlab chiqish zarur. Aligner vositalari bu modellarga kerakli parallel matnlarni moslashtirishda yordam beradi, bu esa ijobiy yoki salbiy kayfiyatni saqlab qolgan holda tarjima qilish imkonini beradi.

O‘zbek-ingliz parallel korpusida sentiment tahlilini amalga oshirish.

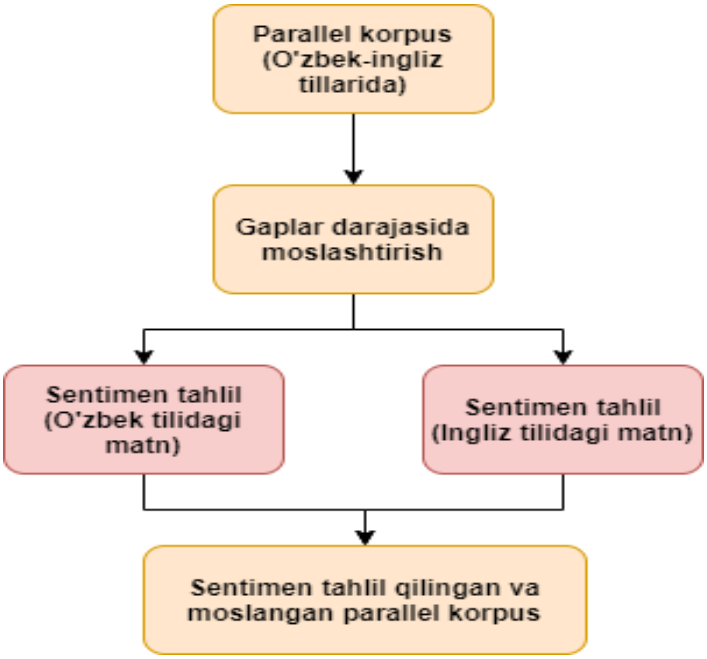
O‘zbek-ingliz paralell korpusida sentiment tahlil jarayonini amalga oshirish bir nechta bosqichda amalga oshiriladi (Qarang: 1-rasmga).

Dastlab parallel korpus shakllantirilishi lozim. Bunda SL va TLda fikrni ifodalovchi gaplar yoki qisqa matnlar jamlanadi. Matnlarga ishlov berish bosqichidan keyin ular alohida SL matnlar va TL matnlar shaklida alohida fayllar ko‘rinishida shakllantiriladi.

Korpus uchun matnlar jamlangach moslashtirish bosqichiga o‘tiladi. Bu bosqich o‘zbek-ingliz “Aligner” vositasi yordamida gaplar darajasida moslikni

aniqlash bosqichi amalga oshiriladi. Bu qadam til o‘zaro farqlarni yengillashtirish va keyingi sentiment tahlilida to‘g‘ri ma’nolarni aniqlash uchun muhimdir.

“Aligner”dan olingan moslangan ma’lumot asosida sentiment tahlili amalga oshiriladi. Bu yerda klassifikator yoki sentiment leksikonidan foydalangan holda, har bir segmentning ijobiy, salbiy yoki neytral ekanligi aniqlanadi.



1-rasm. Parallel korpusni moslashtirish va sentiment tahlil qilish bosqichi.

Moslashtirilgan va tahlil qilingan korpus natijasi jadval ko‘rinishida taqdim etiladi. Bunda ijobiy mazmunga ega gaplar *positive*, salbiy mazmundagilar *negative* va neytral mazmundagi gaplar *neutral* belgisi asosida belgilanadi.

1-jadval. Moslangan parallel korpus va sentiment tahlil natijasi.

ID	Inglizcha matn	O‘zbekcha matn	Sentiment tahlil natijasi
1	The book elicited a highly favorable impression from the reader.	Kitob o‘quvchida nihoyatda ijobiy taassurot uyg‘otdi.	positive
2	The quality of service at the restaurant was critically inadequate.	Restorandagi xizmat sifati juda past darajada bo‘lgan.	negative
3	Meteorological conditions appear to be stable and pleasant today.	Bugungi meteorologik sharoitlar barqaror va yoqimli kechmoqda.	neutral
4	The user expressed complete satisfaction with the provided service.	Foydalanuvchi ko‘rsatilgan xizmatdan to‘liq mamnunligini bildirdi.	positive

5	The recurrent tardiness of the individual provoked irritation.	Shaxsning muntazam kechikishi g‘azabni keltirib chiqardi.	negative
6	The cinematic experience was acceptable, yet lacked distinctiveness.	Kinematografik tajriba qoniqarli bo‘lsa-da, o‘ziga xoslikdan yiroq edi.	neutral
7	The previous day’s meeting was perceived as a pleasant encounter.	Kechagi uchrashuv yoqimli muloqot sifatida baholandi.	positive
8	The culinary offerings at this venue failed to meet expectations.	Ushbu maskandagi taomlar kutilgan darajada emas edi.	negative
9	The accommodation maintained high standards of cleanliness and comfort.	Turar joy tozalik va qulaylik mezonlariga to‘liq javob berdi.	positive
10	The application encountered frequent crashes, causing frustration.	Ilova tez-tez nosozlikka uchrab, foydalanuvchida noqulaylik tug‘dirdi.	negative

Xulosa.

Sentiment tahlilini kam resursli tillarda amalga oshirishda parallel korpuslar va ularni aniqlik bilan moslashtiruvchi “Aligner”lar asosiy vosita bo‘lib xizmat qiladi. Aynan o‘zbek tilida bu yo‘nalishdagi tadqiqotlar hali boshlang‘ich bosqichda bo‘lsa-da, parallel korpus tuzish, ularni yuqori aniqlikda moslashtirish va bu orqali sentimentni tahlil qilish istiqbolli va dolzarb ilmiy yo‘nalish hisoblanadi.

Foydalanilgan adabiyotlar ro‘yhati.

1. Liu B. Sentiment Analysis and Opinion Mining. Morgan and Claypool Publishers. 2012.
2. Barnes J., Klinger R., Schulte im Walde S. Projecting Embeddings for Domain Adaptation: Joint Modeling of Sentiment Analysis in Diverse Languages. In Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), 2018. – P. 818–830.
3. Bojar O., Chatterjee R., Federmann C., Graham Y., Haddow B., Huck M., Birch, A. Findings of the 2016 Conference on Machine Translation (WMT16). In Proceedings of the First Conference on Machine Translation, Berlin, Germany. 2016.
4. Och F., Ney H. A systematic comparison of various statistical alignment models. Computational Linguistics, 2003, 29(1), – P. 19–51.
5. Varga D., Halácsy P., Kornai A., Nagy V., Németh L., Trón V. Parallel corpora for medium density languages. In Proceedings of RANLP 2007, Borovets, Bulgaria, 2007. – P. 1–8.

6. Dou Z., Neubig G. Word Alignment by Fine-tuning BERT with Multilingual Objectives. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL 2021), 2021. – P. 8502–8512.

7. Abdurahmonova N. Mashina tarjimasining lingivistik asoslari. Akademnashr. Toshkent, 2012. – 30 b.

8. Xamroyeva Sh., Matyakubova N. Korpus lingvistikasida “Aligner” dasturlarning o‘rni / “Fan va Jamiyat” ilmiy-uslubiy jurnal. Nukus, 2023. – B. 74-76.

9. Matlatipov S., Rahimboeva H., Rajabov J., Kuriyozov E. Uzbek Sentiment Analysis Based on Local Restaurant Reviews. CEUR Workshop Proceedings, 3315, 2022. – P. 126–136.

10. Allanazarova, S. Korpus va lug‘atga asoslangan matnning sentiment tahlili. Xorazm Ma’mun akademiyasi axborotnomasi, 2024. 7(4), – P. 28-31.

SIFAT SO‘Z TURKUMINI GRAMMATIK POS TEGLASH

Mansurova Shaxinabonu Najmiddin qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
tiladabiyot@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada sifat so‘z turkumining grammatik jihatdan qismlariga ajratilishi va tabiiy tilni qayta ishlash (NLP) doirasida so‘z turkumlarini belgilash (POS tagging) jarayoni tahlil qilinadi. Sifatlar til tuzilishida muhim o‘rin tutib, ularning morfologik va sintaktik xususiyatlarini to‘g‘ri aniqlash matnlarni avtomatlashtirilgan qayta ishlashda katta ahamiyatga ega. Maqolada sifatlarning darajalari, shakllari va qo‘llanilishiga asoslangan POS teglash usullari hamda zamonaviy sun‘iy intellekt algoritmlari yordamida amalga oshiriladigan yondashuvlar haqida ma’lumot beriladi.

Annotation. This article analyzes the grammatical classification of the adjective part of speech and the process of part of speech tagging (POS tagging) within the scope of Natural Language Processing (NLP). Adjectives play an important role in language structure, and correctly identifying their morphological and syntactic properties is crucial for the automated processing of texts. The article provides information on POS tagging methods based on the degrees, forms, and usage of adjectives, as well as approaches implemented using modern artificial intelligence algorithms.

Аннотация. В данной статье анализируется грамматическая классификация прилагательного как части речи и процесс тегирования частей речи (POS tagging) в рамках обработки естественного языка (NLP). Прилагательные играют важную роль в языковой структуре, и правильное определение их морфологических и синтаксических свойств имеет большое значение для автоматизированной обработки текстов. В статье представлена информация о методах тегирования частей речи на основе степеней, форм и использования прилагательных, а также подходах, реализуемых с использованием современных алгоритмов искусственного интеллекта.

Kalit so‘zlar: *sifat, so‘z turkumlari, grammatik tahlil, POS tagging, tabiiy tilni qayta ishlash, morfologik tahlil, sintaktik tahlil, NLP, sun‘iy intellekt, mashinani o‘rganish.*

Kirish

Sifat so‘z turkumi, lingvistikada sifat sifatida tanilgan so‘zlar to‘plami bo‘lib, ob‘ektlar yoki hodisalarni ta’riflash va ularga xususiyatlar berish uchun ishlatiladi. Tilshunoslikda sifatlar odatda o‘zi bilan bog‘liq bo‘lgan nomlarni ta’riflash uchun

ishlatiladi, masalan, "yaxshi kitob", "baland tog‘" va "quyoshli kun". Bu so‘zlar ma’lum bir ob‘ektning xususiyatlarini aniqlab, uni boshqalardan ajratib turadi. Shuning uchun sifatlar tabiiy tilni qayta ishlash va matnni avtomatik tahlil qilishda muhim ahamiyatga ega.[1]

Grammatik POS teglash (Part-of-Speech Tagging, POS tagging) – tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining muhim vazifalaridan biri bo‘lib, unda har bir so‘zning grammatik rolini belgilash orqali matnni avtomatik ravishda tahlil qilish amalga oshiriladi. Grammatik pos teglash tilni analiz qilishda tilning morfologik va sintaktik xususiyatlarini aniqlashga yordam beradi va bu jarayon ko‘plab tilshunoslik va kompyuter fanlarida keng qo‘llaniladi. Ushbu jarayon yordamida har bir so‘zning grammatik kategoriyasi, masalan, ot, fe‘l, sifat, predlog va boshqa so‘z turkumlari aniqlanadi

Grammatik POS teglash jarayonida sifatlar matndagi o‘rniga va uning grammatik funktsiyasiga qarab teglar bilan belgilanishi kerak. Bu jarayon tillarning morfologik va sintaktik xususiyatlariga asoslanadi, va tahlil qilishda tilning turli xususiyatlari, masalan, sintaktik roli, kontekst va ma’no o‘zgarishi inobatga olinadi. Sifat so‘zlarining grammatik teglari, tilning tuzilishiga qarab, turli xil bo‘lishi mumkin, masalan, "yaxshi" sifatining "JJ" yoki "ADJ" teglariga ega bo‘lishi mumkin.

Bugungi kunda sifatlarni POS teglash tizimlarida to‘g‘ri aniqlash uchun bir qator ilg‘or metodlar mavjud. Masalan, mashina o‘rganish, chuqur o‘rganish (deep learning) va tabiiy tilni qayta ishlashning boshqa zamonaviy usullari sifat so‘zlarni aniqlashda yordam berishi mumkin. Ayniqsa, chuqur o‘rganish algoritmlari va neyron tarmoqlarining qo‘llanilishi POS teglash jarayonini yanada samarali va aniqlik bilan bajarishga yordam beradi. Shu bilan birga, sifatlar o‘zgarishi mumkin bo‘lgan kontekstda ham to‘g‘ri tahlil qilish zarurati paydo bo‘ladi.

Shu sababli, bu maqola sifat so‘z turkumi va uning grammatik POS teglashini amalga oshirishdagi qiyinchiliklar va hozirgi kunda mavjud bo‘lgan usullarni o‘rganish uchun mo‘ljallangan. U, shuningdek, tilni qayta ishlashda sifatlarni to‘g‘ri belgilashning ahamiyatini va bu jarayonni yaxshilashda qo‘llaniladigan zamonaviy texnologiyalarni tahlil qilishni maqsad qiladi.[2]

Muhokama va natijalar

Sifatni o‘rganish tizimi materiali leksik va grammatik jihatdan izchillikni saqlagan holda rivojlantirib borish, shuningdek, murakkablashtirishni nazarda tutadi. 1-sinfda o‘quvchilar sifatning leksik ma’nosini va uning qanday? va qanaqa? so‘roqlariga javob berishini o‘rganadilar. 2-sinfda sifat so‘z turkumi sifatida to‘liq o‘zlashtiriladi, 3-sinfda esa ilgari o‘rganilganlarni takrorlab, grammatik materialga asoslangan -roq qo‘shimchasi bilan yasalgan sifatlarning yozilishi o‘rgatiladi

(terminlar ishlatilmaydi). Ona tili va o‘qish darslarida o‘quvchilarning nutqi yangi sifatlar bilan boyitiladi va ilgari o‘rganilgan sifatlarning ma’nosiga aniqlik kiritiladi.

Sifat so‘z turkumining grammatik teglanishi quyidagicha:

Yaxshidan bog‘ qoladi, yomondan dog‘.

Keltirilgan gapda, yaxshi va yomon so‘zlari sifat so‘z turkumiga mansub bo‘lib, yaxshi(xususiyat sifati, asliy sifat, oddiy daraja, sodda sifat), yomon(xususiyat sifati, asliy sifat, oddiy daraja, sodda sifat) ko‘rinishida grammatik teglanadi.

Sifatni o‘rganish metodikasi uning lingvistik xususiyatlariga asoslanadi, chunki sifat predmetning belgilarini (rangi, shakli, hajmi, ko‘rinishi, ta’mi, hidi va h.k.) bildiradi. Sifatning leksik ma’nosini tushunish uchun 1-sinfda bolalar sifatning ot bilan bog‘lanishini o‘rganadilar. O‘quvchilar predmetning belgisini aytib, so‘rov yordamida sifat va ot birikmasini aniqlashni o‘rganadilar. Keyingi sinflarda bu bog‘liqlik yanada chuqurlashtiriladi. Shunday qilib, sifatni o‘rganish leksik va grammatik (morfologik va sintaktik) jihatdan amalga oshiriladi.

Boshlang‘ich sinflarda sifat quyidagi izchillikda o‘rganiladi: [3]

1. Sifat bilan tanishtirish (1-sinf),
2. Sifat haqida tushuncha berish (2-sinf),
3. Grammatik materialga bog‘liq sifatlar yozilishini o‘zlashtirish (3-sinf).

Birinchi bosqichda sifatning leksik ma’nosi va so‘roqlariga diqqat qaratiladi. Predmetning belgilarini, uning rangini, mazasi va shaklini ta’riflash uchun sifatlar tanlanadi. Shunday qilib, o‘quvchilar sifatlarni tanib olish uchun predmetlar yoki rasmlar asosida mavzularni muhokama qiladilar. Bu darslarda ko‘rgazmalar, mashqlar va suhbatlar keng qo‘llaniladi.

Ikkinchi bosqichda o‘quvchilarga sifatning grammatik xususiyatlari va uning nutqdagi rolini o‘rganish muhimdir. O‘quvchilar sifatlarni mazmunga mos ravishda ishlatishni, fikrni aniq ifodalash uchun sifatlardan foydalanishni o‘rganadilar. Ayniqsa, sifatlarning belgilovchi va bog‘lanuvchi xususiyatlari ta’kidlanadi. O‘quvchilar predmetning rangini, shaklini, mazasi va boshqa sifatlarini tanlab, ularni tasvirlashda foydalanadilar.

Uchinchi bosqichda sifatlar ustida ishlash davom etadi, ayniqsa, –roq qo‘shimchasi bilan yasalgan sifatlar va ba’zi maxsus sifatlarni to‘g‘ri yozish o‘rgatiladi. Bu bosqichda o‘quvchilarning mustaqil ishlash qobiliyati oshadi, mashqlar asta-sekin murakkablashtiriladi. Mashq materialida yaxshiroq, aqilliroq kabi sifatlar hamda tip-tiniq, sap-sariqiq kabi sifatlar ishlatiladi. Bu jarayon sifatni o‘rganishda uning ot bilan o‘zaro bog‘lanishini aniq tushunishni ta’minlaydi.

Sifat so‘z turkumining grammatik POS teglashidagi ahamiyati tilshunoslikda alohida o‘rganilayotgan sohalardan biridir. Sifatlar tilning eng muhim qismlaridan biri bo‘lib, ular nafaqat ob‘ektning xususiyatlarini ta’riflaydi, balki uning boshqa ob‘ektlar bilan taqqoslanishiga imkoniyat yaratadi. Sifatlarni to‘g‘ri tanib olish va belgilash, so‘zlarni to‘g‘ri tushunish va tilni qayta ishlashda aniq natijalarga erishish uchun zarur hisoblanadi.

1-jadval. O‘zbek tilidagi POS teglar ro‘yxati (qisqa variant)[4]

№	POS teg	Ingliz	Belgilanishi
1	Ot	Noun	N
2	Sifat	Adjective	JJ
3	Fe‘l	Verb	VB
4	Son	Number	NUM
5	Ravish	Adverb	RR
6	Olmosh	Pronoun	P
7	Bog‘lovchi	Conjunction	C
8	Ko‘makchi	preposition	II
9	Yuklama	particle	Prt
10	Modal so‘z	modal words	MD
11	Undov so‘z	exclamatory words	UH
12	Taqlid so‘zlar	imitation words	IM

Grammatik POS Teglashning muhimligi

POS teglash (Part of Speech tagging) – bu tabiiy tilni qayta ishlashning eng muhim qadamlaridan biridir. Bu jarayon so‘zlarni ma’lum bir grammatik turkumlarga ajratishni anglatadi. Sifat so‘zlarining to‘g‘ri aniqlanishi, tilni qayta ishlash tizimlari uchun asosiy maqsad bo‘lib, so‘zning sintaktik rolini aniqlash va uning kontekstiga qarab grammatik funktsiyasini belgilashni talab qiladi. Sifatlar odatda, nomlar bilan bog‘lanib, o‘zgaruvchi xususiyatlarni ko‘rsatadi, va bu xususiyatlarni to‘g‘ri tasvirlash uchun zamonaviy tizimlar yuqori aniqlikda ishlashini ta’minlash zarur.[5]

Sifatlarni belgilashdagi qiyinchiliklar

Sifat so‘zlarini to‘g‘ri belgilashda bir qancha qiyinchiliklar mavjud. Masalan, ba’zi sifatlar ko‘p ma’nolilikka ega bo‘lishi mumkin. Bu holatlarda kontekstni hisobga olgan holda sifatning qanday funktsiya bajarayotgani aniqlanishi kerak. Misol uchun, "katta" so‘zi biror narsa o‘lchamining kattaligini bildirishi mumkin, ammo ba’zan bu sifat “katta odam” kabi iboralarda ob‘ektning ijtimoiy yoki siyosiy holatini ko‘rsatishi mumkin. Shunday qilib, sifatlarni kontekstga qarab to‘g‘ri belgilash muhimdir.

Yana bir qiyinchilik shundaki, ba’zan sifatlar boshqa so‘zlar bilan bog‘lanib, boshqa grammatik funktsiyalarni bajarishi mumkin. Misol uchun, "yaxshi" so‘zi

sifat sifatida ishlatilsa-da, ba’zan u modifikator sifatida ishlatilishi mumkin. Bu holatlarda POS teglash tizimi sifatning qanday rolda ishlashini aniqlashda qiyinchiliklarga duch kelishi mumkin.

Zamonaviy texnologiyalar va yondashuvlar

Bugungi kunda sifatlarni grammatik POS teglash jarayonida qo‘llaniladigan texnologiyalar yangi imkoniyatlar yaratmoqda. Mashina o‘rganish (machine learning) va chuqur o‘rganish (deep learning) algoritmlari sifatlarni aniqlashda yuqori samaradorlikni ta’minlashga yordam beradi. Neyron tarmoqlarining rivojlanishi POS teglash tizimlarining aniqligini sezilarli darajada oshirgan. Ayniqsa, chuqur o‘rganish usullari, matnning ma’nosini yaxshiroq tushunish va sifatlarning to‘g‘ri funktsiyasini aniqlashga imkon beradi. Biroq, bu texnologiyalarni samarali qo‘llash uchun katta hajmdagi ma’lumotlar va kuchli hisoblash quvvatiga ehtiyoj bor.

Tadqiqotlar shuni ko‘rsatdiki, sifatlarni grammatik POS teglash jarayonida qo‘llaniladigan zamonaviy yondashuvlar yuqori darajada samarali bo‘lishi mumkin. Ayniqsa, chuqur o‘rganish va mashina o‘rganish usullari yordamida sifatlarning to‘g‘ri aniqlanishi va tahlil qilinishi tilni qayta ishlash tizimlarining aniqligini oshiradi. Biroq, bu usullarni qo‘llashda kontekstual omillarni, ya’ni so‘zlarning ma’nosini va sintaktik roli o‘zgarishini inobatga olish zarur. Bu, o‘z navbatida, POS teglash tizimlarining aniqligini va samaradorligini oshirishga imkon beradi.

Shuningdek, sifatlar o‘zgaruvchan va ko‘p ma’noli bo‘lgani sababli, POS teglash tizimlari turli kontekstlarni hisobga olib, ularni aniq ajratib belgilashni talab qiladi. Bu, o‘z o‘rnida, texnologiyalarni takomillashtirish va yanada yuqori aniqlikni ta’minlashga olib keladi.[6]

Takliflar:

1. **Grammatik belgilash tizimlarini takomillashtirish**: POS teglash tizimlarining samaradorligini oshirish uchun chuqur o‘rganish usullarini qo‘llashni davom ettirish va ularni yanada takomillashtirish zarur.

2. **Kontekstual yondashuvni kuchaytirish**: Sifatlarning ma’nosini aniqlashda kontekstni hisobga olish tizimning aniqligini oshiradi.

3. **Ma’lumotlar to‘plamlarini kengaytirish**: Yangi va yanada diversifikatsiyalangan ma’lumotlar to‘plamlarini yig‘ish sifatlarni to‘g‘ri aniqlashda yordam beradi.

4. **Mashina o‘rganish modellarini yanada takomillashtirish**: Sifatlarni grammatik belgilashda mashina o‘rganishning yangi metodlarini qo‘llash va ularni yuqori darajada avtomatlashtirish tizim samaradorligini oshiradi.

Natijada, sifatlarni grammatik POS teglashda zamonaviy texnologiyalarni to‘g‘ri qo‘llash tilni qayta ishlash va lingvistik tahlilni yanada samarali va aniq qilish imkonini beradi.[7]

Sifatlarni tahlil qilish, umumiy grammatik ma’noni belgilashni boshlaydi. Sifatlar o‘z xususiyatlariga ko‘ra ikki turga bo‘linadi: 1) asliy sifat, 2) nisbiy sifat. Sifat turkumi sintaktik jihatdan keng emas, lekin uning grammatik xususiyatlari muhimdir. Sifat turkumida egalik kategoriyasi "keyingi sifatni oldingi so‘zga bog‘lash, mansubiyat va xoslik ma’nolarini ifodalash" orqali aniqlanadi. Kesimlik kategoriyasi sifat turkumida juda cheklangan darajada namoyon bo‘ladi va ko‘pincha bog‘lama vositasi orqali ifodalanadi. Sifat so‘z turkumining teglanishi kengaytirilgan tizimda quyidagicha tasniflanadi: – o‘lik: [sifat], [nisbiy sifat], [oddiy daraja], [yasama sifat], [sodda sifat], [jismoniy harakat LMG]; – tuzuk: [sifat], [asliy sifat], [oddiy daraja], [tub sifat], [yasama sifat], [xususiyat LMG]; – yiroq: [sifat], [asliy sifat], [oddiy daraja], [sodda sifat], [tub sifat], [xususiyat LMG]; – yo‘g‘onroq: [sifat], [asliy sifat], [qiyosiy daraja], [sodda sifat], [tub sifat], [shakl LMG]. O‘tkazilgan tajribalar natijalariga ko‘ra, o‘zak ma’lumotlarni sintaktik vazifalar bilan birlashtirish morfologik jihatdan boy til uchun POS teglash natijalarini yaxshilaydi, bu esa NLP vazifalarini samarali hal qilishga yordam beradi. O‘zbek tili korpusi matnlarini POS teglash xususiyatlari va kichik kategoriyalar ro‘yxati keltirilgan. Fe‘l va sifat turkumlari grammatik kategoriyalari, tasnifi va misollar izohlangan. Maqola o‘rganilgan usullarni o‘zbek tili morfologik analizatori va boshqa murakkab NLP vazifalarida qo‘llash mumkinligini ko‘rsatadi.[8]

Xulosa

Sifat so‘z turkumini grammatik POS teglash jarayoni tilshunoslikda muhim o‘rin tutadi, chunki sifatlar predikatdagi boshqa so‘zlar bilan bog‘lanib, ularning semantik va sintaktik rolini aniqlashga yordam beradi. Sifatlarni to‘g‘ri teglash orqali o‘quvchilar va tadqiqotchilar lingvistik tuzilmalarning yanada aniqroq va samarali tahlilini amalga oshiradilar. Sifatlarning leksik va grammatik xususiyatlarini o‘rganish, ularni nutqdagi o‘rni va qo‘llanilishining izchil tarzda o‘rganilishi, tilning murakkab strukturasi haqida kengroq tasavvurga ega bo‘lishga imkon yaratadi.

Shuningdek, sifatlarni grammatik teglash jarayonida so‘zning morfologik xususiyatlari, jumladan, ularning qo‘shimchalar yordamida qiyosiy yoki orttirma darajalarga o‘tishini, shuningdek, sifatlar va otlarning o‘zaro bog‘lanishini aniqlash muhimdir. Bu metodik yondashuv til o‘qituvchilariga va tilshunoslarga sifatlarni o‘rganishda to‘g‘ri va samarali yo‘nalishlarni belgilashga yordam beradi.

Shunday qilib, sifat so‘z turkumini grammatik POS teglash, nafaqat tilshunoslik, balki til o‘rganish va ta’lim jarayonlarida ham samarali vosita sifatida

xizmat qiladi, o‘quvchilarga sifatlarning nutqda qanday ishlashini o‘rgatishda muhim o‘rin tutadi.

Foydalanilgan adabiyotlar:

1. Ahmadaliyev, B. (2020). O‘zbek tili grammatikasi. Toshkent: O‘zbekiston Milliy Universiteti.
2. Karimov, S. (2018). Tildagi morfologik va sintaktik tuzilmalar. Toshkent: Sharq.
3. Tohirov, M. (2019). O‘zbek tilida sifat so‘z turkumining sintaktik o‘rni. Toshkent: Ma’naviyat.
4. Mirzayev, A. (2021). Grammatik POS teglash va uning tilshunoslikdagi ahamiyati. Tashkent: Science and Education.
5. Vaxabov, A. (2017). Tilmashinalar va morfologiya. Toshkent: Universitet.
6. Shamsiyev, Z. (2022). Sifat va uning grammatik xususiyatlari. Toshkent: O‘zbekiston Yozuvchilar Uyushmasi.
7. Smith, J. (2016). Linguistic Analysis of Adjectives in Syntax and Semantics. Oxford University Press.
8. Gorman, J. (2015). Understanding POS Tagging in Natural Language Processing. Cambridge: Cambridge University Press.

DISKURSNI ANIQLASHDA KORPUSDAN FOYDALANISH

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Sulaymonova Mahfuza Shavkat qizi,
tayanch doktarant
ToshDO‘TAU

Annotatsiya. Mazkur maqolada zamonaviy tilshunoslik va kommunikativ metodika nuqtayi nazaridan diskurs kompetensiyasi fenomenining mazmun-mohiyati chuqur tahlil etiladi. Diskurs kompetensiyasi nafaqat grammatik va leksik bilimlar majmui, balki kontekstga mos muloqot qurish, ijtimoiy-madaniy omillarni hisobga olib, fikrni izchil va maqsadga yo‘naltirilgan tarzda ifodalash ko‘nikmalarini ham qamrab oladi. Maqolada til o‘rgatish jarayonida diskurs kompetensiyasining roli, uni shakllantirishda kommunikativ yondashuvning ustunliklari hamda interaktiv metodlarning samaradorligini asosli misollar va nazariy qarashlar orqali yoritiladi. Shuningdek, maqolada diskurs kompetensiyasining til tafakkuri va madaniy kompetensiya bilan uzviy bog‘liqligi, uning shaxslararo muloqotdagi funksional ahamiyati keng qamrovda ko‘rib chiqiladi. Tadqiqot natijalari asosida ta'lim amaliyoti uchun tavsiyalar ham ilgari suriladi.

Abstract. This article provides an in-depth analysis of the phenomenon of discourse competence from the perspectives of modern linguistics and communicative methodology. Discourse competence encompasses not only grammatical and lexical knowledge but also the ability to construct contextually appropriate communication, taking into account socio-cultural factors and expressing ideas coherently and purposefully. The paper highlights the significance of discourse competence in language teaching, emphasizing the effectiveness of communicative approaches and interactive methods through well-grounded examples and theoretical perspectives. Furthermore, it explores the intrinsic link between discourse competence, linguistic thinking, and cultural competence, as well as its functional relevance in interpersonal communication. Based on the research findings, the article offers practical recommendations for educational practice.

Аннотация. В данной статье проводится глубокий анализ феномена дискурсивной компетенции с точки зрения современной лингвистики и коммуникативной методики. Дискурсивная компетенция охватывает не только знание грамматики и лексики, но и умение строить уместное в конкретном контексте общение, учитывая социокультурные факторы и логично, целенаправленно выражать мысли. В статье освещается роль дискурсивной

компетенции в процессе обучения языку, акцентируется внимание на преимуществах коммуникативного подхода и эффективности интерактивных методов, подтвержденных теоретическими положениями и практическими примерами. Кроме того, рассматривается взаимосвязь дискурсивной, языковой и культурной компетенций, а также их функциональное значение в межличностной коммуникации. На основе результатов исследования предложены рекомендации для образовательной практики.

Kalit so‘zlar: *diskurs konpetensiyasi, konseptual va kommunikativ yondashuv, til o‘rganuvchi, izchillik-koherensiya.*

Kirish. Bugungi globallashuv davrida kommunikativ kompetensiyaning tarkibiy qismi sifatida diskurs kompetensiyasi til o‘rganish va o‘rgatish jarayonida alohida ahamiyat kasb etmoqda. Ushbu kompetensiya orqali shaxs ijtimoiy-madaniy kontekstda muloqotga kirisha oladi, mazmunli va izchil fikr bildira oladi hamda kommunikativ vazifalarni samarali bajara oladi. Zamonaviy til o‘rgatish metodikalarida diskurs kompetensiyasining ahamiyatiga alohida e’tibor qaratilmoqda, chunki u faqatgina til birliklarini bilish emas, balki ularni real nutqiy holatlarda qo‘llash qobiliyatini ham o‘z ichiga oladi. Zamonaviy til o‘rgatish metodikasida diskurs kompetensiyasi tilshunoslikning markaziy tushunchalaridan biri sifatida ajralib turadi. U nafaqat til tizimini bilish, balki til vositalaridan kontekstga mos va funksional tarzda foydalanish ko‘nikmasini anglatadi. Diskurs kompetensiyasi mavjud bo‘lmagan sharoitda shaxslararo muloqotning mukammalligi haqida gapirish qiyin. Bu kompetensiya til o‘rganish jarayonida o‘quvchining nutqiy rivojlanish darajasini aniqlovchi asosiy omil sifatida qaraladi. Til o‘rganishda grammatik bilimlar yetarli bo‘lmay, ularni amaliy kommunikativ vaziyatlarga mos tarzda qo‘llay olish muhimdir. Diskurs kompetensiyasi aynan shu jihatga urg‘u beradi: til vositalaridan real nutqiy vaziyatda maqsadga yo‘naltirilgan tarzda foydalanish- bu kompetensiyani rivojlantirish orqali o‘quvchi fikrni izchil ifodalash, suhbatdosh bilan muloqot davom ettirish, nutqiy strategiyalardan samarali foydalanish imkoniyatiga ega bo‘ladi. Diskurs kompetensiyasi ijtimoiy va madaniy omillar bilan uzviy bog‘liq. Har bir til madaniyat, qadriyat, urf-odat va ijtimoiy me’yorlar bilan uyg‘unlashgan bo‘lib, bu jihatlar muloqot mazmuniga bevosita ta’sir qiladi. Masalan, biror fikrni qanday uslubda ifodalash, qanday murojaat shakllaridan foydalanish, voqeani qanday bayon etish – bularning barchasi madaniy va ijtimoiy kompetensiyalarni talab qiladi. Diskurs kompetensiyasi esa bu bilimlarni amalda tatbiq etish vositasidir.

Metodologiya. Zamonaviy tilshunoslikda, ayniqsa kompyuter lingvistikasi sohasida, diskurs tahlili alohida ahamiyat kasb etadi. Diskurs faqatgina alohida gaplar emas, balki ular orasidagi bog‘liqlik, kontekst va ma’noni ham o‘z ichiga oladi. Bugungi kunda katta hajmdagi matnlar jamlanmasi – korpus va undan foydalanish vositalari yordamida diskursni aniqlash yanada aniq, empirik va tizimli

tus oldi. Korpus lingvistikasi – bu real hayotdagi til materiallari asosida tahlil olib boriladigan yo‘nalish bo‘lib, u yozma yoki og‘zaki nutq namunalarini katta hajmda jamlab, ularni maxsus dasturlar yordamida o‘rganadi. Diskurs tahlili esa tilning ma’no, kontekst, pragmatika, aloqa shakllari kabi yuqori darajadagi jihatlarini o‘rganadi. Korpus lingvistikasi yordamida diskurs tahlili subyektivlikdan xoli, ishonchli statistik asosda olib borilishi mumkin.

- Korpus lingvistikasi – til materiallarini to‘plab tahlil qiladi;
- korpuslar – real nutq namunalari yig‘indisidir;
- diskurs tahlili – korpusdan olingan ma’lumotlar asosida ma’no, uslub, muloqot shakllarini aniqlaydi.

Korpus vositalari va diskursni aniqlashdagi roli:

Vosita/usul	**Diskurs tahlilidagi vazifasi**	**Amaliy qo‘llanishi**
Konkordanser-Antconc	*Kalit so‘zlar va ularning kontekstlarini aniqlaydi*	*“Ammo” so‘zining ilmiy matnlardagi ishlatilishini o‘rganish*
Kollokatsiya tahlili	*Tez-tez birga uchraydigan so’zlarni aniqlaydi*	*Siyosiy nutqlarda takrorlanadigan iboralarni topish*
So‘z turkumlarini belgilash	*Gap tuzilishini aniqlash*	*Bahsli matnlardagi Grammatik murakkablikni tahlil qilish*
Noma’lum birliklarni aniqlash-NER	*Ismlar, joy nomlari, hodisalarni belgilaydi*	*Yangiliklarda eslatiladigan subyektlarni tahlil qilish*
Sentiment-kayfiyat tahlili	*Matnning ijobiy yoki salbiy kayfiyatini o‘lchaydi*	*Ijtimoiy tarmoqlardagi munosabatlarni tahlil qilish*
Mavzu modellash-LDA va b.	*Katta matnlarda asosiy mavzularni aniqlaydi*	*Bloglar va maqolalardagi mavzularni guruhlarga ajratish*

Korpus asosidagi diskurs tahlili mashina tarjimasi, suhbatbotlar, matnlarni avtomatik tahlil qilish va mazmuniy izlanishlarda katta rol o‘ynaydi. Masalan, sun’iy intellekt asosida matn yaratishda diskursning tuzilishi (sabab-oqibat, qarama-qarshilik va b.) tahlil qilinishi orqali matnning mantiqiyligi va uzviyligi ta’minlanadi. Korpus lingvistikasi diskursni aniqlashda ilmiy asos va aniq ma’lumotlar bilan ta’minlaydi. Kompyuter lingvistikasi vositalari yordamida bu tahlillar tez, aniq va samarali tarzda amalga oshiriladi. Natijada, nafaqat nazariy, balki amaliy til ishlatishda ham yangicha yondashuvlar yuzaga kelmoqda.

Diskurs kompetensiyasi atamasi lingvistik termin sifatida XX asrning ikkinchi yarmida tilshunoslikda keng qo‘llanila boshlandi. U birinchi navbatda

kommunikativ kompetensiya bilan bog‘lanadi, chunki diskurs kompetensiyasi kommunikatsiyaning samarali tashkil etilishida muhim rol o‘ynaydi. Shu bilan birga, diskurs kompetensiyasi nafaqat grammatik bilimlar va tilning formal tuzilmalari haqida, balki tilning real kommunikativ vaziyatlarda qanday ishlashi, maqsadga yo‘naltirilgan va kontekstual asosda qanday ifodalanishi haqida ham gapiradi.Diskurs kompetensiyasini ta'riflashda ko‘plab ta'riflar mavjud, ammo umumiy qarashga ko‘ra, bu tushuncha shaxsning o‘zaro muloqotda tilni nafaqat grammatik jihatdan to‘g‘ri, balki mantiqan izchil va to‘g‘ri kontekstda ishlatish malakasini anglatadi. Boshqacha aytganda, diskurs kompetensiyasi – bu til vositalarini faqat to‘g‘ri qo‘llash emas, balki maqsadli va samarali tarzda ulardan foydalanish qobiliyatidir. Bu, o‘z navbatida, shaxsning tilni faqat akademik nuqtai nazardan bilishini emas, balki real hayotdagi muloqotda ham foydali qo‘llashini o‘z ichiga oladi.

Tilshunoslikda diskurs atamasi ko‘plab o‘zgarmalar, muloqotdagi vaziyatlar va ko‘plab kommunikativ ko‘nikmalarni o‘z ichiga olgan keng ma'noga ega. Diskurs kompetensiyasi shunday til ko‘nikmasidirki, o‘rganuvchi nafaqat til tizimini, balki tilning o‘rganilgan normativlariga asoslangan holda, muloqot jarayonini to‘g‘ri boshqarish qobiliyatiga ega bo‘ladi. Ma'lumotlar va nutqning samarali shakllanishi, til o‘rganish va undan foydalanish ko‘nikmasi diskurs kompetensiyasining asosiy tarkibiy qismlaridan biridir. Diskurs kompetensiyasi faqatgina tilning grammatik va leksik jihatlarini qamrab olmaydi.Shuningdek, u kommunikativ muloqotda kontekstga mos ravishda tilni to‘g‘ri va samarali ishlatishni talab qiladi. Diskurs kompetensiyasining tarkibiy qismlarini ko‘rib chiqish zarur, chunki ular o‘quvchining va til foydalanuvchisining kommunikativ vaziyatlarda to‘g‘ri va mantiqan izchil nutq qurish imkoniyatini beradi.

Diskurs kompetensiyasining asosiy tarkibiy qismlaridan biri – ***izchillik*** (koherensiya)dir. Koherensiya matnning mazmun jihatdan bir butunlikka ega bo‘lishini ta'minlaydi, ya’ni fikrlar bir-biriga mantiqan bog‘langan va izchil bo‘lishi kerak. Bu jarayonni izchil va aniq tarzda ifodalash muloqotning samaradorligini oshiradi. Koherensiya nutqning to‘g‘ri strukturalanishi va mantiqiy ketma-ketligini tashkil etadi. Bundan tashqari, kohesiya muhim ahamiyatga ega. Kohesiya, o‘z navbatida, matndagi qismlar o‘rtasidagi grammatik va leksik bog‘liqlikni ta'minlaydi. Kohesiv elementlar – ulardan foydalanish, ya’ni zamon, shaxs, joy, va vaqt ko‘rsatkichlarini, bog‘lovchilarni va boshqa grammatik vositalarni to‘g‘ri ishlatish – muloqotda samarali va izchil nutqni tashkil etadi.

Natijalar. Diskurs kompetensiyasining yana bir muhim qismi – kommunikativ niyatni aniqlash va unga mos ravishda nutq qurish. Muloqotda ishtirok etayotgan shaxslar o‘z niyatlarini aniq ifodalashlari va suhbatdoshlarning niyatini tushunishlari lozim. Nutq strategiyalari orqali o‘rganuvchi fikrni qanday izchil ravishda ifodalash, qanday argumentlar bilan o‘z pozitsiyasini himoya

qilishni, va qanday qilib vaziyatga mos ravishda nutqni moslashtirishni o‘rganadi. Madaniy moslashuvchanlik ham diskurs kompetensiyasining tarkibiy qismi hisoblanadi. O‘quvchilar faqat tilni o‘rganish bilan cheklanmay, balki boshqa madaniyatlarga moslashish qobiliyatiga ega bo‘lishlari kerak. Muloqotda turli ijtimoiy va madaniy me’yorlarga moslashish, munosabatlarni to‘g‘ri qurish va til vositalaridan madaniy jihatdan mos tarzda foydalanish zarur. Bu, o‘z navbatida, global muloqotda tilni ishlatish ko‘nikmalarini rivojlantiradi. Diskurs kompetensiyasi va kommunikativ kompetensiya o‘rtasidagi bog‘liqlik to‘g‘ridan-to‘g‘ri aloqada bo‘lib, ular bir-birini to‘ldiradi va kuchaytiradi. Hymesning (1972) kommunikativ kompetensiya modelida til o‘rganish jarayoni faqat grammatik bilimlarni o‘z ichiga olmaydi, balki tilning real kommunikativ vaziyatlarda qanday qo‘llanishini ham o‘z ichiga oladi. Diskurs kompetensiyasi, aynan shu nuqtada, tilni kommunikativ jihatdan qanday to‘g‘ri qo‘llashni o‘rgatadi. Diskurs kompetensiyasi kommunikativ kompetensiyaning markaziy elementlaridan biri bo‘lib, nutqning ko‘plab funktsiyalarini bajaradi. Kommunikativ kompetensiyaning shakllanishi diskurs kompetensiyasiga bevosita bog‘liqdir. Agar kommunikativ kompetensiya shaxsning muloqotga tayyorligini bildirsa, diskurs kompetensiyasi bu tayyorgarlikni amalda qo‘llashga imkon yaratadi. Diskurs kompetensiyasi muloqot jarayonini boshqarishda, fikrlarni izchil ifodalashda, suhbatdoshlar bilan to‘g‘ri aloqada bo‘lishda samarali vosita sifatida ishlaydi.

Kommunikativ kompetensiyaning samarali shakllanishi uchun, tilni to‘g‘ri ishlatishdan tashqari, uni kontekstga moslashtira olish, madaniy va ijtimoiy normalarni inobatga olish zarur. Diskurs kompetensiyasi, shuningdek, suhbatdoshlarning ehtiyojlari va talablariga mos ravishda nutqni tashkil qilishni talab qiladi. Bu jarayonning muvaffaqiyatli bo‘lishi tilni faqat grammatik jihatdan bilish emas, balki real va dinamik vaziyatlarda qanday qo‘llashni anglatadi. Shu sababli, diskurs kompetensiyasi nafaqat tilshunoslik nuqtai nazaridan, balki muloqotdagi samaradorlikni oshirish uchun ham muhim hisoblanadi. Til o‘rganuvchilarni nafaqat tilning grammatik qoidalarini, balki uni ijtimoiy va madaniy kontekstlarda to‘g‘ri va maqsadga muvofiq ishlatishni o‘rgatish zarurdir. Bu esa o‘z navbatida, tilni mukammal o‘rganish va undan samarali foydalanish ko‘nikmalarini rivojlantiradi.Til o‘rgatuvchi metodikalar, shu jumladan kommunikativ va interaktiv yondashuvlar, o‘rganuvchilarga turli kontekstlarda, jumladan, ish joyida, oilada yoki akademik muhitda muloqot qilishda samarali bo‘lish uchun zarur bo‘lgan ko‘nikmalarni shakllantiradi. Diskurs kompetensiyasining bu jihati, tilni o‘rganish jarayonining eng muhim qismlaridan biri hisoblanadi, chunki u muloqot jarayonini oddiy so‘zlashuvdan tashqari, jamiyatning ijtimoiy va madaniy sharoitlarini tushunishga asoslanadi. Nutq turlari va muloqotning maqsadi har doim o‘ziga xos til vositalarini talab qiladi. Diskurs kompetensiyasi o‘rganuvchiga bu nutq turlarini ajrata olish va har bir tur uchun

to‘g‘ri til shakllarini tanlash imkonini beradi. Misol uchun, rasmiy va norasmiy nutq, ilmiy va og‘zaki muloqotning har biri o‘ziga xos strukturaga ega. Rasmiy nutqda, odatda, yanada aniq va izchil ifodalanish talab etiladi, bunda ortiqcha emotsional rang berish yoki noaniq iboralardan saqlanish zarur. Norasmiy nutqda esa, ko‘proq shaxsiy va erkin ifodalar qo‘llaniladi, bunda hissiy ohanglar va suhbatdosh bilan yanada bevosita muloqot o‘rnatish mumkin. Diskurs kompetensiyasi, o‘z navbatida, tilni o‘rganuvchiga ijtimoiy-madaniy faktorlarga moslashish ko‘nikmasini beradi. Misol uchun, ba'zi madaniyatlarda norasmiy nutqdagi hazil-mutoyiba yoki ta’kidlashlar ijtimoiy qabul qilinadi, boshqalarida esa bu qabul qilinmasligi mumkin. Diskurs kompetensiyasi yordamida til o‘rganuvchi, o‘z nutqini va ifodalarini har bir madaniy kontekstga mos ravishda o‘zgartirishni o‘rganadi. Bu esa nafaqat tilni to‘g‘ri ishlatish, balki jamiyatda to‘g‘ri va muvaffaqiyatli muloqot qilishni ham anglatadi. Tilni o‘rganishda ijtimoiy-madaniy omillarni hisobga olish, muloqotning barcha jihatlarini to‘liq tushunishga yordam beradi. Ijtimoiy-madaniy moslashuv orqali o‘rganuvchilar, tilni faqat nazariy o‘rganish bilan cheklanmay, uni real hayotdagi kommunikativ vaziyatlarda samarali ishlatish imkoniyatini ham qo‘lga kiritadilar. Bu bo‘limda til o‘rganishda diskurs kompetensiyasining ahamiyatini keng yoritdim. Har bir mavzuning ta'rifi va muhim jihatlari ko‘rsatilgan, shuningdek, til o‘rgatishda qanday metodik yondashuvlar yordamida bu kompetensiyalarni rivojlantirish mumkinligi haqida fikrlar bildirildi.

Dialogik va interaktiv mashg‘ulotlar til o‘rgatishda diskurs kompetensiyasini shakllantirishda muhim ahamiyatga ega. Ushbu mashg‘ulotlar o‘quvchining tilni faqat o‘qish va yozish orqali emas, balki suhbatlashish, munozara qilish va o‘z fikrini bildirish orqali o‘rganishiga yordam beradi. Diskurs kompetensiyasi – bu tilni samarali ishlatish ko‘nikmasi bo‘lib, uni shakllantirish uchun o‘quvchiga real hayotdagi vaziyatlarga o‘xshash sharoitda muloqot qilish imkoniyatini berish zarur. Dialogik mashg‘ulotlar o‘quvchilarga o‘zaro fikr almashish, argumentatsiya qilish va turli nuqtai nazarlarni inobatga olish imkonini beradi. Shu bilan birga, interaktiv mashg‘ulotlar o‘quvchilarning muloqotda faol ishtirok etishlarini ta’minlab, ularni samarali nutq tuzish, muloqotda fikrni izchil va to‘g‘ri ifodalashga o‘rgatadi. Diskurs kompetensiyasini shakllantirishda bu yondashuvlar o‘quvchining tilni nafaqat grammatika va leksika asosida, balki real va dinamik kommunikatsion vaziyatlarda ishlatish qobiliyatini oshiradi.

Xulosa qilib aytadigan bo‘lsak, diskurs zamonaviy tilshunoslik va ta’lim metodikasida muhim ahamiyatga ega bo‘lgan fenomendir. Ushbu maqolada diskurs kompetensiyasining ta’lim jarayonidagi o‘rni va uning shakllantirish metodlari keng tahlil etildi. Maqolada til o‘rganishda diskurs kompetensiyasining asosiy komponentlari, uning kommunikativ kompetensiyada tutgan o‘rni, hamda turli metodlarning samaradorligi chuqur o‘rganildi. Bu kompetensiyani shakllantirish, o‘quvchilarning nafaqat grammatik va leksik bilimlarini, balki ijtimoiy-madaniy

kontekstda samarali muloqot qilish ko‘nikmalarini ham rivojlantiradi. Diskurs kompetensiyasi nafaqat tilni to‘g‘ri ishlatish, balki muloqotning turli vaziyatlarida aniq, izchil va maqsadga yo‘naltirilgan fikrlarni ifodalash qobiliyatini ham o‘z ichiga oladi. Kommunikativ yondashuv, dialogik va interaktiv mashg‘ulotlar, autentik materiallarning qo‘llanilishi bu jarayonni samarali tarzda rivojlantirishga yordam beradi. Shu bilan birga, madaniy kompetensiya va diskurs kompetensiyasining o‘zaro bog‘liqligi, til tafakkuri, nutq strategiyalari va interkultural muloqotdagi roli ham ma’lum qilindi. Ta’lim jarayonida diskurs kompetensiyasini shakllantirishda samarali metodlarning qo‘llanilishi, o‘quvchilarga tilni real hayotda ishlatish ko‘nikmalarini o‘rgatadi. Kommunikativ yondashuvning asosiy maqsadi – o‘quvchiga tilni nafaqat rasmiy muloqotda, balki ijtimoiy va madaniy kontekstda ham samarali ishlatishni o‘rgatishdir. Autentik materiallar, interaktiv mashg‘ulotlar va tilni turli kommunikativ vaziyatlarda qo‘llash imkoniyatlari o‘quvchilarni real dunyoga tayyorlashga yordam beradi. Diskurs kompetensiyasining rivojlanishi, o‘z navbatida, tilni o‘rgatish jarayonini yanada chuqurlashtiradi va o‘quvchilarni faqat til bilimiga ega bo‘lish bilan cheklanmay, balki tilni kommunikativ jarayonda to‘g‘ri va samarali ishlatishga tayyorlaydi. Bu borada o‘quv dasturlarining boyitilishi, diskursga asoslangan baholash tizimlarining joriy etilishi va o‘qituvchilar uchun metodik tavsiyalarning ishlab chiqilishi muhim ahamiyatga ega.

Shu bilan birga, diskurs kompetensiyasining rivojlanishi madaniy kompetensiyaning shakllanishi va interkultural muloqotdagi muvaffaqiyatli ishtirok etishni ta'minlaydi. O‘quvchilar tilni faqat texnik jihatdan emas, balki madaniy, ijtimoiy va semantik jihatlardan ham to‘g‘ri ishlatishni o‘rganadilar. Bu esa, o‘z navbatida, ularning muloqotda yanada samarali va aniq bo‘lishlariga yordam beradi.Shu sababli, ta’lim jarayonida diskurs kompetensiyasini shakllantirish o‘quvchilarni nafaqat tilni o‘rganishga, balki uni real hayotdagi muloqotda muvaffaqiyatli ishlatishga tayyorlashga qaratilgan muhim vazifadir. Bu jarayonni samarali tarzda amalga oshirish uchun metodik yondashuvlar, materiallar va o‘qituvchilar malakasini oshirish bo‘yicha tavsiyalarni ishlab chiqish zarur. Diskurs kompetensiyasining rivojlanishi, til o‘rganishda o‘quvchilarning muloqot ko‘nikmalarini yuqori darajada shakllantirishga, o‘z fikrlarini aniq va izchil tarzda ifodalashga imkon yaratadi. Shu bilan birga, bu jarayon madaniy kompetensiyaning shakllanishi va interkultural muloqotdagi muvaffaqiyatli ishtirok etishning muhim omili bo‘ladi.

Foydalanilgan adabiyotlar:

1. Abdunazarova N. “Diskurs”ni talqin qilishga turlicha yondashuvlar. Ta'lim fanlari bo‘yicha akademik tadqiqotlar. 2021, №. 4

2. Usmonova Sh. Til madaniyati. - T.: 2014 yil.

3. Matn va diskurs tushunchalari munosabati Sarbarova Dilnoza Nosirjon qizi, O‘zbekiston Milliy universiteti O‘zbek tilshunosligi kafedrasi o‘qituvchisi dilnozasarbarova1@gmail.com https://doi.org/ 10.5281/zenodo. 7998832

4. Saidova, M. U. (2020). Lexicographic and ethymological analysis of the concise oxford dictionary of literary terms by. https://namdu.research commons.org/cgi/viewcontent cgi?article= 2723&context=journal

5. Yunusova, A. (2023). Современные подходы к теории языковой личности и сущности дискурса. центр научных публикаций (buxdu.Uz),40(40).https://journal.buxdu.uz/index.php/journals_ buxdu/ article/view/10591.

6. Yunusova, A. (2023). Functional and pragmatic aspects of inversion. центр научных публикаций (buxdu.Uz), 30(30). https://journal.buxdu.uz /index.php/journals_ buxdu/article/view /9178.

7. Canale, M., & Swain, M. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics,1980. 1(1), 1-47.

8. Bachman, L. F. Fundamental considerations in language testing. Oxford University Press.1990.

9. Savignon, S. J. Interpreting Communicative Language Teaching. Yale University Press.2002.

10. Hymes, D. On communicative competence. In J.B. Pride & J. Holmes (Eds.), Sociolinguistics. Penguin Books.1972.

11. Celce-Murcia, M. Rethinking the role of communicative competence in language teaching. In E. Alcón Soler & M. P. Safont Jordà (Eds.), Intercultural Language Use and Language Learning.2007.

O‘ZBEK TILI MILLIY KORPUSI TUZILISHI: ASOSIY KOMPONENTLAR VA AXBOROT TURLARI

Primova Mastura Hakim qizi,
O‘qituvchi
primovamastura@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. O‘zbek tili milliy korpusi tilshunoslik, adabiyotshunoslik va sun’iy intellekt sohalarida qo‘llaniladigan murakkab tizim bo‘lib, uning tuzilishi matnlar to‘plami, annotatsiya qatlamlari va ma’lumotlar bazasi arxitekturasi asosida shakllantirilgan. Korpusning asosiy komponentlariga turli davrlar (zamonaviy, tarixiy), janrlar (badiiy, ilmiy, rasmiy) va dialektlardagi matnlar, shuningdek, ularga qo‘shilgan morfologik, sintaktik va semantik annotatsiyalar kiradi. Ushbu maqolada korpusning tuzilishi, asosiy komponentlari va undagi axborot turlari atroflicha o‘rganiladi. Tadqiqot korpus lingvistikasi usullariga asoslanib, matnlarning turli janrlari va manbalarini tahlil qiladi. Korpus matnlar to‘plami, morfologik, sintaktik va semantik annotatsiya, metama’lumotlar va qidiruv tizimi kabi asosiy qismlardan iborat ekani aniqlanadi. Natijalar shuni ko‘rsatadiki, korpus turli davrlar va janrlarga oid matnlarni qamrab, tilshunoslik tadqiqotlari va sun’iy intellekt loyihalari uchun katta imkoniyatlar beradi. Korpusning foydalanuvchilar uchun qulay interfeysi va qidiruv vositalari uning qo‘llanilishini yanada samarali qiladi. Maqolada korpusning kelajakda rivojlanishi, xususan, annotatsiya sifatini oshirish va yangi texnologiyalar bilan integratsiya qilish bo‘yicha takliflar beriladi. Ushbu tadqiqot o‘zbek tili milliy korpusining ahamiyatini ochib beradi va tilshunoslar hamda texnologlar uchun qimmatli resurs bo‘lib xizmat qiladi.

Kalit so‘zlar: *O‘zbek tili milliy korpusi, korpus lingvistikasi, leksik-grammatik teglash, korpus platformalari, kollokatsiya, KWIC (Key Word in Context), lingvistik teglash, metama’lumotlar (metadata).*

Kirish

Til korpusi – bu ma’lum bir tilga oid yozma (va nutq) matnlarning elektron shaklda jamlangan ulkan to‘plami bo‘lib, u foydalanuvchi uchun maxsus qidiruv tizimi vazifasini ham o‘taydi[1]. Ya’ni, korpus – raqamlashtirilgan matnlarni lingvistik jihatdan teglangan (annotatsiyalangan) holda saqlovchi va ulardan turli xil axborot olish imkonini beruvchi kompleks tizimdir. Milliy korpuslar odatda o‘z ichiga millionlab so‘zlarni qamrab olgan matnlar majmuini oladi; masalan, Britaniya milliy korpusi (BNC) taxminan 100 million so‘zli ingliz tilining turli janrdagi matnlarini o‘z ichiga oladi, Rus milliy korpusi (RNC) esa 1 milliarddan ortiq so‘z shakllarini jamlagan holda doimiy ravishda kengayib borayapti. Korpus **tarkibidagi matnlar turli manbalardan**, jumladan **badiiy adabiyotlar, ilmiy maqolalar,**

ommaviy axborot vositalari materiallari va **og‘zaki nutq yozuvlari**dan tanlab to‘planadi [2]. Elektron korpus oddiy matn arxivigina emas, balki undagi har bir so‘zni va gapni lingvistik jihatdan tahlillab belgilovchi hamda foydalanuvchi so‘rovlariga tezkor javob beruvchi axborot tizimidir.

Korpus tuzilmasining darajalari. Har qanday zamonaviy til korpusi bir necha darajadagi asosiy komponentlar yig‘indisidan tashkil topadi:

1) **matn bazasi va lingvistik teglash** – korpusdagi barcha matnlarning struktura va til birligidagi belgilarini o‘z ichiga oladi;
2) **texnik arxitektura** – ma’lumotlarni saqlash, qayta ishlash va qidiruvni ta’minlovchi dasturiy-apparat tizimlarini qamrab oladi;
3) **foydalanuvchi interfeysi** – korpus bilan ishlash uchun qulay grafikali yoki veb-ishchi muhit bo‘lib, unda qidiruv, filtr, statistik chiqish kabi funksiyalar mujassam.

Quyida ushbu har bir komponent batafsil yoritiladi, shuningdek, korpuslardan olinadigan axborot turlari – lingvistik belgilash, metama’lumotlar va statistik ma’lumotlar alohida tahlil qilinadi.

Matn bazasi va lingvistik komponentlar

Matnlar va ularning tuzilishi. Korpusning asosini turli manbalardan to‘plangan matnlar bazasi tashkil qiladi. Korpusga kiritilayotgan har bir matn odatda oldindan raqamli ko‘rinishga keltiriladi va maxsus formatga solinadi (masalan, UTF-8 kodida matn). Matn ichidagi birliklar eng kichik **tokenlarga ajratiladi** – token bu so‘z, son, tinish belgisi kabi ajratib olinuvchi minimal elementdir. Tokenlarga ajratish (tokenizatsiya, tokenlash) natijasida **matn gaplarga** va **gaplar so‘zlarga** bo‘linadi. Har bir so‘z tokeniga lingvistik xususiyatlar bog‘lanadi: **uning lemmasi** (lug‘aviy asosi yoki boshlang‘ich shakli) **aniqlanadi** va **so‘z turkumi, grammatik kategoriya** kabi teglar bilan belgilab chiqiladi. Masalan, Rus milliy korpusida 1 milliarddan ziyod so‘z shakllari avtomatik lemma va morfologik teglar bilan belgilangan bo‘lib, har bir so‘zning barcha mumkin bo‘lgan grammatik tahlillari unga biriktirilgan. Shu tariqa, korpus matnlari ichida qidiruv faqat xom tekst bo‘yicha emas, balki lemmatlangan va teglangan shaklda amalga oshiriladi – foydalanuvchi so‘zning aynan qaysi grammatik turkumda kelishini ham tanlab izlashi mumkin.

Leksik-grammatik teglash. Korpus tarkibidagi so‘z tokenlariga biriktiriladigan lingvistik teglash (annotatsiya) bir necha qatlamdan iborat bo‘lishi mumkin. Eng asosiy qatlam – **morfologik teglash**, ya’ni har bir so‘zning qaysi so‘z turkumi (masalan, **ot, fe’l, sifatl– yoki part-of-speech, POS**) ekanini ko‘rsatish va uning grammatik grammemalarini (kelishik, zamon kabi kategoriyalarni) belgilashdir. Korpus yaratuvchilari til strukturasi murakkabliklarini e’tiborga olgan

holda bunday teglar tizimini ishlab chiqadilar. Masalan, RNC misolida har bir so‘z shakli uchun uning lemmasi va so‘z turkumi, so‘ngra so‘z turkumiga xos doimiy grammatik xususiyatlar va ayni o‘sha kontekstdagi o‘zgaruvchi grammatik shakllari ko‘rsatiladi [3]. Korpusning teglangan qismida shu to‘liq morfologik tahlil foydalanuvchiga ko‘rsatiladi, qolgan qismida esa faqat lemma va so‘z turkumi aks ettiriladi.

Sintaktik struktura. Lingvistik teglashning keyingi darajasi – sintaktik annotatsiya bo‘lib, ba’zi korpuslarda har bir **gapning tuzilishi (parsing)** ham saqlanadi. Bu sintaktik bog‘lanishlar ko‘rinishida (odatda, daraxt struktura shaklida) ifodalanadi: gap tarkibidagi so‘zlar orasidagi sintaktik munosabatlar (boshlovchi, to‘ldiruvchi, aniqlovchi va hokazo) **maxsus teglar** bilan bog‘langan graflar tarzida beriladi. Masalan, Rus milliy korpusining **SynTagRus** deb nomlangan maxsus sintaktik subkorpusi mavjud bo‘lib, u 1 million atrofidagi gaplarga to‘liq sintaktik daraxt tuzilmalarini biriktirgan. Bunday daraxtbank (treebank) korpuslar til sintaksisini o‘rganish va mashinada tushunishni rivojlantirishda muhim manba hisoblanadi. Ingliz tilida **Penn Treebank**, nemis tilida **TIGER Corpus** kabi mashhur sintaktik teglangan korpuslar mavjud. Sintaktik teglash tufayli foydalanuvchi korpusdan nafaqat alohida so‘zlarni, balki ma’lum bir **sintaktik konstruksiyalarni** ham izlab topishi mumkin bo‘ladi (masalan, **“ot+fe’l+ot”** tuzilmasidagi gaplar va hokazo).

Agglyutinativ tillarda belgilash muammosi. Ta’kidlash joizki, turkiy agglyutinativ tillarda (so‘zga ko‘plab qo‘shimchalar ketma-ket qo‘shiluvchi tillarda) korpus uchun morfologik analiz va so‘zni asos (lemma)ga keltirish masalasi murakkabroq bo‘ladi. O‘zbek tilini ham qamrab oluvchi tadqiqotlarda aynan POS-teglash va stemming (so‘zdan qo‘shimchalarni ajratib, negiz shaklini topish) muammolari dolzarb ekani qayd etilgan. Jumladan, B. Elov, Sh. Hamroyeva va boshqalar o‘zlarining ilmiy ishlarida agglyutinativ tillarda so‘zning grammatik kategoriyasini avtomatik aniqlash qiyinligi, ba’zan qo‘shimchalar tufayli ko‘p ma’noli holatlar paydo bo‘lishi va ularni bartaraf etish uchun maxsus algoritmlar zarurligini ta’kidlaydilar [4]. Shu bois, zamonaviy o‘zbek korpuslarini yaratishda **morfologik analizatorlar** va **lemmatizatorlar** ishlab chiqish alohida e’tibor talab qilmoqda. Masalan, o‘zbek tili uchun yaratilgan **https://uznatcorpara.uz/** kabi onlayn morfologik analizatorlar korpusdagi so‘zlarni tahlil qilishga xizmat qilmoqda.

Texnik arxitektura va platforma tizimi

Arxitektura ahamiyati. Korpusning muvaffaqiyatli ishlashi uchun uning ichki texnik arxitekturasi juda muhimdir. Faqatgina matn hajmini oshirish bilan cheklanib bo‘lmaydi – *agar korpus arxitekturasi zaif bo‘lsa* yoki u *lingvistik jihatdan teglanmagan* bo‘lsa, foydasi cheklangan bo‘lib qoladi. Mark Deyvis

(COCA korpusi asoschisi) ta’kidlaganidek, korpus hajmi **“qancha katta bo‘lsa shuncha yaxshi”** degan qoidaga faqat to‘g‘ri arxitektura va to‘liq annotatsiya bilan erishish mumkin; aks holda 100 million so‘zli yaxshi belgilangan korpus 1 milliard so‘zli ammo belgilanishi sust korpusdan afzalroq natija berishi mumkin [5]. **Optimal arxitektura** matnlarni samarali indekslash, tezkor qidiruv va katta hajmda ham barqaror ishlashni ta’minlaydi.

Ma’lumotlar bazasi tuzilishi. Korpus ichki tizimi ko‘pincha maxsus ma’lumotlar bazasi yoki indekslash usullariga tayangan holda quriladi. Korpusdagi har bir so‘z alohida yozuv (record) sifatida saqlanishi, unga lemma va teglar bog‘lanishi mumkin. Zamonaviy yirik korpuslarda tezkor qidiruv uchun ko‘pincha **relatsion ma’lumotlar bazasi arxitekturasi qo‘llaniladi** (masalan, COCA va boshqa English-Corpora.org korpuslari shu usulda). Bunda har bir so‘z maxsus sonli identifikator (ID) bilan bazaga kiritiladi, so‘zlarning matndagi ketma-ketligi indekslar orqali saqlanadi. Masalan, COCA korpusida so‘zlar raqamlar ko‘rinishida *“corpus”* jadvalida saqlanib, alohida *“lexicon”* jadvalida IDga mos so‘zning o‘zi, lemmasi va POS-tegi saqlanadi; shu tariqa massiv matnlarda ham ma’lumotga murojaat juda tez bajariladi [6]. Bunday arxitektura yordamida foydalanuvchi milliardlab so‘zli korpuslarda ham bir necha soniya ichida murakkab so‘rovlarni bajarishi mumkin – masalan, berilgan so‘zdan oldin keluvchi eng ko‘p 100 sifatni butun korpus bo‘ylab sanash kabi og‘ir so‘rovlar 1-2 soniyada natija beradi. Shuningdek, **klasterlangan indekslar** va **ma’lumotlarni ixtiyoriy kontekst bo‘yicha saqlash** kabi usullar qidiruvni optimallashtiradi. Korpus arxitekturasi to‘g‘ri tashkil etilgan bo‘lsa, hatto maxsus korpus dasturlari (masalan, Sketch Engine)dan ham tezroq natija olish mumkinligi tajribada ko‘rsatilgan.

Korpus platformalari. Bugungi kunda korpus tuzish uchun tayyor dasturiy platformalar va tizimlar mavjud. Masalan, jahon tajribasida **Sketch Engine, CQPweb** kabi platformalar ko‘plab korpuslar uchun universal yechim sifatida qo‘llanadi – ularda **matnlarni kiritish, avtomatik teglash** va **qidiruv interfeysi** birgalikda ta’minlanadi. Ayrim milliy korpuslar esa o‘ziga mos maxsus platformada yaratilgan: masalan, RNC uchun Maxsus qidiruv dasturi (Yandex bilan hamkorlikda) ishlab chiqilgan bo‘lib, u korpusning morfologik va sintaktik qidiruvlarini ta’minlaydi [4]. O‘zbek tilining milliy korpusi (https://uzschoolcorpara.uz/) ham maxsus veb-platforma sifatida ishga tushirilgan: unda matn bazasi serverda SQLServer ma’lumotlar bazasida saqlanadi, foydalanuvchi so‘rovlari Python yordamida qayta ishlanib, natijalar sahifada ko‘rsatiladi (loyiha tafsilotlarida shu keltirilgan). Korpus platformasining muhim qismi – bu lingvistik ishlov berish modullaridir: masalan, so‘zlarni avtomatik teglash uchun morfologik analizator yoki maxsus modellar ulanadi. RNC misolida Mystem va Dialing morfologik dasturlari integratsiya qilingan bo‘lib, ular matnlardagi so‘zlarning formasini aniqlashda ishlatiladi. O‘zbek tilida ham shunga o‘xshash

tarzda tayyor dasturiy modullar (https://uznatcorpara.uz/) korpusga ulanmoqda. Umuman, texnik arxitektura jozibadorligi shundaki, yaxshi optimizatsiya qilingan tizim katta hajmdagi korpuslarda ham foydalanuvchiga bir zumda javob qaytara oladi va yangi matnlarni qo‘shish ham oson kechadi.

Foydalanuvchi interfeysi va qidiruv imkoniyatlari

Interfeys dizayni va funksional. Korpusdan samarali foydalanish uchun uning foydalanuvchi interfeysi qulay va funksional bo‘lishi lozim. Interfeys foydalanuvchiga oson tushunarli bo‘lishi, turli ehtiyojlarga mos qidiruv turlarini taklif qilishi zarur. Odatda korpusning veb-interfeysi orqali quyidagi asosiy imkoniyatlar taqdim etiladi:

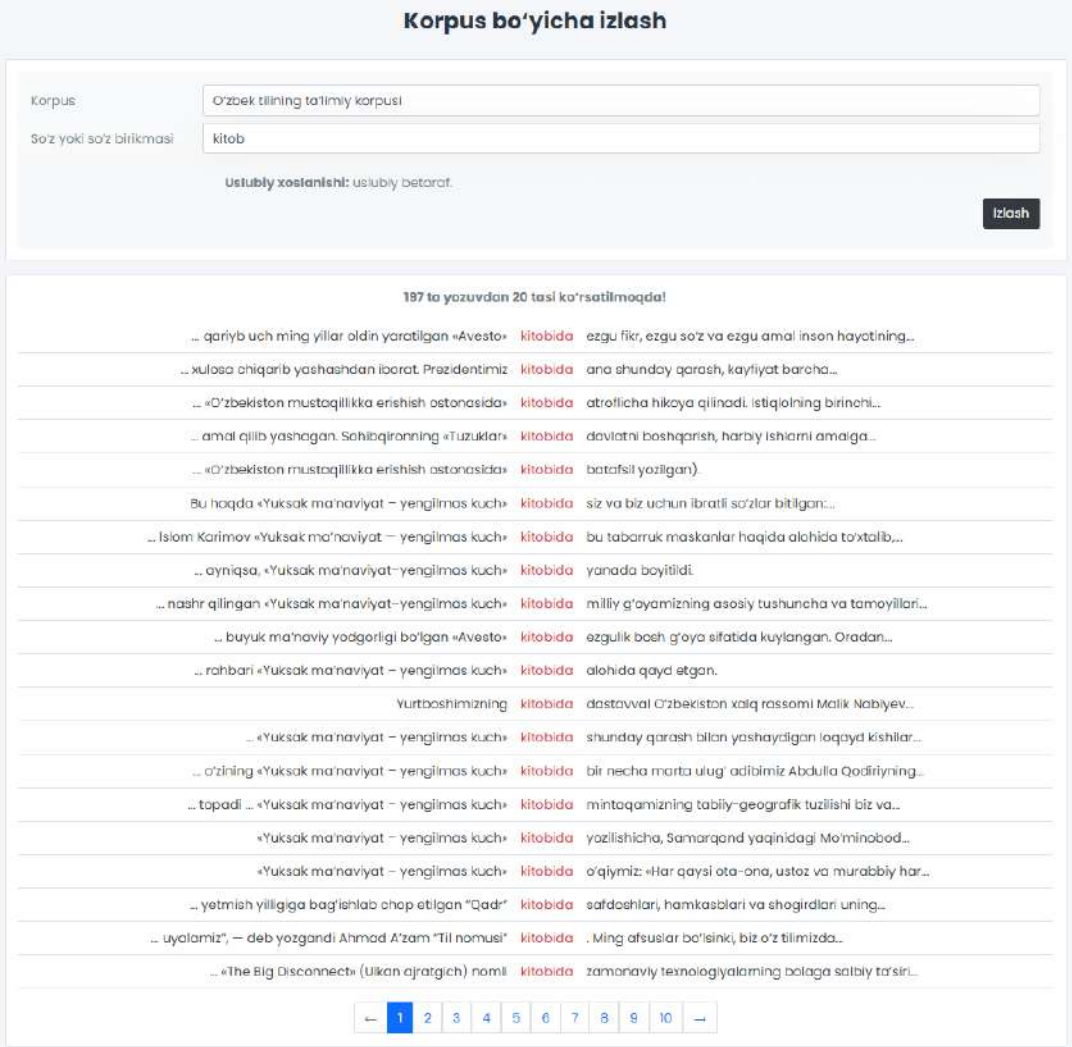
1) **Oddiy qidiruv** – matndan ma’lum bir so‘z yoki iborani izlash (bu Google qidiruviga o‘xshash, lekin korpus ichida amalga oshiriladi). Foydalanuvchi qidiruv satriga so‘z kiritadi va korpusdan ushbu so‘z uchragan barcha kontekstlar olinadi.

2) **Kengaytirilgan qidiruv** – foydalanuvchi lemma bo‘yicha (ya’ni so‘zning barcha shakllarini qamrab) yoki ma’lum so‘z turkumi bo‘yicha so‘rov berishi mumkin. Masalan, faqat fe’l lemmasini qidirish, yoki *“kitob_Noun”* kabi so‘z turini ko‘rsatib izlash imkoniyati. Bunday qidiruv POS-teglar yordamida amalga oshadi. RNC interfeysida maxsus *“Лексико-грамматический поиск”* degan bo‘limda foydalanuvchi so‘z yonida grammatik xususiyatlarni belgilashi mumkin – masalan, *“OT, birlik, Qaratqich kelishigi”* kabi.

3) **Murakkab qidiruv** – bir nechta so‘zlarning o‘zaro yaqinligini yoki bir gap ichida birga kelishini qidirish. Masalan, *“A va B so‘zlari bir gapda masofasi 5 tokendan oshmaydigan holda uchrasin”* kabi murakkab shartlar qo‘yilishi mumkin. Bunday qidiruvlar **CQL (Corpus Query Language)** yoki grafik forma orqali qo‘llab-quvvatlanadi.

4) **Filtrlash** – bu qidiruv natijalarini yoki qidiruvning o‘zini ma’lum metama’lumotlarga ko‘ra cheklashdir. Masalan, foydalanuvchi *faqat badiiy adabiyot janridagi matnlardan qidirmoqchi* bo‘lsa, kerakli filtrni tanlaydi. Yoki *ma’lum yillar oralig‘idagi matnlardan izlash*, yoki *muallifi ayol bo‘lgan matnlar* doirasidagina qidirish kabi. Jahon korpuslarida bunday filtrlash oddiy *“bir necha bosqichda”* amalga oshadi: masalan, COCA interfeysida foydalanuvchi **janr, yil, manba bo‘yicha saralab olish** orqali o‘ziga xos virtual kichik korpus yaratishi mumkin. Inglizcorpora saytida bir nechta belgi tanlash bilan, masalan, *“2000-yillarning jurnallari ichida dasturlash mavzusidagi matnlar”* kabi tanlangan bo‘limni bir necha soniyada shakllantirish imkoni borligi ko‘rsatilgan. O‘zbek tilining milliy korpusida ham matnlar **uslub (janr) bo‘yicha (ilmiy, badiiy, publitsistik, rasmiy, so‘zlashuv), yaratilgan vaqti bo‘yicha (sovet davri, mustaqillik davri kabi)** filtrlanishi nazarda tutilgan.

Natijalarni chiqarish shakllari. Korpus interfeysi qidiruv natijalarini foydalanuvchiga qulay tarzda namoyish qilishi ham muhimdir. Odatda **natijalar konkordans ko‘rinishida taqdim etiladi** [7]. **Konkordans** – bu kalit so‘zning atrofidagi kontekst bilan bir qatorda bir necha qatorda ko‘rsatilishidir. Masalan, foydalanuvchi “ilm” so‘zini izlasa, korpus dasturi ushbu so‘z uchragan har bir gapni alohida qator qilib, so‘zning o‘zi markazda ko‘rinadigan qilib chiqaradi. Bu usul **KWIC (Key Word in Context)** deb ham ataladi. Konkordans foydalanuvchiga qidirilgan so‘zning turli kontekstlarda qanday ishlatilishini tezda ko‘zdan kechirishga yordam beradi – muhim lug‘aviy va grammatik namunaklarni aniqlash osonlashadi. Misol uchun, 1.2.1-rasmda keltirilgan konkordansda *“day by day”* iborasi har xil misollarda keltirilgan bo‘lib, har bir qator korpusning boshqa bir hujjatidan olingan (bir qator ikkinchi qator bilan bevosita bog‘liq matn emas).

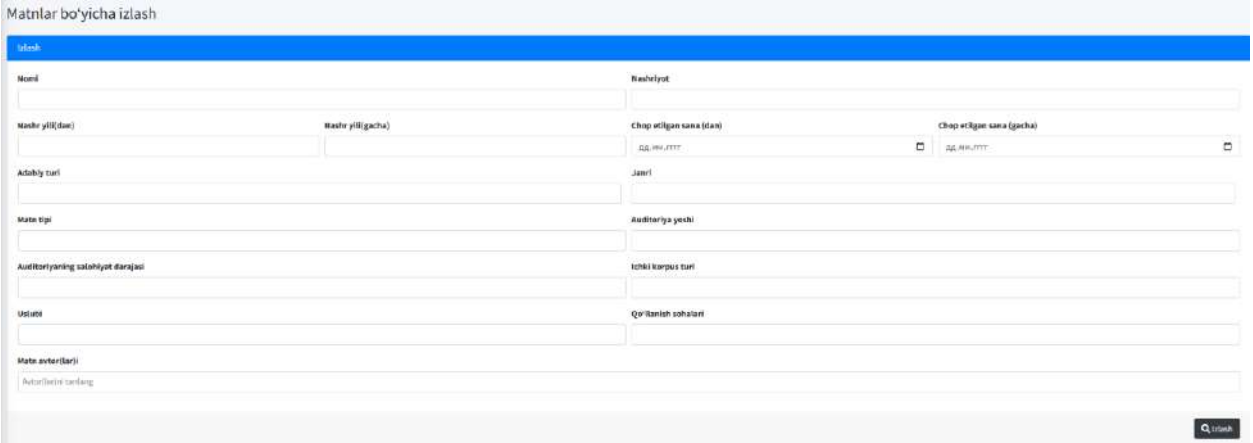


1-rasm. KWIC (Key Word in Context)

Korpus interfeysi odatda bir vaqtning o‘zida **30–50 ta** konkordans qatordan natijani ko‘rsatadi, foydalanuvchi esa keyingi sahifalarga o‘tib barcha topilmalarni

ko‘rib chiqishi mumkin. Bundan tashqari, har bir konkordans qator yonida matn manbasi ko‘rsatiladi (masalan, “Xalq so‘zi gazetasi, 2018 yil”) va u ustiga bosib to‘liq matnni ko‘rish imkoni ham bo‘lishi mumkin.

Statistik funksiyalar. Korpus foydalanuvchilari ko‘pincha individual misollar bilan cheklanmasdan, matnlar bo‘yicha statistik ma’lumotlarni olishga ham qiziqadilar. Shuning uchun zamonaviy korpus interfeyslari **so‘z chastotasi, kollokatsiyalar, ngramlar** kabi vositalarni ham taqdim etadi. Misol uchun, COCA korpusida foydalanuvchi istalgan so‘zni qidirganda, uning umumiy chastotasi (korpusda necha marta uchragani) hamda janrlar kesimidagi chastotasi (masalan, og‘zaki nutqda 500 marta, ilmiy matnlarda 200 marta kabi) grafigi bilan ko‘rsatiladi.



2-rasm. Korpus qidiruv tizimi

Bu bilan so‘zning qaysi uslubda ko‘proq ishlatilishi yoki vaqt o‘tishi bilan chastotadagi o‘zgarishi tahlil qilinadi. **Kollokatsiya** – tanlangan so‘zning yon-atrofida eng ko‘p uchraydigan so‘z birikmalaridir; masalan, korpus tahlili natijasida *“ilm”* so‘zi bilan eng ko‘p birga keluvchi so‘zlar *“fani”, “sohasida”, “arbobi”* kabi kollokatsiyalar ekani aniqlanishi mumkin. Korpus dasturlari buni aniqlash uchun statistik o‘lchovlardan (MI – o‘zaro axborot miqdori yoki t-skori kabi) foydalanadi va foydalanuvchiga Top-20 eng muhim kollokatsiyalar ro‘yxatini bera oladi. Masalan, COCA korpusida biror so‘z uchun Collocates bo‘limi mavjud bo‘lib, u so‘zning chap yoki o‘ng tomonida berilgan masofa ichida eng ko‘p uchragan so‘zlarni va ularning ko‘rsatkichlarini chiqaradi. Yuqorida aytib o‘tilganidek, maxsus **so‘rov orqali ham kollokatsiyani topish mumkin:** masalan, SQL so‘rov tili misolida “kitob” so‘zidan oldin keluvchi sifatlar topilib, ular orasidan 100 tasi saralanishi misoli keltirilgan edi – amalda foydalanuvchi buni oddiy interfeys orqali bir necha belgi tanlab bajaradi, tizim esa ortda bunday kodni ishlatadi.

N-grammalar esa korpusdagi so‘zlarning eng ko‘p uchraydigan ketma-ketliklaridir (N so‘zdan iborat frazalar). Korpusdan, masalan, eng ko‘p qo‘llanilgan **ikki so‘zli birikmalar (bigram)** yoki **uch so‘zli birikmalar (trigram)** avtomatik

chiqarib olinishi mumkin. Bu tilning *iboralari, maqol* va *iqtiboslarini* aniqlashda foydali. Masalan, o‘zbek tilining ilk milliy korpusi matnlari asosida n-gram usuli orqali til modelini qurish tajribasi amalga oshirilgan bo‘lib, bunda korpusdagi eng ko‘p uchraydigan so‘z ketma-ketliklari aniqlangan. Natijada, masalan, o‘zbek tilida eng ko‘p takrorlanadigan uch so‘zli iboralar (*“aloqa vositalari orqali”*, *“Prezidenti matbuot xizmati ma’lum qildi” singari*) ajratib olinishi mumkin – bu ma’lumotlar til xususiyatlarini o‘rganishda yoki mashina tarjimasida qo‘l keladi[8].

Yuqoridagi funksiyalar barchasi korpus interfeysi orqali birlashtirilgan holda taqdim etiladi [9]. Yaxshi interfeys oddiy foydalanuvchilar uchun sodda qidiruvni, lingvistlar uchun esa murakkab tahliliy so‘rovlarni amalga oshirishga imkon yaratadi. Misol tariqasida, Britaniya milliy korpusining avvalgi BNCweb interfeysi ham, hozirgi Sketch Engine dagi versiyasi ham foydalanuvchiga yuqorida sanalgan barcha imkoniyatlarni beradi – **so‘zlarni regex (muntazam ifodalar) bo‘yicha qidirishdan tortib**, matnlarni metama’lumotlar bo‘yicha ajratish va hatto tezavror (thesaurus) usulida o‘xshash so‘zlarni topishgacha. Rus milliy korpusining onlayn tizimi esa rus tilining o‘ziga xos grammatikalarini (masalan, fe’lning vid kategoriyasi, otning irglik/nisbat kategoriyasi) ham qidiruv shartlari sifatida qo‘shishga ixtisoslashgan. O‘zbek tilining elektron korpusi uchun ham shunday imkoniyatlarni bosqichma-bosqich joriy etish rejalangan – xususan, UzbCorpus.uz saytining 2020 yilgi ilk versiyasida token, lemma bo‘yicha oddiy qidiruvlar ishga tushirilgan bo‘lsa, keyinchalik kengaytirilgan qidiruv turlari ham qo‘shib borilmoqda.

Korpusdagi axborot turlari

Korpus tizimida turli axborot turlari qatlamlari mavjud bo‘lib, ularni uch asosiy guruhga ajratish mumkin:

1) **lingvistik axborot** – matnlarning lingvistik tavsifi va teglanishi (morfologik, sintaktik va boshqalar);

2) **metama’lumotlar** – har bir matnning tavsiflovchi ma’lumotlari (yaratilgan vaqti, janri, muallifi, manbasi, tili va hokazo);

3) **statistik axborot** – korpusdan foydalanuvchi so‘rov natijasida olinadigan chastotaviy yoki taqqoslovchi ko‘rsatkichlar (so‘zlar chastotasi, konkordans natijalari, kollokatsiyalar, n-gramma ro‘yxatlari va boshqalar).

Yuqorida biz aslida bu turdagi axborotlarning barchasiga to‘xtaldik – quyida ularni yana tizimli ravishda sharhlaymiz.

Lingvistik teglash qatlami

Lingvistik axborot korpus matnlarining annotatsiya qilingan qatlamidir. Uning tarkibiga yuqorida ta’riflangan **morfologik teglar, sintaktik strukturalar,**

shuningdek, ba’zan **semantik teglash** ham kiradi. Lingvistik teglash tufayli korpus *“oddiy matnlar to‘plami”*dan haqiqiy **til modellari bazasiga** aylanadi – chunki har bir so‘z va gap haqida mashina o‘qiy oladigan boy ma’lumotlar biriktirilgan bo‘ladi[10]. Morfologik teglashning ahamiyati shundaki, u *korpusdan ma’lum grammatik shakllarni izlash*, *grammatik hodisalar tezligini aniqlash* kabi ilmiy ishlarni amalga oshirishga imkon beradi. Masalan, lingvist olim ingliz tilida fe’lning zamon shakllari qo‘llanish dinamikasini o‘rganmoqchi bo‘lsa, to‘g‘ridan-to‘g‘ri korpusdan *“Past Perfect”* bo‘lib kelgan fe’llarni yillar kesimida sanab bera oladi – buning uchun korpusdagi har bir fe’lning zamon kategoriyasi belgilangan bo‘lishi kifoya. Rus milliy korpusida hatto so‘zlarning **semantik xususiyatlari** ham qisman belgilangan: masalan, unda *“inson”* yoki *“hayvon”* semantik sinfiga mansub otlar alohida belgiga ega, yoki **onomastik** (*shaxs ismi, joy nomi*) otlar alohida ajratilgan. Bunday chuqur belgilash qatlami keyinchalik korpusdan, masalan, *“hayvonlarni ifodalovchi otlar ishtirok etgan frazeologizmlar”* kabi nozik so‘rovlarni amalga oshirishda asqotadi.

O‘zbek tilining elektron korpuslari hozircha to‘liq lingvistik teglashga ega boshlang‘ich bosqichda. B.Elov va boshq. (2024) ishlanmalarida o‘zbek matnlarini avtomatik teglash va tahlil qilish ustida ishlar olib borilmoqda [11]. Jumladan, yaratilgan korpus imkoniyatlarini kengaytirish uchun xorijiy tajribadan foydalanilayotgani, xususan, og‘zaki matnlar korpusini ham yaratish va uni teglash masalalari o‘rganilayotgani qayd etiladi. Bu shuni ko‘rsatadiki, yaqin yillarda o‘zbek korpuslarida ham morfologik belgilash bilan birga sintaktik va semantik annotatsiya qatlamlari paydo bo‘lishi kutilmoqda. Shu tariqa, lingvistik axborot qatlami korpusning *“yuragi”* bo‘lib, u orqali korpus ilmiy-tadqiqotlarda qo‘llanadi va tilning raqamli modelini yaratadi. Aynan lingvistik belgilash tufayli korpus **milliy tilni chuqur o‘rganishda beqiyos manba bo‘la oladi** – bu haqda mahalliy tadqiqotchilar (B. Elov, Sh. Hamroyeva va boshq.) ham o‘z ishlarida ta’kidlaydi.

Metama’lumotlar (metadata)

Har bir korpus matni qandaydir manbadan olingan bo‘ladi va uning o‘ziga xos metama’lumotlari mavjud. **Metama’lumot** – bu matn haqidagi ma’lumot. Korpus dizaynida matnlarni tanlash va tasniflash juda muhim bo‘lgani sababli, har bir kiritilayotgan matnga tegishli metama’lumotlar biriktiriladi: **yaratilgan sanasi, janri, uslubi, mavzusi, muallifi, manbasi (masalan, qaysi jurnal yoki sayt), tili (agar original matn boshqa tilda bo‘lsa yoki o‘zbekcha lotin yozuvida yoxud kirill yozuvida bo‘lishi)** kabi ko‘plab tavsiflar ko‘rsatiladi. Bu ma’lumotlar korpus tuzishda ikki sababga ko‘ra muhim: birinchidan, **balanslangan korpus** tuzish uchun – ya’ni korpus tarkibida turli davr va turli janrlarning ulushi muvozanatli bo‘lishini nazorat qilish; ikkinchidan, keyinchalik foydalanuvchi bu metama’lumotlar orqali saralash va filtrlashni amalga oshirishi uchun. Misol tariqasida, BNC korpusi maxsus balanslash rejasi asosida tuzilgan bo‘lib, unda matnlar 1960–1990 yillardagi

ingliz tilini qamrab oluvchi 100 million so‘z hajmida yig‘ilgan, shundan 10% og‘zaki (nutq transkriptlari) va 90% yozma matnlar; yozma matnlar esa o‘z navbatida ilmiy, publitsistik, badiiy adabiyot, rasmiy hujjatlar kabi toifalarga bo‘lingan. Shu tarzda BNC ichida har bir matnga uning manbasi (masalan, “The Times gazetasi, 1991 yil, yangilik xabari” kabi) haqida batafsil metama’lumot bog‘langan. Rus milliy korpusida ham har bir matn uchun **bibliografik ma’lumotlar** berilgan: **asarning nomi, muallifi, yozilgan yili, janri, uslubi (badiiy, ilmiy va h.k.), matn turi (og‘zaki/yozma), hatto og‘zaki nutq bo‘lsa, gapiruvchining jinsi, yoshi, ijtimoiy holati** kabi qo‘shimcha ma’lumotlar kiritilgan. O‘zbek tilining milliy korpusi uchun B.Elov va boshq. (2023) tadqiqotlarida o‘zbek adabiy tilining besh uslubga mansub manbalari alohida jamlangani ta’kidlanadi (*so‘zlashuv uslubi, badiiy, ilmiy, rasmiy, publitsistik*) [12]. Bu shuni anglatadiki, har bir matnga qaysi uslubga tegishli ekanini ko‘rsatuvchi belgi qo‘yilgan. Kelgusida foydalanuvchi, masalan, faqat publitsistik uslubdagi matnlardan misollar izlashni xohlasa, *“janr: publitsistika”* filtrini tanlab qidirishi mumkin bo‘ladi.

Metama’lumotlarning yana bir foydasi – **korpus tarkibini tahlil qilish** va **taqqoslash** imkonini berishi. Masalan, korpusdan avtomatik ravishda *“badiiy adabiyot”* va *“ilmiy uslub”* bo‘yicha ikki alohida kichik korpus ajratib olib, ularning lug‘at tarkibini taqqoslash mumkin. Yoki vaqt bo‘yicha dinamik o‘zgarishni kuzatish uchun korpusni yillarga ajratish mumkin: masalan, COCA korpusida matnlar yillar va besh yilliklar bo‘yicha belgilangan, shuning uchun foydalanuvchi so‘zning 1990-yillardagi va 2010-yillardagi tezligidagi farqni oson ko‘ra oladi (grafigi orqali). Korpus interfeyslari metama’lumotlar bilan ishlashni intuitiv oson qilishga intiladi – misol uchun, English-Corpora (COCA) tizimida *“Virtual Corpus”* degan funksiya mavjud bo‘lib, foydalanuvchi bir necha soniya ichida tanlagan metama’lumotlar asosida shaxsiy kichik korpus tuzishi va unga tegishli so‘zlar statistikasini ajratib olishi mumkin. Xuddi shunday, Sketch Engine dasturida ham *“Subcorpus”* funksiyasi bor: foydalanuvchi masalan, muallifi Oybek bo‘lgan matnlar to‘plamini yoki 2018 yildan keyin chiqqan internet maqolalari to‘plamini ajratishi mumkin – buning uchun muallif va yil metama’lumotlarini filtrlaydi xolos, tizim uning uchun alohida korpus shakllantirib beradi.

Umuman olganda, **metama’lumotlar** korpusning *tasnifiy axboroti* bo‘lib, u korpus tuzilishini yanada boy qiladi. Chunki oddiy matnlar to‘plamidan farqli o‘laroq, korpusda matnlarning kelib chiqishi va xususiyatlari to‘g‘risidagi ma’lumotlar ham bazaga kiritiladi. Bu ma’lumotlar ilmiy tadqiqotlarda masalan, *stilistika, sotsiolingvistika, diatopik tilshunoslik (lahja va shevalar bo‘yicha)* tadqiqotlarni o‘tkazishga imkon yaratadi. Korpus tuzishda shu sababli metama’lumotlarni to‘plash va to‘g‘ri formatda kiritish alohida bosqich sifatida qaraladi.

Statistik axborot va foydalanuvchi ma’lumotlari

Statistik axborot – korpusdan foydalanuvchi olishi mumkin bo‘lgan turli tahliliy ma’lumotlar. Bular korpusga oldindan kiritilib qo‘yilgan emas, balki foydalanuvchi so‘roviga binoan **hisoblab chiqariladigan ko‘rsatkichlardir**. Ular qatoriga yuqorida tilga olingan **so‘zlar chastotasi, kollokatsion lug‘atlar, ngramlar** kiradi. Bundan tashqari, foydalanuvchi korpusdan o‘zi istalgan ikki holatni solishtirish natijasida statistik xulosa ham olishi mumkin – masalan, *“A matnlar korpusida so‘z X chastotasi B korpusidagidan 2 barobar ko‘p”* degan kabi ma’lumot.

Chastotali lug‘atlar – korpus asosida tuzilgan so‘zlarning ro‘yxatidir. Korpus hajmi katta bo‘lgani tufayli bunday ro‘yxatlar tilning eng keng qo‘llaniladigan so‘zlarini aniqlashda muhim manba bo‘ladi. Masalan, BNC korpusi asosida tuzilgan BNC1000 degan ro‘yxatda ingliz tilidagi eng ko‘p ishlatiladigan 1000 ta so‘z keltirilgan va u til o‘rganuvchilar uchun qo‘llanma sifatida mashhur. Korpusdan chastotani olish odatda dasturiy ravishda “so‘zlar ustida oddiy sanaq” ishidir, lekin katta hajmda bu oson emas – masalan, 1 milliard so‘zli korpusda turli so‘z shakllari soni (types) yuz minglarni tashkil qiladi. Shunga qaramay, Leipzig korpus to‘plami kabi loyihalar dunyoning 20 dan ortiq tili uchun avtomatik ravishda 1 million so‘zdan iborat korpuslar tuzib, ularning **chastotali lug‘atlarini** ham e’lon qilgan [12]. Ularning maqsadi resursi kam tillar uchun ham asosiy korpus statistikalarini ochiq taqdim etishdir. O‘zbek tilida ham ilk bor shunday chastotali lug‘at B.Elov tomonidan tuzilgan bo‘lib, https://uzschoolcorpara.uz/ uchun jami so‘zlar chastotasi hisoblab chiqilgan (natijada o‘zbek tilida eng ko‘p qo‘llaniladigan so‘z **“va”** ekanligi kabi qiziqarli faktlar aniqlandi). Chastota axboroti lingvistika tadqiqotlarida muhim: masalan, **Zipf qonunini tekshirish**, **so‘z boyligini baholash**, **leksikografik baza tuzish** kabi ishlarda korpusdan olingan chastotalar asos bo‘ladi.

Kollokatsiyalar va n-gramlar – bular korpus matnlaridan hosil qilinadigan **birikmalar statistikasidir**. Kollokatsiya ko‘pincha ikki yoki undan ortiq so‘zning bir-biriga yaqin yoki bir gap ichida kutilganidan ko‘ra ko‘proq birga uchrashini bildiradi. Korpus buni aniqlash uchun eng yaxshi manbadir: masalan, korpus tahlili shuni ko‘rsatishi mumkinki, *“ijtimoiy”* so‘zi eng ko‘p *“tarmoqlar”* so‘zi bilan yonma-yon qo‘llanilar ekan (natijada “ijtimoiy tarmoqlar” birikmasi kollokatsion bog‘lanish kuchiga ega fraza deyish mumkin). Bunday kollokatsiyalarni korpusdan chiqarib olish uchun dastur har bir so‘z uchun unga eng ko‘p yaqin turgan so‘zlarni sanaydi va **MI o‘lchovi** orqali umumiy uchrashuv ehtimoli bilan solishtiradi. Natijada esa foydalanuvchiga, masalan, tanlangan **X so‘zining 20 ta eng yaqin kollokati ro‘yxati** va **ularning MI skorlari** taklif etiladi. Ushbu axborot tilshunoslikda **valentlik va frazeologizmlarni** o‘rganishda, avtomatik tarjimada sinonimik birikmalarni tanlashda va boshqa sohalarda qo‘l keladi.

N-grammalar esa kengroq tushuncha bo‘lib, u korpusdagi eng ko`p uchrayotgan N uzunlikdagi so‘z zanjirlarini ifodalaydi. Masalan, **2 so‘zli zanjirlar – bigrammalar**; **3 so‘zli – trigrammalar** va hokazo. Korpusdan avtomatik ravishda chastotasi yuqori n-grammalarni chiqarish orqali tilga xos iboralar aniqlanadi. Masalan, ingliz tilida *“at the end of the day”* 5 so‘zli birikmasi juda ko‘p qo‘llaniladi – buni katta korpuslarsiz aniqlash qiyin, lekin korpus statistikasi darhol ko‘rsatadi. O‘zbek tilida korpus yordamida *“degan edi”, “qayd etib o‘tdi”, “shuni ham ta’kidlash joiz”* kabi og‘zaki nutq va rasmiy uslubda keng uchraydigan birikmalarni aniqlash mumkin.

Foydalanuvchi statistikasi. Ayrim korpus tizimlari foydalanuvchilarning qaysi so‘zlarni ko‘p qidirayotgani, eng ko‘p so‘rovlar qaysi tillar uchun berilayotgani kabi meta-statistik ma’lumotlarni ham yuritadi. Bu, albatta, korpusning ichki til ma’lumotidan tashqari, foydalanishiga oid axborotdir. Masalan, milliy korpus administratorlari qaysi davrda qidiruvlar soni oshganini, ko‘proq qaysi turkumdagi so‘zlar izlanayotganini tahlil qilib borishi mumkin. Bu tur ma’lumotlar korpusni yanada takomillashtirish (masalan, eng ko‘p so‘ralayotgan funksiyani qo‘shish) uchun xizmat qiladi.

O‘zbek va jahon korpuslarini qurish tajribasi

Yuqorida keltirilgan komponent va axborot turlari zamonaviy korpus tuzishning **umumiy tamoyillaridir**. Korpus tuzishda maqsad tilning imkon qadar to‘liq va balansli ko‘rinishini yaratish, uni ishonchli lingvistik belgilash bilan ta’minlash va foydalanuvchiga qulay foydalanish imkoniyatini berishdir. Jahonda milliy korpuslar soni yildan yilga ortib bormoqda: ma’lumotlarga ko‘ra, hatto 1990-yilga kelib dunyo tillari uchun 600 ga yaqin korpus mavjud bo‘lgan, hozirda esa minglab korpus loyihalari borligi taxmin qilinadi (bunga turli ixtisoslashgan korpuslar ham kiradi) [13]. Korpuslarning hajmi ham oshib bormoqda – masalan, ingliz tilining Amerikaner korpusi COCA yiliga 20 million so‘zga kengayib, hozirda 1 milliarddan oshdi, Britaniya milliy korpusi esa hali ham 100 million so‘z hajmida **“statik”** holda qolmoqda. Bu shuni ko‘rsatadiki, **monitor korpuslar (doimiy yangilanib boruvchi)** va **statik korpuslar** farqlanadi. COCA singari monitor korpus zamonaviy til o‘zgarishlarini muntazam aks ettirib boradi, BNC kabi statik korpus esa muayyan davrning til hosilasi sifatida tarixiy ahamiyatga ega bo‘ladi. O‘zbek tilida ham istiqbolda monitor korpus yaratish masalasi ko‘ndalang turibdi – chunki tilimiz juda tez yangi so‘zlar bilan boyib, ommaviy axborot til uslubi o‘zgarib bormoqda.

Milliy korpus loyihasi (UzbCorpus.uz). O‘zbek tilining birinchi elektron korpusi sifatida UzbCorpus.uz platformasi 2021 yilda ishga tushirildi. Bu korpus O‘zbekistonda davlat miqyosidagi ilmiy loyiha doirasida boshlangan bo‘lib, dastlab badiiy asarlar matnlarini jamlashdan boshladi. Keyinchalik, B.Elov boshchiligida

ToshDO‘TAU tadqiqotchilar jamoasi ushbu korpusni rivojlantirish ishlari davom ettirildi – unga publitsistik va ilmiy matnlar qo‘shildi, hajmi oshirildi. Hozirda **UzbCorpus.uz** tarkibida 10 milliondan ortiq so‘z mavjud (aniq statistik ma’lumotlar e’lon qilinmagan, lekin ilmiy maqolalarda shu atrofida ko‘rsatilgan). Korpus tarkibi besh funksional uslubni qamrab olgan matnlardan iborat ekani ilmiy adabiyotda qayd etilgan. Korpusning interfeysi uch tilda (o‘zbek, ingliz, rus) taqdim qilingan bo‘lib, unda token va lemma bo‘yicha qidiruv, konkordans natijalarini ko‘rish imkoniyati bor. Mazkur korpus haligacha rivojlanishda davom etmoqda. Bu shuni ko‘rsatadiki, o‘zbek korpusini yanada multimediya (matn + audio) shaklida boyitish rejalangan, bu esa jahon korpusshunosligidagi zamonaviy tendensiyalarga mos keladi (masalan, rus tilida ham audio va sheva korpuslari alohida subkorpuslar sifatida kiritilganligi ma’lum).

Jahon korpuslari bilan qiyos. O‘zbek tilini korpuslash tajribasini jahon yutuqlari bilan qiyoslash foydali. Britaniya va Amerika ingliz tilining korpuslari (BNC va COCA) o‘zbek korpusiga o‘rnak bo‘la oladigan bir nechta xususiyatga ega: ularda matnlar juda keng ko‘lamda jamlangan va aniq balanslangan; har bir so‘z to‘liq morfologik tegangan (BNC uchun CLAWS POS-teglar, COCA uchun Penn Treebank tagset asosida); foydalanuvchi interfeysi esa nihoyatda boy funksionalga ega (lemma bo‘yicha izlash, sinonim bo‘yicha izlash, hatto WordNet ma’lumotlariga bog‘langan holda ma’nodoshlik bo‘yicha qidiruv va h.k. – COCA misolida). Rus milliy korpusi esa o‘zbek tiliga ko‘proq strukturaviy jihatdan yaqin bo‘lgani uchun qiziqarli tajriba beradi: unda agglyutinativ xususiyatlar bo‘lmasa-da, boy flektiv morfologiyaga ega rus tilini teglashda katta mehnat qilingan va ayni paytda ushbu korpus ochiq tarzda internetda foydalanishga imkon beradi. RNC tarkibida, yuqorida aytilganidek, sintaktik daraxtlar subkorpusi, she’riy matnlar subkorpusi, parallel korpuslar kabi turli komponentlar borligi e’tiborga loyiq – o‘zbek tilida ham kelgusida shunga o‘xshash parallel korpuslar (masalan, o‘zbek-ingliz, o‘zbek-rus tarjima korpusi) hamda ixtisoslashgan soha matnlari korpuslarini yaratish rejasi bor.

Yana bir jahon loyihasi – **Leipzig Corpora Collection (LCC)** – o‘zbek tili korpusini rivojlantirishda foydali bo‘lishi mumkin. LCC doirasida 250 dan ortiq til uchun bir xil formatda korpus ma’lumotlari va ularning asosiy statistik ko‘rsatkichlari (wordlist, bigram list va hokazo) bepul tarqatiladi. Masalan, LCCda o‘zbek tilidan ham 1 million so‘zli matnlar to‘plami va uning so‘z chastotasi ro‘yxati mavjud. Bu kichik hajmli bo‘lsa-da, qiyosiy tadqiqotlarda as qotadi. Kelgusida o‘zbek milliy korpusini ham Leipzig kabi ochiq formatda, xalqaro standartlarga mos holda eksport qilish va xalqaro korpuslar qatoriga qo‘shish maqsad qilinmoqda – bu o‘zbek tilining dunyo miqyosida o‘rganilishiga, turkiy tillarni solishtirishda ishtirok etishiga xizmat qiladi.

Xulosa. Xulosa o‘rnida, korpus tuzilishi va axborot turlari bo‘yicha yuqoridagi tahlillar shuni ko‘rsatadiki, zamonaviy elektron korpus – bu ko‘p qavatli

murakkab tizimdir. Unda matnlar lingvistik jihatdan boy belgilangan, texnik bazasi kuchli arxitekturaga tayangan va foydalanuvchi uchun keng qidiruv hamda tahlil imkoniyatlarini taqdim etadi. Optimal korpus yaratish uchun ushbu uch asos – **matn bazasi, texnik arxitektura va interfeys –** birdek mustahkam bo‘lishi lozim [14]. Shu bilan birga, korpus bilan bog‘liq **barcha axborot turlari** – lingvistik, metama’lumot va statistik – to‘liq va izchil ravishda o‘zaro bog‘langan holda ishlaydi. O‘zbekiston tilshunoslari tomonidan boshlangan milliy korpus loyihasi ham aynan shu tamoyillarga intilayotgani ko‘zga tashlanadi.

B.Elov, N.Abdurahmonova kabi olimlar korpus lingvistikasi sohasida dastlabki ishlarni amalga oshirib, o‘zbek tilini raqamli korpuslash yo‘lida muhim qadamlar qo‘yishgan – bu ishlarda jahon korpus tajribasiga tayanilgan va milliy til xususiyatlariga mos yechimlar taklif qilingan. Kelayotgan yillarda o‘zbek tili korpusi yanada kengayib, to‘liq belgilangan, milliardlab so‘zli zamonaviy korpusga aylanishi va jahon korpuslaridan kam bo‘lmagan ilmiy ahamiyat kasb etishi kutilmoqda. Bu borada olib borilayotgan ishlar – lingvistik teglashning avtomatlashtirilishi, arxitekturaning optimallashtirilishi, yangi matnlar bilan boyitish – korpus tuzilishini takomillashtirib, foydalanuvchilar va tadqiqotchilar uchun yanada qulay muhit yaratib beradi. Korpus tuzilishi va undagi axborot qatlamlari qanchalik boy bo‘lsa, tilshunoslikda shunchalik chuqur va ishonchli tahlillar qilish imkoni tug‘iladi. Bu esa ona tilimizni ilmiy jihatdan teranroq anglash va kelgusida turli sun’iy intellekt tizimlarida qo‘llash uchun mustahkam asos yaratadi.

Foydalanilgan adabiyotlar:

11. Primova, M. (2023). Til korpuslarida matnlarni annotatsiyalash: afzallik va kamchiliklari. *Uzbekistan: Language and Culture*, *4*(4).
12. Копотев, М. В., & Янда, Л. (2006). Национальный корпус русского языка. *Вопросы языкознания*, (5), 149-155.
13. Avgustinova, T., & Zhang, Y. (2009, September). Exploiting the Russian national corpus in the development of a Russian Resource Grammar. In *Proceedings of the workshop on adaptation of language resources and technology to new domains* (pp. 1-11).
14. Elov, B. B., Khamroeva, S. M., Alayev, R. H., Khusainova, Z. Y., & Yodgorov, U. S. (2023). Methods of processing the uzbek language corpus texts. *International Journal of Open Information Technologies*, *11*(12), 143-151.
15. Suhr, C., Nevalainen, T., & Taavitsainen, I. (Eds.). (2019). *From data to evidence in English language research*. Leiden: Brill.
16. Zeldes, A. (2021). Corpus architecture. In *A practical handbook of corpus linguistics* (pp. 49-73). Cham: Springer International Publishing.

17. Bolgün, M. A. (2013). The significance of data-driven descriptions of forms in explicit grammar instruction. *Procedia-Social and Behavioral Sciences*, *70*, 430-446.
18. https://tsuull.uz/sites/default/files/b_elov_a_abdullayev_n_xudaybergano v_ozbek_tili_korpusi_matnlari_0.pdf
19. Elov, B., Alayev, R., & Abdullayev, A. (2024). Tabiiy tilning statistik modellari. *digital transformation and artificial intelligence*, *2*(6), 178-189.
20. https://www.academia.edu/124652038/Korpus_lingvistikasi
21. Elov, B., & Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. *Computer Linguistics: problems, solutions, prospects*, *1*(1).
22. Elov, B., Hamroyeva, S., Alayev, R., Xusainova, Z., & Yodgorov, U. (2023). O ‘zbek tili korpusi matnlarini qayta ishlash usullari. *Digital transformation and artificial intelligence*, *1*(3), 117-129.
23. Biemann, C., Heyer, G., Quasthoff, U., & Richter, M. (2007). The Leipzig corpora collection-monolingual corpora of standard size. *Proceedings of Corpus Linguistic*, *2007*.
24. Abduraxmonova, N. Z. Q., & Urazaliyeva, M. Y. (2022). O‘zbek tili elektron korpusida (http://uzbekcorpus. uz/) og ‘zaki matnlar korpusini yaratishning nazariy va amaliy masalalari. *Academic research in educational sciences*, *3*(3), B. 644-650.
25. Davies, M. (2015). Corpora: an introduction. *The Cambridge handbook of English corpus linguistics*, 11-31.

TIL KORPUSI MATNLARIDA IMLO TUZATISH

Botir Elov Boltayevich,
texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Ahmedova Maftuna Kaxramon qizi,
Kompyuter lingvistikasi mutaxasisligi magistranti
MaftunaAhmedova1997@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqola til korpusi matnlarida imlo xatolarini aniqlash va tuzatish muammosiga bag‘ishlangan. Tadqiqot kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohasidagi zamonaviy yondashuvlarni tahlil qiladi. Korpus matnlaridagi imlo xatolarini avtomatik aniqlash va tuzatish algoritmlari, usullari va dasturiy vositalari ko‘rib chiqiladi. Tadqiqot natijasida imlo tuzatishning statistik, qoidaga asoslangan va mashina o‘rganishiga asoslangan usullarining samaradorligi taqqoslanadi. Xususan, til korpuslarida imlo xatolarini aniqlash va tuzatish jarayonida qo‘llaniladigan zamonaviy yondashuvlar va ularning o‘zbek tili korpuslariga tatbiq etish masalalari muhokama qilinadi.

Abstract. This article deals with the problem of detecting and correcting spelling errors in language Corpus texts. The study analyzes modern approaches in the field of computer linguistics and natural language processing. Algorithms, methods and software tools for automatic detection and correction of spelling errors in Corpus texts are considered. The study compares the effectiveness of statistical, rule-based, and machine learning-based methods of spelling correction. In particular, the language Corps discusses the modern approaches used in the process of detecting and correcting spelling errors and their application to the Corps of the Uzbek language.

Аннотация. Данная статья посвящена проблеме выявления и исправления орфографических ошибок в текстах языкового корпуса. В исследовании анализируются современные подходы в области компьютерной лингвистики и обработки естественного языка. Рассмотрены алгоритмы, методы и программные средства автоматического обнаружения и исправления орфографических ошибок в текстах корпуса. В ходе исследования сравнивается эффективность статистических методов коррекции орфографии, основанных на правилах, и методов машинного обучения. В частности, обсуждаются современные подходы к выявлению и исправлению орфографических ошибок в языковых корпусах и их применение в корпусах узбекского языка.

Kalit so‘zlar: *imlo tuzatish, til korpusi, tabiiy tilni qayta ishlash, mashina o‘rganishi, statistik modellar*.

KIRISH

Zamonaviy axborot texnologiyalari asrida til korpuslarining sifati va to‘g‘riligi juda muhim ahamiyatga ega. Til korpusi – bu muayyan tildagi matnlarning elektron to‘plami bo‘lib, tilshunoslik tadqiqotlari, kompyuter lingvistikasi, tabiiy tilni qayta ishlash, statistik tahlil va boshqa ko‘plab sohalarda qo‘llaniladi.

Til korpusi – bu tilning haqiqiy foydalanuvchilari tomonidan ishlab chiqarilgan va so‘zlar,iboralar va umuman, til qanday ishlatilishini tahlil qilish uchun ishlatiladigan juda katta matnlar to‘plami. U tilshunoslar, leksikograflar, ijtimoiy olimlar, gumanitar fanlar, tabiiy tillarni qayta ishlash bo‘yicha mutaxassislar va boshqa ko‘plab sohalarda qo‘llaniladi. Korpus shuningdek, dasturiy ta’minotni ishlab chiqishda ishlatiladigan turli til ma’lumotlar bazalarini yaratish uchun ishlatiladi, masalan, bashoratli klaviaturalar, imloni tekshirish va tuzatish, matn/nutqni tushunish tizimlari,matndan nutqqa modullar, mashina tarjimasi tizimlari va boshqalar[1].

Til korpusi – bu biror tilni to‘g‘ri va kompleks tarzda o‘rganish uchun zarur bo‘lgan ma’lumotlarni o‘z ichiga olgan matnlar to‘plamidir. Bunday matnlar, odatda, turli xil yozuvlar va nutq namunalarini o‘z ichiga oladi, masalan, kitoblar, maqolalar, she’rlar, onlayn matnlar, ijtimoiy tarmoq postlari va hokazo. Korpuslar tilshunoslar va kompyuter lingvistikasi mutaxassislari tomonidan tilni tahlil qilish, o‘rganish, semantik va sintaktik strukturalarni aniqlash hamda tilni qayta ishlashning turli texnologiyalarini ishlab chiqish uchun foydalaniladi[2].

Biroq korpus matnlari turli manbalardan olingan bo‘lib, ko‘pincha kiritish xatolari, noto‘g‘ri so‘zlar, texnik xatolar va boshqa nuqsonlarni o‘z ichiga oladi. Bu xatolar korpus asosida ishlaydigan tizimlar ishini sezilarli darajada yomonlashtirishi mumkin. Shuning uchun korpus matnlaridagi imlo xatolarini aniqlash va tuzatish – kompyuter lingvistikasining dolzarb muammolaridan biridir [3].

Turli tillar uchun imlo tuzatish tizimlari mavjud bo‘lsa-da, o‘zbek tili kabi resurslar cheklangan tillar uchun bu masala hali ham to‘liq hal qilinmagan. O‘zbek tilining o‘ziga xos xususiyatlari – lotin va kirill alifbolaridan foydalanish, morfologik boylik, so‘z yasalishining murakkabligi kabi omillar imlo tuzatish tizimlarini yaratishda qo‘shimcha qiyinchiliklar tug‘diradi [4].

Ushbu maqolada til korpusi matnlaridagi imlo xatolarini aniqlash va tuzatish bo‘yicha mavjud yondashuvlar tahlil qilinadi, ularning afzalliklari va kamchiliklari aniqlanadi, hamda o‘zbek tili korpuslarida imlo tuzatish tizimlarini takomillashtirish bo‘yicha takliflar beriladi.

METODOLOGIYA VA ADABIYOTLAR TAHLILI

Tadqiqot metodologiyasi tizimli adabiyotlar tahlili (systematic literature review) usulini qo‘llashga asoslangan. Imlo tuzatish tizimlarini yaratishda asosan uch xil yondashuv qo‘llaniladi: qoidaga asoslangan, statistik va mashina o‘rganishiga asoslangan usullar[5].

Qoidaga asoslangan yondashuvlar til grammatikasi va imlo qoidalariga tayangan holda xatolarni aniqlaydi va tuzatadi. Qoidaga asoslangan tizimlar aniq va tushunarli natijalar bersa-da, barcha hollarda ham yaxshi natija bermaydi. Chunki tilning barcha grammatik qoidalarini ifodalash murakkab. Bundan tashqari, yangi so‘zlar, chet tilidan kirgan so‘zlar va adabiy tilda qo‘llanilmaydigan so‘zlar va eskirgan so‘zlar bilan ishlashda qiyinchiliklar yuzaga keladi.

Statistik yondashuvlar katta hajmdagi korpuslar asosida so‘zlarning chastotasi va ularning kontekstdagi ishlatilishini tahlil qiladi. Bu usul “N-gram” modellari, Noisy Channel modellarni o‘z ichiga oladi. Statistik usullar faqat katta hajmdagi ma’lumotlar mavjud bo‘lgandagina yaxshi natija bergani uchun kam resursli tillar uchun statistik yondashuvni qo‘llash qiyin.

So‘nggi yillarda mashina o‘rganishi va chuqur o‘rganish (deep learning) usullari korpus matnlaridagi imlo xatolarini aniqlash va tuzatishda keng qo‘llanilmoqda. Bu yondashuvda neyron tarmoqlar, xususan rekurrent neyron tarmoqlar (RNN), uzun-qisqa muddatli xotira (LSTM) va transformerlar kabi modellar ishlatiladi. Masalan, I. Petrov, Smirnov va V.Ivanov “Transformer-based spelling correction for Russian text corpora” nomli maqolasida transformer arxitekturasiga asoslangan modelni rus tili korpuslaridagi imlo xatolarini tuzatish uchun taklif etganlar va bu model 92% aniqlikka erishgan[6]. Mashina o‘rganishi usullari yuqori aniqlikka erishishi mumkin, lekin bu usullarni qo‘llash uchun katta hajmdagi annotatsiyalangan ma’lumotlar to‘plami (dataset) talab etiladi.

Bundan tashqari, gibrid (hibrid) yondashuvlar ham mavjud bo‘lib, ular yuqoridagi usullarning kombinatsiyasidan foydalanadi.

Kompyuter lingvistikasi mutaxassislari yaxshi biladilarki, imloviy sifati past bo‘lgan korpus uning asosida yaratiladigan dasturiy mahsulotlar sifatiga salbiy ta’sir qiladi. Masalan, neyron tarmoqlar asosida ishlaydigan sun’iy intellektni xatolari ko‘p matn bilan mashq qildirilsa, u ana shu matndagi xatolarni tilning ajralmas qismi deb qabul qilishga va shundan kelib chiqib ishlashga o‘rganadi. Oddiy misol: sun’iy intellektga berilgan korpusda “tatbiq” so‘zi 40% hollarda “tadbiq” deb yozilgan bo‘lsa, u shu so‘z o‘zbek tilida ikki xil usulda yozilar ekan deb xulosa chiqaradi va shu xulosa asosida ishlaydi. Bu juda soddalashtirilgan misol bo‘lgani bilan, uning o‘zi lingvistik dasturni “*o*‘qitish” uchun ishlatilgan korpus sifati uning yakuniy ishlash sifatiga qanday ta’sir qilishini tushunib yetish uchun yetarli[7].

O‘zbek tili korpuslaridagi imlo xatolarini tuzatish masalasi bo‘yicha maxsus tadqiqotlar juda kam. Mavjud tadqiqotlar ko‘proq rus, ingliz,ispan va boshqa keng tarqalgan tillarga qaratilgan. Bu esa o‘zbek tili korpuslari matnlarini takomillashtirish uchun yangi yondashuvlarni ishlab chiqish kerakligini ko‘rsatadi.

Adabiyotlar tahlili natijasida til korpusi matnlarida imlo tuzatishning turli usullari samaradorligi aniqlandi va qiyosiy tahlil qilindi. 1-jadvalda imlo tuzatish usullarining qiyosiy tahlili keltirilgan.

1-jadval. Imlo tuzatish usullarining qiyosiy tahlili

Usul	**Afzalliklari**	**Kamchiliklari**	**O‘rtacha aniqlik (%)**	**O‘zbek tiliga moslashish imkoniyati**
Qoidaga asoslangan	Tushunarliligi, izohlanishi oson,kichik korpus uchun qo‘llanilishi mumkin.	Yangi so‘zlarni qamrab ololmasligi.	75-80%	Yuqori, lekin morfologik qoidalarni to‘liq kiritish talab etiladi.
Statistik	Kontekstga asoslanganligi, yangi so‘zlarni ham qamray olishi.	Katta hajmdagi ma’lumotlar talab etilishi.	80-85%	O‘rta, katta hajmli korpus talab etiladi.
Mashina o‘rganishi	Yuqori aniqlik, kontekstni yaxshi tushunishi, yangilanish imkoniyati.	Katta annotatsiyalangan ma’lumotlar bazasi talab etilishi, hisoblash resurslarining ko‘pligi	85-92%	Past-o‘rta, annotatsiyalangan ma’lumotlar yetishmasligi.
Gibrid	Turli usullarning afzalliklarini birlashtirishi.	Tizimning murakkabligi, sozlash qiyinligi.	88-94%	O‘rta-yuqori, moslashtirilgan holda yaxshi natija beradi.

Jadvaldan ko‘rinib turibdiki, zamonaviy yondashuv sifatida gibrid tizimlar eng yuqori aniqlikka erishadi, ammo ularni yaratish va sozlash murakkab jarayondir. O‘zbek tili uchun qoidaga asoslangan tizimlar nisbatan oson moslashtirish imkoniyatiga ega bo‘lsa-da, ularning aniqligi pastroq. Mashina o‘rganishiga asoslangan usullar yuqori aniqlikka erishishi mumkin, lekin o‘zbek tili uchun annotatsiyalangan ma’lumotlar yetishmasligi bu usulni qo‘llashni qiyinlashtiradi.

2-jadvalda esa turli tillarda imlo tuzatish tizimlari uchun erishilgan aniqlik ko‘rsatkichlari taqqoslangan.

2-jadval. Turli tillarda imlo tuzatish tizimlarining aniqligi

Til	**Qoidaga asoslangan (%)**	**Statistik (%)**	**Mashina o‘rganishi (%)**	**Gibrid (%)**
Ingliz	82%	88%	94%	96%

Rus	80%	85%	92%	94%
Turkcha	78%	83%	90%	93%
O‘zbek	75%	80%	83%	88%

Jadval ma’lumotlaridan ko‘rinib turibdiki, resurslari cheklangan tillar, xususan turkiy tillarda imlo tuzatish tizimlarining aniqligi ingliz va rus tillariga nisbatan pastroq. Bu ushbu tillar uchun korpuslar hajmining kichikligi, annotatsiyalangan ma’lumotlarning yetishmasligi va tilning o‘ziga xos murakkabliklari bilan bog‘liq.

O‘zbek tili korpuslarida imlo tuzatish tizimlarini takomillashtirish uchun quyidagi yondashuvlar taklif etiladi:

1. Morfologik tahlilga asoslangan qoidalar tizimini takomillashtirish. O‘zbek tilining morfologik boyligi va so‘z yasalishining o‘ziga xos xususiyatlarini hisobga olgan holda qoidalar tizimini taklif etish lozim.

2. “O‘zbek tili milliy korpusi” ma’lumotlaridan foydalangan holda keng qamrovli statistik modellarni yaratish. Bu korpus hajmini kengaytirish, sifatini oshirish orqali amalga oshirilishi mumkin.

3. Transfer learning usulidan foydalanib, rus, turk kabi yaqin tillarda o‘rgatilgan modellarni o‘zbek tiliga moslashtirish. Bu usul orqali ma’lumotlar yetishmasligi muammosini qisman hal qilish mumkin.

4. O‘zbek tili xususiyatlarini hisobga olgan gibrid tizimlarni yaratish: morfologik tahlil, statistik modellar va neyron tarmoqlarni birlashtiradigan yondashuvlarni ishlab chiqish.

Oldimizda turgan yana bir masalalardan biri:o‘zbek tili imlo qoidalarini bir standartga keltirish, alifbo masalasini uzil-kesil hal qilishdan iboratdir.

O‘zbek tili korpuslarida uchraydigan imlo xatolarini tahlil qilganda, xatolarning quyidagi asosiy turlari ko‘p uchraydi:

Lotin va kirill alifbolari orasidagi transpozitsiya xatolari (masalan, “sh” o‘rniga “ш” yoki “w” aksincha)

1. H va X harflari bilan bog‘liq xatolar (masalan, “h” va “x” harflarining almashtirilishi).

2. Qo‘shimchalarning noto‘g‘ri qo‘shilishi (“mening kitobim” emas “meni kitobim”).

3. Qo‘shma so‘zlarning alohida yozilishi bilan bog‘liq xatolar (“oqqush” emas “oq qush” tarzida).

4. Olinma so‘zlarni yozishda yuzaga keladigan xatolar (“multimedia” shaklida emas, “multimediya” shalida yozilish holatlari) ko‘p uchraydi.

Til korpuslarida imlo tuzatish jarayonini avtomatlashtirish uchun kontekstual ma’lumotlardan foydalanish muhim ahamiyatga ega. Chunki ko‘p so‘zlar kontekstga qarab to‘g‘ri yoki noto‘g‘ri yozilgan bo‘lishi mumkin. Masalan, “boshog‘riq” va “bosh og‘riq” so‘zlari kontekstga qarab bir-birining o‘rniga ishlatilishi mumkin, lekin ular turli ma’noga ega.Birinchisida “muammo” ma’nosida kelsa,ikkinchisida boshdagi og‘riq ma’nosida keladi.Bunday holatlarda kontekstual ma’lumotlardan foydalanish imlo tuzatish tizimining aniqligini sezilarli darajada oshiradi.

Zamonaviy transformer asosli modellar (BERT, GPT, T5) kontekstual ma’lumotlarni yaxshi tushunishi va saqlashi bilan ajralib turadi. Ushbu modellarni o‘zbek tili korpuslariga moslashtirish va imlo tuzatish tizimlarida qo‘llash istiqbolli yo‘nalish hisoblanadi.

Korpus matnlarida imlo tuzatish jarayoni faqat xatolarni aniqlash va tuzatishdan iborat emas. Bu jarayon matn sifatini oshirish, uni standartlashtirish va keyingi ishlov berish uchun tayyorlashning muhim bosqichi hisoblanadi. Sifatli tuzatilgan korpus matnlari tilshunoslik tadqiqotlari, mashinali tarjima, avtomatik annotatsiyalash va boshqa ko‘plab vazifalar uchun muhim ahamiyatga ega.

O‘zbek tili korpuslarida imlo tuzatishning hozirgi holati va istiqbollari tahlil qilinganda, quyidagi muammolar va yechimlar aniqlanadi:

1. Annotatsiyalangan ma’lumotlar yetishmasligi – bu muammoni kraudsorsing, mavjud korpuslar asosida sun’iy xatolar yaratish va qo‘shni tillar ma’lumotlaridan foydalanish orqali hal qilish mumkin.

2. O‘zbek tilining o‘ziga xos xususiyatlari – morfologik tahlil, qo‘shimcha qo‘shish qoidalari va so‘z yasalish qonuniyatlarini hisobga olgan algoritmlar yaratish.

3. Lotin va kirill alifbolari orasidagi transliteratsiya muammolari – ikki alifboda ham ishlay oladigan universal tizimlar yaratish.

4. Tilning doimiy rivojlanishi va yangi so‘zlarning paydo bo‘lishi – tizimni doimiy yangilab turish mexanizmlarini ishlab chiqish.

Ushbu muammolar va yechimlarni hisobga olgan holda, o‘zbek tili korpusi matnlarida imlo tuzatish tizimlarini takomillashtirish uchun tizimli yondashuv talab etiladi. Bu yondashuv tilshunoslik, kompyuter dasturlash, statistika va mashina o‘rganishi sohalarini birlashtirgan holda amalga oshirilishi lozim.

XULOSA. Til korpusi matnlarida imlo tuzatish tizimlari zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash sohalarining muhim qismi hisoblanadi. Ushbu tadqiqot natijasida quyidagi xulosalar shakllantirildi:

birinchidan, o‘zbek tili korpuslarida imlo tuzatish tizimlari hozirgi kunda rivojlanish bosqichida bo‘lib, qoidaga asoslangan, statistik va mashina o‘rganishiga asoslangan usullarning kombinatsiyasi eng samarali natija beradi;

ikkinchidan, resurslari cheklangan tillar, xususan o‘zbek tili uchun imlo tuzatish tizimlarini yaratishda tilning o‘ziga xos xususiyatlarini hisobga olish, morfologik tahlil va kontekstual ma’lumotlardan foydalanish lozim;

uchinchidan, zamonaviy transformer-asosli modellarni o‘zbek tiliga moslashtirish va imlo tuzatish tizimlarida qo‘llash samarali natija beradi;

to‘rtinchidan, imlo tuzatish tizimlarining samaradorligini oshirish uchun annotatsiyalangan ma’lumotlar bazasini kengaytirish zarur.

Ushbu tadqiqot natijalari o‘zbek tili korpusi matnlarida imlo tuzatish tizimlarini takomillashtirish uchun asos bo‘lib xizmat qiladi va bu sohada yangi tadqiqotlar uchun yo‘nalishlar belgilab beradi.

Kelajakda bu sohada qilinadigan tadqiqotlar o‘zbek tili korpuslarining sifatini yanada oshirish, mashina tarjimasi, avtomatik annotatsiyalash va boshqa kompyuter lingvistikasi sohalari uchun muhim ahamiyatga ega bo‘ladi.

Foydalanilgan adabiyotlar:

1. Elov B.B., Yuldashev A.U., Yodgorov U.S. Til korpusi turlari va umumiy xususiyatlari //Xorazm Ma’mun akademiyasi, – 2024 – №12 . B.118-124.
2. Elov B., Samadboyeva M. Language corpora and their importance// Xorijiy lingvistika va lingvodidaktika. – 2025.
3. Kukich Karen. Techniques for automatically correcting words in text // ACM Computing Surveys. – 2019. – T. 24. – № 4. – P. 377-439.
4. Bekmurodov A. va Rixsiyeva G. O‘zbek tilida imlo xatolarini avtomatik aniqlash va tuzatish muammolari // O‘zbek tilshunosligi masalalari. – 2020. – T. 4. – № 2. – B. 67-82.
5. Flor, Michael va Futagi Yoko. Approaches to Automated Spelling Correction: A Comparative Study // Natural Language Engineering. – 2021. – T. 27. – № 1. – P. 1-33.
6. Petrov S., Smirnov I va Ivanov V. Transformer-based spelling correction for Russian text corpora // Computational Linguistics and Intellectual Technologies. – 2023. – T. 22. – № 1. – P. 123-138.
7. Jo‘rayev J. O‘zbek tili korpusini yaratish: Muammolar halqasi va yechimlar. // Kompyuter lingvistikasi:Muammolar,yechim,istiqbollar – 2021.– №. 01. – B. 23-26.

NUSRATULLO JUMAXO‘JA MUALLIFLIK KORPUSI LINGVISTIK TA’MINOTI

Akramova Shohista Islom qizi,
tayanch doktoranti
akramovashohista88@gmail.com
ToshDO‘TAU

Annotatsiya: Nusratullo Jumaxo‘ja davrimizning yetuk ziyolilaridan biri bo‘lib, adabiyot va publitsistika sohalarida boy ilmiy meros qoldirgan. Uning maqolalari turli janrlarda yozilgan bo‘lib, mazmunan chuqur va uslubiy jihatdan boyligi bilan ajralib turadi. Ushbu maqolada N.Jumaxo‘janing maqolalarini janr jihatdan tasniflash hamda ularning uslubiy xususiyatlarini ilmiy tahlil qilish maqsad qilingan.

Abstract: Nusratullo Jumahoja is one of the most prominent intellectuals of our time, leaving behind a rich scientific legacy in the fields of literature and journalism. His articles are written in various genres and are distinguished by their depth of content and methodological richness. This article aims to classify Jumahoja’s articles by genre and scientifically analyze their methodological features.

Аннотатсия: Нусратулла Джумахаджа – один из зрелых интеллектуалов нашего времени, оставивший богатое научное наследие в области литературы и журналистики. Его статьи написаны в различных жанрах и отличаются глубоким содержанием и богатым стилем. Целью данной статьи является классификатsiя статей Джумаходжи по жанрам и научный анализ их методологических особенностей.

Kalit so‘zlar: *Maqola, adabiyotshunoslik, til siyosati, Navoiyshunoslik, ijtimoiy hayot, janr, Matnshunoslik.*

KIRISH

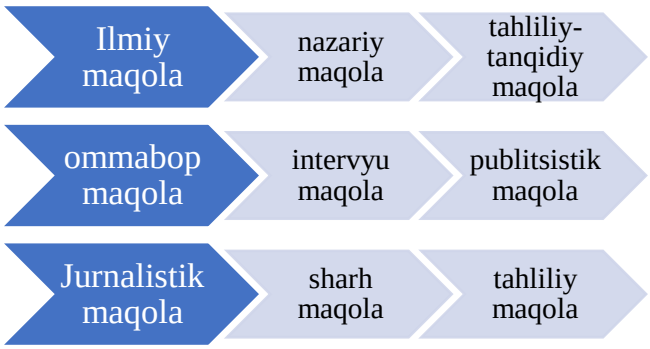
Korpus lingvistikasi – matn (til / lingvistik) korpuslarini ishlab chiqish, yaratish va ulardan foydalanish bilan shug‘ullanadigan soha, aniqroq aytganda, kompyuter lingvistikasining yo‘nalishi. Korpus lingvistikasi (corpus linguistics) termini ingliz tilidagi dastlabki korpuslarning yaratilishi bilan yuzaga kelgan.[1] Bu harakatlarga o‘sha davrda tilshunoslikning zamonaviy yo‘nalishi sifatida qaralgan. Corpus linguistics termini dastlab 1963-yilda ingliz tilidagi korpuslarning yaratilishi bilan foydalanila boshlangan.[2]

Mualliflik korpusi muayyan bir muallif qalamiga mansub matnlar majmuasidan iborat tizim bo‘lib, asosan, yozuvchining yozish uslubi, asarlaridagi til, matn xususiyatlarini axborot texnologiyalari yordamida semantik tasniflab,

tadqiq etishga asoslangan elektron shakdagi ma’lumotlar bazasi sanaladi.[3] Mualliflik korpusining boshqa korpuslardan farqi uning bazasida muayyan bir muallif asarlari jamlanadi, matnlar qayta ishlanadi, grammatik va semantik teglanadi, qidiruv tizimi orqali muallifga tegishli materiallardan zaruriy ma’lumotlar, lingvistik ifodalar aniqlanadi, matnlar asosida statistik ma’lumotlar taqdim etiladi, metama’lumotlar aniq beriladi.[4] Shunday qilib mualliflik korpusi keng imkoniyatli qidiruv tizimiga ega bo‘lib, muallif ijod qilgan barcha tur va janrdagi asarlar qamrab olinadigan, maxsus parametrlar asosida ham qidiruv imkoniyatiga ega, hajmi chegaralanmagan, muallif va uning ijodi bilan bog‘liq manbalarni qulay va tezkor olish mumkin bo‘lgan elektron ma’lumotlar bazasi hisoblanadi.

Nusratullo Jumaxo‘ja zamonaviy o‘zbek adabiyotshunosligi va matnshunosligi sohalarida yetuk olim sifatida tanilgan. Uning ilmiy faoliyati, ayniqsa, Navoiyshunoslik yo‘nalishida muhim o‘rin tutadi. N.Jumaxo‘janing tadqiqotlari o‘zining ilmiy mukammalligi, konseptual yondashuvi hamda tarixiy-madaniy kontekstni chuqur anglashi bilan ajralib turadi. Olim tomonidan yozilgan maqolalar adabiyotshunoslik, va matnshunoslik, Navoiyshunoslik masalalarini qamrab olib, ilmiy jamoatchilikda katta qiziqish uyg‘otgan. Uning “Munis lirikasi va peyzaj”, “Ustoz ijodi – mahorat maktabi (Munis va Ogahiy)”, “Bir nuqtada ming ma’no”, Huvaydo, “Yuksaklik badiiy mahoratda” singari maqolalari adabiyot nazariyasiga oid muhim jihatlarni ochib beradi. Ayniqsa, “Matnshunoslik yangilanmog‘i kerak” nomli maqolasi ilmiy bahs-munozaralarga sabab bo‘lgan bo‘lib, matnshunoslik va filologiya sohasida tub islohotlar zarurati haqida fikr yuritilgan. Nusratullo Jumaxo‘ja ilmiy faoliyatida so‘zga bo‘lgan yuksak hurmat va mas’uliyatni namoyon etadi. Uning “Savodxonlik mezoni”, “Elga e’tibor – tilga e’tibor”, “Til huquqi – el huquqi” kabi maqolalarida til madaniyati, uning to‘G‘ri qo‘llanishi va ilmiy izlanishlar orqali til taraqqiyoti masalalari yoritilgan. Bu esa olimning faqat navoiyshunos emas, balki keng qamrovli filolog sifatida shakllanganligini tasdiqlaydi. N.Jumaxo‘janing ilmiy izlanishlarida Alisher Navoiy asarlarining matnshunoslik tahliliga alohida e’tibor qaratiladi. Uning “Ul sarvi gulro‘ kelmadi”, “G‘azali tadqiqi”, “Navoiy asarlarida mulkdorlik ma’naviyati” kabi maqolalari shoir ijodini yangicha ilmiy talqin etish borasida dolzarb ahamiyat kasb etadi. Xususan, “Navoiyni anglash – o‘zlikni anglash” nomli tadqiqotida Navoiy merosining o‘zbek milliy o‘zligini shakllantirishdagi roli chuqur tahlil qilingan. Nusratullo Jumaxo‘janing ilmiy uslubi puxta va dalillarga asoslangan bo‘lib, uning har bir tadqiqoti adabiy merosni yangicha anglash va baholashga xizmat qiladi. Uning yondashuvi adabiyotshunoslik sohasida innovatsion metodlarni joriy qilishga intilishini ko‘rsatadi. Shu sababli, olimning ilmiy merosi nafaqat o‘zbek filologiyasi balki, umuman, turkiyshunoslik va sharqshunoslik tadqiqotlari uchun ham muhim ahamiyatga ega. Nusratullo Jumaxo‘ja mualliflik korpusini yaratish orqali adabiy

merosni tizimli o‘rganishga erishiladi. Nusratullo Jumaxo‘janing ilmiy va publitsistik maqolalari, navoiyshunoslikka oid tadqiqotlari, matnshunoslikka bag‘ishlangan asarlari keng qamrovli bo‘lib, ularning barchasini yagona tizimga solish ilmiy tahlilni osonlashtiradi. Mualliflik korpusida muallif tilini to‘liq, batafsil, obyektiv ko‘rsatish imkoniyati bor, shuning uchun bunday korpuslar boshqa axborot bankidan afzalligi bilan ajralib turadi. Korpus turli xildagi tadqiqot uchun asos, manba, vosita vazifasini o‘tay oladi[5]. Bunday korpuslarning afzalligi yana shundaki, ular yordamida nafaqat biron bir so‘z yoki jumlaning, balki butun boshli asar tillarini bilish imkoniyati mavjud. Mualliflik korpuslaridagi ma’lumotlarning ilmiy manbalar asosida tahrir qilinganligi, unda berilgan ma’lumotlarning aniqliligi va ishonchliligini kafolatlaydi hamda lingvistik hodisalarning butun spektrini to‘liq aks ettirishi bilan ularni har tomonlama va obyektiv o‘rganishga imkon beradi. Bundan tashqari, til va uslubiy xususiyatlarni tahlil qilish orqali olimning asarlarida ilmiy uslub, lingvistik konstruksiyalar, terminologiya va konseptual qarashlar, olimning ilmiy tili va uslubi chuqur o‘rganilishi mumkin. [6] Yana bir asosiy sababi esa ma’lumki, Nusratullo Jumaxo‘janing ilmiy qarashlari va yondashuvlari nafaqat Navoiy asarlarini sharhlashda, balki umumiy matnshunoslik tamoyillarini ishlab chiqishda ham muhim ahamiyatga ega bo‘lib, korpus orqali olimning Navoiyshunoslikka qo‘shgan hissasi aniqroq yoritishga erishish mumkin. Bugungi Sun’iy intellekt va raqamli filologiya rivojlanib borayotgan bir paytda olimning mualliflik korpusi kelajakda sun’iy intellekt asosida matn tahlili, stilometriya, avtomatik tarjima va matnni annotatsiya qilish kabi sohalarda foydalanish uchun zarur bo‘ladi. Nusratullo Jumaxo‘janing ilmiy merosini kelgusi tadqiqotchilar uchun tizimli holda taqdim etish, uning asarlarini yanada keng targ‘ib qilish va yangi ilmiy izlanishlarga zamin yaratish uchun muhimdir.Yuqoridagi sabablar Nusratullo Jumaxo‘ja mualliflik korpusini yaratishning dolzarbligini tasdiqlaydi va uni amalga oshirish adabiyotshunoslik va matnshunoslik sohalari uchun katta ilmiy ahamiyatga ega ekanligini ko‘rsatadi. Olimning maqolalarini tahlil qilar ekanmiz, ularni quyidagi chizma asosida guruhlarga ajratdik.



Maqola bu muayyan mavzudagi tahliliy, ilmiy yoki jurnalistik fikrlarni bayon etuvchi matn turi bo‘lib, turli maqsadlarda yoziladi va o‘z mazmuniga ko‘ra bir

necha toifalarga bo‘linadi. Olimning ilmiy maqolalari ilmiy tadqiqot natijalarini yoritish, nazariy tahlillar qilish va yangi ilmiy qarashlarni ilgari surish ma’lum bir mavzu yoki ilmiy tushunchani nazariy jihatdan chuqur tahlil qilishga asoslansa, tahliliy-tanqidiy maqolalarida esa mavjud ilmiy g‘oyalarga tanqidiy yondashib, ularni tahlil qilish va baholashga qaratiladi.

Ommabop maqolalari keng ommaga mo‘ljallangan bo‘lib, odatda jamiyat uchun dolzarb bo‘lgan mavzularni yorituvchi, olimning shaxsiy fikrlari va tahliliy yondashuvi aks etadi. N.Jumaxo‘janing maqolalarida quyidagi uslubiy xususiyatlar kuzatiladi:

Muallif maqolalarida ortiqcha so‘z ishlatmay, fikrni aniq va lo‘nda bayon etishga intilgan. Matnlarning mantiqiy tuzilishi va tushunarliligi yuqori darajada ta’minlangan. Bundan tashqari, tarixiy faktlar, ilmiy dalillar va hayotiy misollarga tayangan holda yozilgan bo‘lib, bu jihat maqolalarining ishonchli va ilmiy asoslangan material sifatida qabul qilish imkonini beradi. Har bir maqolada muallif o‘z shaxsiy qarashlarini ochiq bayon etib, o‘quvchiga ta’sir o‘tkazishga harakat qilishi maqolalarning publitsistik ta’sirchanligini oshirishga xizmat qilgan.

Nusratullo Jumaxo‘janing maqolalari janr jihatidan xilma-xilligi, ilmiy va tahliliy yondashuvlarning uyg‘unligi bilan ajralib turadi. Ushbu maqolalar nafaqat o‘z davri uchun, balki bugungi kun o‘quvchilari uchun ham dolzarb ahamiyat kasb etadi. N. Jumaxo‘janing davlat tili bo‘yicha olib borgan ilmiy-amaliy tadqiqotlari milliy til siyosati va ta’lim tizimi uchun muhim ilmiy asos bo‘lib xizmat qiladi desak, ayni haqiqatni aytgan bo‘lamiz. Nusratullo Jumaxo‘ja maqolalarining mavzulari xilma- xil va rang-barang bo‘lib, olim nafaqat adabiyot olamidagi har bir hodisaga, balki jamiyatdagi bo‘layotgan hodisalarga, til siyosatiga ham befaqr qarab turolmaydi, o‘z qarashini asosli va aniq yetkazadi. Nusratullo Jumaxo‘ja maqolalarini mavzulariga ko‘ra quyidagicha guruhlarga ajratdik:

Adabiyotshunoslikka oid maqolalarda Bobur, Uvaysiy, Ogahiy, Mashrab, Huvaydo, Turdi FaroG‘iy Munis, Feruz va ko‘plab mumtoz adabiyot namoyondalari ijodini o‘rganadi va tahlil qiladi. Bundan tashqari, zamonaviy o‘zbek adabiyotining rivoji, adabiy jarayonlar va badiiy asarlar haqida yozgan maqolalarida olimning klassik va zamonaviy adabiyot tahliliga katta e’tibor berishini kuzatishimiz mumkin.

Olim Alisher Navoiy merosini chuqur o‘rganib, uning asarlarining mazmuni, badiiy va lingvistik xususiyatlari haqida yozgan. Bu maqolalarda Navoiy ijodining zamonaviy adabiyotga ta’siri ham yoritiladi.

Navoiyshunoslikka oid mavzularida Navoiyning bir baytidan tortib, yirik dostonlarigacha, shaxsiyati va Navoiyshunoslikda olib borilayotgan ishlar va ularning salmog‘igacha o‘zing qarashlarini keltirib o‘tadi. “Navoiy va mulkdorlik ma’naviyati”, “Navoiyni anglash”, “Navoiy xazinasi”, “Navoiy qomusi yaratilmoqda”, “Lutfiyning lutfiga sazovor”, “Abadiy daho” kabi maqolalar aynan shular jumlasidan.

Matnshunoslikka oid maqolalarida eski o‘zbek adabiyoti matnlarini ilmiy nashrga to‘g‘ri tayyorlash, qo‘lyozmalarni tadqiq qilish va matnlarning tarixiy-lingvistik xususiyatlarini o‘rganishga oid “G‘alati mashhurni ko‘paytirmaylik”, “Bahslar behuda ketmasin”, “Matnshunoslik umumfilologik faoliyat”, “Matn buzilsa”, “Matnshunoslik yangilanmog‘i kerak”, “Adash”tirmang” bag‘ishlangan.

Adabiy tanqidga oid maqolalarda zamonaviy o‘zbek adabiyoti vakillarining ijodi balki navoiyshunos olimlarning ijodiga ham tanqidiy fikr bildirib, adabiy jarayonga oid qarashlari ifodalangan.

Til siyosati va nutq madaniyati yo‘nalishidagi maqolalar ko‘pchilikni tashkil qiladi. Bu maqolalarda olim til siyosati, davlat tilining rivoji va nutq madaniyatining muhim jihatlariga e’tibor qaratadi. “Davlat tili millat sha’ni va G‘ururi”, “Istiqlol va ona tilimiz”, “Elga e’tibor, tilga e’tibor”, “Istiqlol tildan boshlangan”, “Jayxunning asliyati” kabi maqolalarini aynan shu guruhga kiritdik.

Jumaxo‘ja taniqli ijodkorlar va olimlar haqida yodnomalarini, shuningdek, yosh va taniqli ijodkorlarning ilmiy va badiiy asarlarga taqrizlar ham bergani uchun bir guruhni shunday nomladik.

Ijtimoiy hayotga oid “Simyog‘ochga chiqish sharmti?,” “Jenisbekning jonu jahoni”, “Ko‘chirmakashlik illat”, “Adolatli test taraqqiyot me’zoni”, “Mohi Ramazonki”, “Maktab islohoti-davr talabi”, “Huquqimiz alifbosi”, “Hashar majburiymi?” maqolalarida ijtimoiy adolat, madaniyat, ta’lim va boshqa dolzarb masalalar haqida fikr yuritilgan.

Nusratullo Jumaxo‘janing maqolalari keng qamrovli bo‘lib, ularning har biri o‘zbek adabiyoti va tilshunosligining turli sohalariga ba‘ishlangan. Uning asarlari bugungi kunda ham qimmatli ilmiy manba hisoblanadi.

XULOSA

Nusratullo Jumaxo‘janing maqolalarini o‘rganish, avvalo, uning ilmiy va publitsistik qarashlarini chuqur tahlil qilish, o‘zbek tilshunosligi va adabiyotshunosligi rivojiga qo‘shgan hissasini aniqlash uchun muhimdir. Olimning maqolalari adabiyotshunoslik, matnshunoslik, navoiyshunoslik, adabiy tanqid, til siyosati va nutq madaniyati hamda ijtimoiy hayot kabi dolzarb masalalarni qamrab oladi. Bu yo‘nalishlarning har biri bugungi ilmiy tadqiqotlar uchun muhim bo‘lib, til va adabiyot taraqqiyoti, ijtimoiy-madaniy jarayonlarning tahlili va tushunilishiga xizmat qiladi.

Shuningdek, olimning mualliflik korpusini yaratish ilmiy-tahliliy tadqiqotlar uchun alohida ahamiyatga ega. Mualliflik korpusi N.Jumaxo‘janing ilmiy uslubi, til xususiyatlari va mavzu tanlashdagi o‘ziga xos yondashuvlarini tizimli o‘rganish imkonini beradi. Bu orqali uning ilmiy merosi nafaqat alohida maqolalar darajasida, balki umumiy ilmiy-madaniy jarayondagi o‘rni nuqtayi nazaridan ham baholanadi.

Umuman olganda, Nusratullo Jumaxo‘janing maqolalarining janr xususiyatlari va uslubi, ularning mavzu doirasi, dolzarbligi hamda mualliflik korpusining shakllantirilishi nafaqat filologik tadqiqotlar, balki til siyosati va adabiy tanqid sohalarida ham muhim ilmiy manba bo‘lib xizmat qiladi. Shu sababli, uning asarlarini o‘rganish o‘zbek adabiyotshunosligi va Navoiyshunosligining yanada rivojlanishiga katta hissa qo‘shadi.

Foydalanilgan adabiyotlar:

1. Mengliev B., Hamroeva Sh. Mualliflik korpusining maqsad va vazifalari / Til va adabiyot ta’limida zamonaviy axborot va pedagogik texnologiyalar. Respublika ilmiy-amaliy anjumani materiallari. – Toshkent, 2018. – B. 170-172.
2. G‘ulomova N.S. Mualliflik korpuslarining ta’lim jarayonidagi ahamiyati. /Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” mavzusidagi Xalqaro ilmiy-amaliy konferensiya to‘plami. – Toshkent, 2022. B. 113-118.
3. A. Pawlowski, “Chronological corpora: Challenges and opportunities of sequential analysis. The example of ChronoPress corpus of Polish,” in *Conference: Digital Humanities 2016*, 2016, no. July, pp. 311–313.
4. Rahimov A. Kompyuter lingvistikasi asoslari. – T.: Akademnashr, 2011. – B. 78.
5. Abdullaeva O. O‘zbek tilining internet axborot matnlari korpusini shakllantirishning nazariy va amaliy asoslari. Filol. fan. bo‘yicha fal. dok. (PhD) diss. aftoref. – Toshkent, 2022. – 54 b.
6. Abduraxmonova N. Elektron korpuslarning kompyuter modellari: Filol. fan. dok. diss. avtoref. – Toshkent, 2021. – 74 b.

AUDIOMATNLARDA MORFOLOGIK TAHLIL BOSQICHI

Urazaliyeva Mavluda Yangiboyevna,
mustaqil izlanuvchisi
urazaliyevam@nuu.uz
O‘zbekiston Milliy universiteti

Annotatsiya. Ushbu maqolada audiomatnlar asosida o‘zbek tilining morfologik tahlil bosqichlarini chuqur o‘rganishga qaratilgan lingvistik va texnologik yondashuvlar tahlil qilinadi. Tadqiqotda audiomatnlarning murakkab fonetik va sintaktik xususiyatlari hisobga olingan holda, avtomatik morfemik segmentatsiya, leksik analiz, grammatik kategoriya aniqlash va kontekstual morfologik disambiguatsiya bosqichlari alohida ilmiy metodologiya asosida bayon etilgan. Xususan, BERT arxitekturasiga asoslangan chuqur o‘rgatish modellarining o‘zbek tilidagi og‘zaki nutq korpusi (uzbekcorpus.uz hamda kun.uz transkriptlari) asosida sinovdan o‘tkazilishi natijasida yuqori aniqlikka ega morfologik tahlil tizimining yaratilishi nazariy va amaliy jihatdan asoslab berilgan.

Abstract. This article analyzes linguistic and technological approaches aimed at an in-depth study of the morphological analysis stages of the Uzbek language based on audio texts. The research describes the stages of automatic morphemic segmentation, lexical analysis, grammatical category determination, and contextual morphological disambiguation, taking into account the complex phonetic and syntactic features of audio texts, based on a distinct scientific methodology. In particular, the development of a high-precision morphological analysis system has been theoretically and practically substantiated through testing deep learning models based on the BERT architecture using the Uzbek language oral speech corpus (transcripts from uzbekcorpus.uz and kun.uz).

Аннотация. В данной статье анализируются лингвистические и технологические подходы, направленные на углубленное изучение этапов морфологического анализа узбекского языка на основе аудиотекстов. В исследовании описываются этапы автоматической морфемной сегментации, лексического анализа, определения грамматических категорий и контекстуальной морфологической дисамбигуации с учетом сложных фонетических и синтаксических особенностей аудио текстов на основе специальной научной методологии. В частности, разработка высокоточной системы морфологического анализа была теоретически и практически обоснована путем тестирования моделей глубокого обучения на базе архитектуры BERT с использованием корпуса устной речи узбекского языка (транскрипции с uzbekcorpus.uz и kun.uz).

Kalit so‘zlar: *audiomatn, morfologik tahlil, fonetika, segmentatsiya, BERT, tabiiy tilni qayta ishlash (NLP), stilistika.*

Zamonaviy tilshunoslik, ayniqsa raqamli lingvistika va avtomatik nutqni qayta ishlash yo‘nalishlarida, audiomatnlarni tahlil qilish alohida ahamiyatga ega. Bunda nutq birligi sifatida shakllangan og‘zaki matnlar lingvistik struktura, grammatik qurilish va semantik qatlam nuqtai nazaridan chuqur tahlil qilinadi. Ayniqsa, avtomatik nutqni aniqlash (ASR – Automatic Speech Recognition) va tabiiy tilni qayta ishlash (NLP – Natural Language Processing) tizimlari uchun audiomatnlarning grammatik strukturasini aniq belgilash texnologik jarayonning poydevoridir.

Audiomatnlarni morfologik jihatdan tahlil qilish ularning grammatik tuzilmasini aniqlashda muhim bosqich hisoblanadi. Bu bosqichda audiomatnlar transkripsiya qilinib, so‘zlar morfemalarga ajratiladi va har bir morfema grammatik kategoriyalar nuqtai nazaridan belgilab chiqiladi.

Shu o‘rinda qayd etish lozimki, o‘zbek tili agglutinativ til bo‘lgani sababli morfologik tahlilda affiksal ko‘rsatkichlarning roli beqiyosdir. Audiomatnlarda morfologik tahlil bosqichi ularning tarkibiy qismlarini aniqlash, grammatik birliklar va ularning o‘zaro munosabatlarini ko‘rsatish, shu bilan birga morfemik strukturalarni avtomatlashtirilgan tarzda ajratishga xizmat qiladi. Bu jarayon o‘zbek tilining lingvistik xususiyatlarini chuqur aks ettirib, til korpuslarini boyitishga hamda sun’iy intellekt asosidagi til modellari uchun mustahkam asos yaratadi [1].

Morfologik tahlil natijalari asosida korpusga morfologik teglar biriktiriladi. Bu esa til korpusining yanada boyitilishiga va avtomatik analiz tizimlarining takomillashtirilishiga xizmat qiladi.

Audiomatnlarda morfologik tahlil bosqichi — tilshunoslik, kompyuter lingvistikasi va sun'iy intellekt asosidagi tizimlar uchun til materialining grammatik va strukturaviy xususiyatlarini aniqlovchi eng muhim bosqichlardan biridir. Ushbu tahlilning asosiy vazifasi — og‘zaki nutqda shakllangan til birliklarini transkripsiya qilish, ularni morfemalarga ajratish va grammatik kategoriyalar asosida teglash orqali ularning funksiyalarini aniqlashdir.

Birinchi bosqich *transkripsiya* bo‘lib, audiomatn fonetik shakldan ortografik (yozma) shaklga ko‘chiriladi. Bu jarayonda real talaffuzdagi ohanglar, urg‘ular, qisqarishlar, eliziyalar va boshqa artikulyator xususiyatlar inobatga olinadi. Masalan, uzbekcorpus.uz platformasidagi audiomatnlarda “ketayotgan” fe’lining og‘zaki shakli [ket’yotgan] tarzida transkripsiya qilinadi. Nutq tezligi, urg‘u o‘zgarishi va regional talaffuz bu bosqichga bevosita ta’sir qiladi [1]. Shu bois, transkripsiyalashda avtomatik vositalardan tashqari lingvistik ekspertizaga asoslangan tekshiruv ham zarur.

Bugungi kunda transkripsiyani avtomatlashtirish bo‘yicha jahon tajribasida bir nechta ilg‘or modellar mavjud. Masalan, Mozilla’ning DeepSpeech, Metaning wav2vec 2.0 yoki Kaldi kabi ochiq kodli ASR (Automatic Speech Recognition) modellar audiomatnni yozma matnga aylantirishda keng qo‘llaniladi [2]. Ushbu modellar o‘zbek tiliga moslashtirilsa, transkripsiya jarayoni yanada avtomatlashtirilgan va sifatli amalga oshirilishi mumkin.

Transkripsiyadan keyingi bosqich *morfemik tahlil* hisoblanadi. Bu bosqichda matndagi har bir so‘z grammatik birliklarga — morfemalarga — ajratiladi va ular grammatik kategoriya bo‘yicha belgilanadi. Masalan, “qilolmayapti” fe’li uzbekcorpus.uz korpusida qil+OL+NEG+PROG+3SG tarzida teglanadi: OL — imkon, NEG – inkor, PROG – hozirgi davomiy zamon, 3SG – uchinchi shaxs, birlik[3]. Bu kabi tahlillar [4] affiksal tizim tasnifiga to‘liq mos keladi.

Yana bir muhim jihat shundaki, fonetik jihatdan qisqargan shakllar morfemik tahlil jarayonida noto‘g‘ri ajratilish xususiyatiga ega. Masalan, “kelayotgan edi” so‘z birikmasi og‘zaki shaklda [kelyatkanidi] tarzida transkripsiya qilinadi. Bunday hollarni aniqlash uchun fonetik-morfologik segmentatsiya modellari (Foneme2Morph) ishlab chiqilmoqda. Ular tovushlar ketma-ketligidan morfemalarni ajratib olish imkonini beradi[5].

Audiomatnlarda morfologik tahlil ikki komponent — fonetik va grammatik komponentlarning uyg‘un ishlashini talab qiladi. Bu uyg‘unlik transkripsiyadan boshlab teglashgacha bo‘lgan barcha bosqichlarda aniqlik, tizimlilik va tilshunoslikka asoslangan yondashuvni talab etadi.

O‘zbek tili turkiy tillar oilasiga mansub bo‘lib, agglutinativ tillar qatoriga kiradi. Agglyutinativlik hodisasi orqali so‘z asoslariga grammatik ma’no ifodalovchi affikslar qatma-qat qo‘shilib, yangi shakllar va ma’nolar hosil bo‘ladi hamda bunda so‘z negizlari o‘zining shakliy va ma’noviy mustaqilligini saqlab qoladi. Har bir affiks o‘zining alohida grammatik funksiyasiga ega bo‘lib, mustaqil morfema sifatida ko‘riladi[4].

Agglutinativ tillarning asosiy xususiyati shundan iboratki, har bir qo‘shimcha faqat bitta grammatik ma’noni ifodalaydi va ularni so‘z tarkibida aniq farqlash mumkin; bu esa morfologik tahlil jarayonini ancha soddalashtiradi. Ushbu jihatdan agglutinativ tillar flektiv tillardan farqlanadi, chunki flektiv tillarda bir qo‘shimcha bir nechta grammatik ma’noni ifodalash xususiyatiga ega bo‘ladi va morfemalar orasidagi chegaralar unchalik aniq emas[6].

Masalan, “boryapman” fe’l shaklini morfologik tahlil qiladigan bo‘lsak, quyidagi morfemalarni ajratish mumkin: bor — fe’l asosi (leksik morfema); -ya — davomiylik morfemasi; -p — shaxs qo‘shimchasidan oldingi fonetik ko‘makchi; -man — 1-shaxs, birlik ko‘rsatkich morfemasi.

O‘zbek tilining agglutinativlik xususiyatini kuzatishda uzbekcorpus.uz platformasida mavjud misollardan foydalanish mumkin bo‘lib, bu til materiallari va ularning tahlillarini yanada boyitishga yordam beradi:

“O‘qiyotganimda” so‘z shakli quyidagicha tahlil qilinadi: o‘qi (fe’l asosi), -yot (davomiylik), -gan (sifatdosh yasovchi qo‘shimcha), -im (egalik), -da (kelishik qo‘shimchasi).

“Bilimdonligimizdan” shakli: bilim (ot), -don (sifat yasovchi qo‘shimcha), -lik (abstrakt ot yasovchi qo‘shimcha), -imiz (egalik ko‘plik shakli), -dan (kelishik).

“Mustaqilligimizning” shakli: mustaqil (sifat), -lik (abstrakt ot yasovchi), -imiz (egalik), -ning (qaratqich kelishik).

Ushbu xususiyat tufayli o‘zbek tilida so‘z hosil qilish imkoniyatlari kengayib, lug‘at tarkibi boyitiladi hamda grammatik shakllar aniq ifodalangan holda yuzaga keladi, shuningdek, so‘z yasash jarayonida aniqlik va tizimlilik ta’minlanadi. Jahon tilshunosligida ham agglutinativ tillarning morfologik tuzilishi va xususiyatlari bo‘yicha ko‘plab tadqiqotlar olib borilgan bo‘lib, bu tadqiqotlar morfemalar orasidagi munosabatlar va ularning morfologik strukturalarini batafsil ochib beradi [7].

Demak, o‘zbek tilining agglutinativ tabiati uning morfologik tizimini boyitib, tilning ekspressivligini va so‘z hosil qilishdagi imkoniyatlarini oshiruvchi muhim omil sifatida xizmat qiladi.

O‘zbek tilida affiksal morfemalar so‘z shakllariga grammatik ma’no yuklashda muhim vosita hisoblanadi. Ular tilshunoslikning morfologiya bo‘limida markaziy o‘rin tutadi. Affiksal ko‘rsatkichlar yordamida so‘zlarning zamon, shaxs, son, egalik, kelishik kabi grammatik kategoriyalarga mansubligi belgilanadi [8].

Birinchidan, affikslar orqali **zamon** kategoriyasi ifodalanadi. Masalan, fe’lga qo‘shilgan -di, -moqda, -adi, -yapti kabi affikslar harakatning vaqt shaklini bildiradi: *bordim*, *ketmoqda*, *boradi*, *kelayapti* kabi.

Ikkinchidan, affiksal morfemalar **shaxs va son**ni ifodalashda xizmat qiladi. Masalan: *keladi* (3-shaxs, birlik), *kelamiz* (1-shaxs, ko‘plik), *kelasanlar* (2-shaxs, ko‘plik) shakllarida bu grammatik ma’no affikslar orqali beriladi [9].

Uchinchidan, **egalik kategoriyasi** o‘zbek tilida egaga tegishlilikni bildiruvchi affikslar yordamida ifodalanadi: *kitobim* (1-shaxs, birlik), *kitobing* (2-shaxs, birlik), *kitoblari* (3-shaxs, ko‘plik) kabi shakllar buning yaqqol namunasidir [10]. Shuningdek, **kelishik kategoriyasi** ham affikslar orqali amalga oshiriladi: *kitobni* (tushum), *kitobga* (jo‘nalish), *kitobdan* (chiqish) va hokazo.

Affikslarning o‘zaro kombinatsiyalanishi esa grammatik jihatdan murakkab sintaktik tuzilmalarni yaratishga imkon beradi. Masalan, *yozayotgan edim* shaklida

-ayotgan davomiylikni, -edi o'tgan zamonni, -m esa 1-shaxs birlikni bildiradi [11]. Bu kabi murakkab morfologik birliklar, nutqiy faoliyatda ifoda aniqligi va ma’no aniqligini ta’minlaydi.

O‘zbek tilidagi affiksal ko‘rsatkichlar zamon, shaxs, son, egalik va kelishik kategoriyalarini shakllantirish va ifodalashda asosiy grammatik vosita hisoblanadi. Bunday tahlillar uzbekcorpus.uz platformasida mavjud audiomatnlar misolida yanada chuqurroq tadqiq qilinishi mumkin bo‘lib, amaliy tilshunoslikda muhim ahamiyatga ega.

Til korpuslari hozirgi zamon lingvistik tadqiqotlarining muhim qismi sifatida tilshunoslik sohasida keng qo‘llanib kelmoqda. Har bir morfemalangan so‘z teglar bilan belgilanishi morfologik tahlilning avtomatlashtirilishi uchun asos hisoblanadi. Ushbu jarayon avtomatik tizimlar orqali amalga oshirilishi esa korpuslarning samarali boyitilishi va leksik-semantik tahlillar sifatining oshishiga xizmat qiladi [12].

Morfologik teglash tizimi so‘zlarning tarkibiy qismlariga ajratilib, morfemalar darajasida tahlil qilinishiga asoslangan bo‘lib, har bir ajralgan morfema maxsus morfologik teglar bilan belgilab chiqiladi. O‘zbek tilida ishlab chiqilgan morfologik teglar tizimi jahon miqyosida keng tarqalgan Universal Dependencies (UD) [13] formatiga asoslangan holda moslashtirilmoqda. Bu esa xalqaro lingvistik standartlarga mos keluvchi, shuningdek, o‘zbek tilining morfologik xususiyatlarini to‘liq aks ettiruvchi tizim yaratilishini ta'minlaydi.

Masalan, quyidagi shaklni ko‘rib chiqaylik:

borayapman → bor+VA+PROG+1SG

Bu yerda:

VA — fe’l asos;

PROG — hozirgi davomiy zamon shakli;

1SG — birinchi shaxs, birlik shakli.

Bu kabi teglash amaliyotining afzalligi shundaki, murakkab grammatik tuzilishlar avtomatik tarzda aniqlanib, tizim yordamida morfologik jihatdan to‘g‘ri segmentatsiyalanadi. Bu o‘z navbatida, so‘zlarning grammatik va leksik xususiyatlarini chuqurroq tahlil qilish imkoniyatini yaratadi.

O‘zbek tilida morfologik tahlil uchun quyidagi boshqa misollarni ham keltirish mumkin:

“yozmoqda” → yoz+VA+PROG+3SG

“kelishmoqda” → kel+VA+PROG+3PL

“o‘qidim” → o‘qi+VA+PST+1SG

uzbekcorpus.uz platformasi asosida olingan quyidagi misollar ham buni tasdiqlaydi:

“Gapirib turibman” → gapir+VA+CONV+tur+VA+PROG+1SG

“Ko‘rib chiqamiz” → ko‘r+VA+CONV+chiq+VA+FUT+1PL

“O‘ylab ko‘raman” → o‘y+VA+CONV+ko‘r+VA+PRES+1SG

Morfologik tahlilning avtomatik tarzda amalga oshirilishi tabiiy tilni qayta ishlash (NLP) sohasidagi ko‘plab vazifalarni (masalan, sintaktik va semantik tahlillar, matnlarni avtomatik ravishda klassifikatsiya qilish va boshqalar) bajarishni ancha osonlashtiradi. Bu jihatdan qaralganda, avtomatik teglash tizimlarining til korpuslari yaratish va rivojlantirishdagi roli nihoyatda katta [14].

Bundan tashqari, morfologik teglash yordamida turli murakkab tuzilmalar va shakllar avtomatik tarzda aniqlanadi, bu esa tilni o‘rganish, mashinali tarjima tizimlarini rivojlantirish va nutqni tanish texnologiyalarini yaratishda muhim imkoniyatlar ochadi [15]. Jumladan, o‘zbek tilida qo‘llanilayotgan zamonaviy morfologik tahlil tizimlari lemmatizatsiya (asos shaklini aniqlash) va annotatsiya jarayonlarini samarali tarzda bajarishga imkon beradi.

Shunday qilib, morfologik teglar va avtomatik teglash tizimlaridan foydalanish o‘zbek tilining lingvistik tadqiqotlarida xalqaro standartlarga muvofiqlikni ta’minlash bilan birga, tahlil sifatini oshirish va tilni chuqurroq tadqiq etishga xizmat qiladi.

Zamonaviy lingvistika fanida til tizimining turli darajalari o‘zaro uzviy bog‘liqlikda o‘rganiladi. Ayniqsa, fonetika va morfologiya o‘rtasidagi o‘zaro aloqadorlik til birliklarini aniqlash, tahlil qilish va sun’iy intellekt asosidagi til texnologiyalarini ishlab chiqishda muhim ahamiyat kasb etadi. Og‘zaki nutqda yuz beradigan fonetik o‘zgarishlar, aksariyat hollarda, morfemalarning shaklan buzilishiga yoki noto‘g‘ri talqin qilinishiga sabab bo‘ladi. Bu holat ayniqsa morfemik segmentatsiya, avtomatik transkripsiya va lemmatizatsiya jarayonlarida ko‘zga tashlanadi. Fonetik variantliklar, ya’ni so‘z yoki so‘z birikmalarining talaffuzda fonologik shakli buzilishi, morfemik chegaralarni aniqlashni murakkablashtiradi.

Tilning fonetik va morfologik qatlamlari o‘zaro uzviy bog‘liq bo‘lib, og‘zaki nutq jarayonida ro‘y beradigan fonetik hodisalar morfemalarning sirtqi shakli, tuzilishi va qo‘llanish sohalariga sezilarli darajada ta’sir qiladi. Bu holat ayniqsa nutqni avtomatik tahlil qilish, morfemalarni aniqlash va matnni tilshunoslik yondashuvida strukturaviy tahlil qilishda muhim ahamiyatga ega. Quyida ushbu fonetik hodisalar **kun.uz** saytida e’lon qilingan materiallar asosida tahlil qilinadi:

1. Qisqarish (eliziya) - bu og‘zaki nutqda tovushlarning tushib qolishi orqali so‘z shaklining o‘zgarishi hodisasi bo‘lib, u ko‘pincha nutq tezligining oshishi, talaffuz qulayligi yoki nutqiy stereotiplar ta’sirida yuzaga keladi. Eliziya natijasida ayrim morfemalar o‘zining to‘liq shaklida emas, balki qisqa yoki o‘zgargan variantda namoyon bo‘ladi. **Misol uchun:** “O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev Samarqandga tashrif buyurdi” matni og‘zaki nutqda ko‘pincha ***[Prezident Shavkat Mirziyoyev Samarqandga tashrif buyurdi]*** tarzida aytiladi. Bu yerda “O‘zbekiston Respublikasi” qisqargan, biroq mazmun o‘zgarmagan. Bu morfologik jihatdan determinativ qismni fonetik jihatdan inkorporatsiyalash natijasidir[16].

2. Assimilatsiya — bu qo‘shni tovushlarning o‘zaro ta’siri natijasida bir tovush ikkinchisiga moslashib o‘zgarishidir. Bu holat ayniqsa morfema chegaralarida ko‘proq kuzatiladi. Fonetik o‘zgarish morfemalarning aniq chegaralarini aniqlashda noaniqliklar tug‘diradi. **Misol uchun:** “...Prezident Shavkat Mirziyoyev respublikamizda amalga oshirilayotgan islohotlar...” jumlasidagi “respublikamizda” so‘zi tez aytilganda ***[rezpublikamizda]*** yoki ***[rezbulikamizda]*** tarzida talaffuz qilinishi mumkin, bu esa [s] tovushining [b] yoki [z] tovushlariga moslashganligini ko‘rsatadi[17].

3. Koitasiya (so‘zlar fonetik birikmasi) – bu so‘zlar orasidagi fonetik chegaraning yo‘qolib, bir butun ohangda aytilishi hodisasidir. Bu hodisa bog‘lovchi morfemalarning nutqdagi ifodalanishiga bevosita ta’sir qiladi. **Misol uchun:** “O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev Samarqandga tashrif buyurdi” jumlasi og‘zaki nutqda ***[O‘zbekistonrespublikasiprezidentishavkatmirziyoyevsamarqandgatoshifbuyurdi]*** tarzida uzluksiz aytilishi mumkin. Bu yerda morfemalar o‘zining fonetik chegarasini yo‘qotib, murakkab fonetik birlikni hosil qiladi[18].

4. Urg‘u o‘zgarishi (intonatsion siljish) – gapdagi yoki so‘zdagi muhim qismlarni ajratib ko‘rsatishga xizmat qiluvchi prosodik vosita hisoblanadi. Urg‘uning o‘zgarishi ayrim morfemalarning semantik yukini oshirish yoki kamaytirishga olib keladi. **Misol uchun:** “O‘zbekiston Respublikasi Prezidenti **Shavkat Mirziyoyev** Samarqandga tashrif buyurdi” gapidagi urg‘uning “Shavkat Mirziyoyev”ga tushishi bu shaxsga urg‘u berilganligini ko‘rsatadi. Ammo matn intonatsiyasida “Prezidenti” so‘ziga urg‘u berilsa, bu lavozimga urg‘u bergan bo‘ladi: ***“O‘zbekiston Respublikasi Prezidenti Shavkat Mirziyoyev...”***[16].

5. Epenteza (tovush qo‘shilishi) — bu fonetik muvozanatni saqlash yoki talaffuzni qulaylashtirish maqsadida so‘z yoki morfema tarkibiga ortiqcha tovushning kirib kelishi hodisasidir. Bu holat ayrim morfemalarning fonetik ko‘rinishini o‘zgartiradi, ayniqsa og‘zaki nutqda aniqlik pasayishiga sabab bo‘lishi mumkin. **Misol uchun:** “Ular yirik investitsiyalarni jalb qilishga erishmoqchi”

jumlasidagi “jalb” fe’li og‘zaki nutqda ***[jalib]*** tarzida talaffuz qilinadi. Bu yerda /i/ tovushining ortiqcha kiritilishi orqali fonetik yumshatish amalga oshgan va bu morfemaning sirtqi shakliga ta’sir qilgan[19].

6. Metateza (tovushlarning o‘rin almashuvi) — bu tovushlarning o‘zaro o‘rin almashtirishi natijasida yuzaga keladigan hodisadir. Og‘zaki nutqda tez va emotsional tarzda so‘zlashish natijasida bunday holatlar kuzatiladi va bu morfemalarning standart shaklidan og‘ishini bildiradi. ***Misol uchun*:** “Shartnomani buzgan tomon javobgarlikka tortiladi” jumlasidagi “shartnoma” so‘zi ayrim hollarda og‘zaki nutqda ***[sharntoma]*** tarzida talaffuz qilinadi. Bu tovushlar o‘rnining almashishi orqali morfemaning fonetik tuzilishi buziladi, bu esa morfologik tahlilda qiyinchilik tug‘diradi[20].

7. Reduksiya (tovushning zaiflashuvi yoki tushishi) — bu urg‘usiz holatda tovushlarning zaiflashuvi yoki butunlay tushib qolishidir. Bu hodisa morfemalarning to‘liq shaklini saqlab qolmasdan, qisqartirilgan shakllarda nutqda ishtirok etishiga olib keladi. **Misol uchun:** “Prezident tashabbusi bilan yangi loyiha ishga tushirildi” gapidagi “tashabbusi” so‘zi tez aytilganda ***[tashabsi]*** tarzida qisqarib aytilishi mumkin. Bu yerda [u] tovushining reduksiyaga uchragani morfemaning sirtqi shaklini buzadi[21].

8. Tilshunoslikda fonetik variantlarning morfologik neutralizatsiyasi. Ba’zi fonetik o‘zgarishlar natijasida bir necha morfema shakllari fonetik jihatdan bir hil eshitilishi mumkin. Bunday holatda morfologik ajratish qiyinlashadi. Bu tilshunoslikda ***morfologik neutralizatsiya*** deb ataladi. **Misol uchun:** “Yangi korxonalar tashkil etiladi” va “Yangi korxonani tashkil etadi” jumlalaridagi “etiladi” va “etadi” fe’llari og‘zaki nutqda tez aytilganda ***[etadi]*** tarzida yangraydi, natijada zamon va nisbat farqlari fonetik jihatdan yo‘qoladi va morfologik ajratish imkonsizlashadi[22].

Yuqorida tahlil qilingan fonetik hodisalar – epenteza, metateza, reduksiya va morfologik neutralizatsiya – morfemalarning og‘zaki nutqdagi shakllanishiga va ulardan mazmuniy birlik sifatida foydalanish jarayoniga chuqur ta’sir ko‘rsatadi. Shu sababli, fonetik o‘zgarishlarning morfologiyaga ta’sirini aniqlash faqat nazariy tilshunoslik uchun emas, balki tabiiy tilni qayta ishlash, avtomatik morfologik tahlil, nutqni tanish va matnni kodlash kabi amaliy sohalar uchun ham muhim hisoblanadi.

O‘zbek tilida og‘zaki nutqda yuzaga keladigan fonetik o‘zgarishlar morfemalarning to‘liq yoki an’anaviy shaklini buzadi. Bu holat, ayniqsa, **korpus lingvistikasi**da avtomatik morfologik tahlil tizimlari uchun muammo tug‘diradi. Masalan**, uzbekcorpus.uz** platformasida matnli ma’lumotlar yozma shaklda mavjud bo‘lsada, ushbu matnlarda aks ettirilgan og‘zaki nutq elementlari (yoki og‘zaki nutqqa xos bo‘lgan fonetik variantlar) tahlil qilinayotganda noto‘g‘ri morfologik ajratmalarga olib kelishi mumkin. Kun.uz saytida e’lon qilingan maqolada:

“Loyihaga jalb etilgan xorijiy investorlar...” Ushbu gap og‘zaki nutqda ko‘pincha quyidagi tarzda ifodalanadi: ***[Loyihaga jalib qilgan xorijiy investorlar...]*** Bu yerda “jalb” fe’lining **epentezaga** uchrashi (ya’ni “jalib”) morfologik tahlil tizimlari uchun bu so‘zni yangi morfema deb qabul qilish ehtimolini kuchaytiradi. Bu esa morfologik xatoliklar sonini oshiradi.

Audiomatnlarda morfologik tahlil bosqichi nutq birliklarini grammatik jihatdan strukturaviy tahlil qilish, ularni yozma korpusga moslashtirish va avtomatik tizimlar uchun ma’lumotlar bazasi sifatida tayyorlash imkonini beradi. Ushbu bosqich, ayniqsa, tabiiy tilni avtomatik qayta ishlash (NLP) sohasida tilshunoslikka oid chuqur bilimlar bilan uyg‘un holda amalga oshirilganda, til birliklarini aniq identifikatsiya qilish, ularning grammatik shakllarini ajratish va tahlil qilishda muhim o‘rin egallaydi.

O‘zbek tilining agglutinativ xususiyati, ya’ni grammatik ma’no va vazifalarni ko‘rsatishda affikslarning izchilligi orqali amalga oshirilishi, morfologik tahlilni yanada murakkablashtiradi. Shu bilan birga, fonetik o‘zgaruvchanlik – ya’ni affikslar yoki asoslarning talaffuzdagi o‘zgarishi (masalan, tovush tushishi, assimilatsiya, dissimilatsiya) – avtomatlashtirilgan tahlil jarayonida hisobga olinishi kerak bo‘lgan muhim omildir. Aynan shu nuqtalarda morfologik tahlilning puxta ishlab chiqilgan modelga asoslanishi zarur bo‘ladi.

uzbekcorpus.uz platformasi bu borada amaliy tayanch nuqtasi bo‘lib xizmat qiladi. U yirik va xilma-xil matnlar asosida tuzilgan bo‘lib, so‘z shakllarining real qo‘llanilish namunalari bilan ta’minlaydi. Bu esa nafaqat morfologik tahlilning avtomatlashtirilgan modelini ishlab chiqishda, balki fonetik o‘zgarishlarning statistik tahlilini olib borishda ham katta imkoniyatlar yaratadi. Platformadagi ma’lumotlar morfologik tahlil bosqichining ishonchliligini oshirish bilan birga, algoritmlarni o‘rgatishda sifatli tayanch korpus sifatida xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Abduraxmonova, N. (2021). O‘zbek tili elektron korpusining kompyuter modellari. Filologiya fanlari doktori (DSc) dissertatsiyasi avtoreferati.
2. Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477.
3. Tursunov, J. (2023). "Audiomatnlar va zamon shakllari." O‘zbek tilshunosligi jurnali, №2, 45–53.
4. Mahmudov, N. (2015). O‘zbek tili morfologiyasi: affiksal tizim va uning amaliy tahlili. Toshkent: O‘zbekiston Milliy ensiklopediyasi. – 45b.
5. Cotterell, R., Kirov, C., Sylak-Glassman, J., & Eisner, J. (2015). Morphological segmentation for low-resource languages: Challenges and strategies.

Proceedings of the NAACL Workshop on Computational Morphology and Phonology.

6. Haspelmath, M., & Sims, A. D. (2010). Understanding Morphology (2nd ed.). London: Hodder Education. – 67b.

7. Johanson, L., & Csató, É. Á. (2021). Turkic Languages. 2nd edition. London: Routledge. DOI: https://doi.org/10.4324/9780429025569 – 120b.

8. Mahmudov N., Nurmonov A. (2007). O‘zbek tilining nazariy grammatikasi. Toshkent: Fan. — 134–137-bb.

9. Saidova Z. (2015). Hozirgi o‘zbek tilining funksional grammatikasi. Toshkent: Akademnashr. — 93b.

10. Abdullayev F. (2021). O‘zbek tilida affikslarning semantikasi. Toshkent: Fan. — 57b.

11. Mirtojiev M. (2010). O‘zbek tilida grammatik shakllar va kategoriyalar. Toshkent: Universitet. — 112-bet.

12. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press. Relevant content: Chapter 11 – Part-of-Speech Tagging. – 341-345b.

13. Nivre, J., de Marneffe, M. C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., ... & Zeman, D. (2016). Universal dependencies v1: A multilingual treebank collection. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), 1659-1666.

14. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). New Jersey: Prentice Hall. – 164-170b.

15. Habash, N. Y. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers. – 61-66b.

16. “Shavkat Mirziyoyev Samarqandga tashrif buyurdi”, kun.uz, 2023-yil 12-oktabr.

17. “Prezident islohotlar haqida fikr bildirdi”, kun.uz, 2024-yil 1-mart.

18. “Prezident Shavkat Mirziyoyev yangi farmonni imzoladi”, kun.uz, 2023-yil 5-noyabr.

19. “Prezident investitsiyalarni jalb qilishga alohida e’tibor qaratmoqda”, kun.uz, 2024-yil 14-fevral.

20. “Shartnoma talablari buzilgan taqdirda...”, kun.uz, 2023-yil 9-iyun. [https://kun.uz/News/2023/06/09/shartnoma]

21. “Prezident tashabbusi bilan qator loyihalar amalga oshirildi”, kun.uz, 2023-yil 20-aprel. [https://kun.uz/News/2023/04/20/tashabbus]

22. “Yangi korxonalar tashkil etiladi”, kun.uz, 2023-yil 7-sentabr. [https://kun.uz/News/2023/09/07/korxona]

23. Aripov, M., Sharipbay, A., Abdurakhmonova, N., Razakhova B.: Ontology of grammar rules as example of noun of Uzbek and Kazakh languages. In:

Abstract of the VI International Conference “Modern Problems of Applied Mathematics and Information Technology - Al-Khorezmiy 2018”, pp. 37–38, Tashkent, Uzbekistan (2018).

24. Kubedinova L. Khusainov A., Suleymanov D., Gilmullin R., Abdurakhmonova N. First Results of the TurkLang-7 Project: Creating Russian-Turkic Parallel Corpora and MT Systems. Proceedings of the Computational Models in Language and Speech Workshop (CMLS 2020) co-located with 16th International Conference on Computational and Cognitive Linguistics (TEL 2020) .2020/11: 90-101.

25. Abdurakhmonova N. Dependency parsing based on Uzbek Corpus. InProceedings of the International Conference on Language Technologies for All (LT4All) 2019.

26. N. Abdurakhmonova, U. Tuliyev and A. Gatiatullin, "Linguistic functionality of Uzbek Electron Corpus: uzbekcorpus.uz," 2021 International Conference on Information Science and Communications Technologies (ICISCT), 2021, pp. 1-4, doi: 10.1109/ICISCT52966.2021.9670043.

TARJIMA NAZARIYASIDA PARALLEL MATNLARNING AHAMIYATI

Asqarova Madina Boymurod qizi,
Tayanch doktorant,
madina.asqarova.315@gmail.com
O‘zbekiston davlat jahon tillari universiteti

Annotatsiya: Ushbu maqolada tarjima nazariyasi va amaliyotida parallel matnlarning muhim jihatlari tahlil qilinadi. Shuningdek, parallel matnlarning turlari, ularning qiyosiy tahlili hamda tarjima sifatini oshirishda uslubiy qo‘llanma ekanligi ko‘rib chiqiladi. Parallel matnlarning tarjima universallarini aniqlash, tarjima texnologiyalarini rivojlantirish va tarjimonlarni tayyorlashdagi ahamiyati muhokama qilinadi. Shu bilan bir qatorda parallel korpuslarning hisoblash lingvistikasi va madaniyatlararo muloqot tadqiqotlariga ta’siri muhokama qilinadi.

Annotation: This article explores the crucial role of parallel texts in translation theory and practice. It examines various types of parallel texts, their applications in contrastive analysis, translation quality assessment, and corpus-based studies. The paper discusses how parallel texts contribute to understanding translation universals, developing translation technologies, and enhancing translator training. Furthermore, it analyzes the impact of parallel corpora on computational linguistics and cross-cultural communication studies.

Аннотация: В данной статье рассматривается ключевая роль параллельных текстов в теории и практике перевода. Исследуются различные типы параллельных текстов, их применение в сопоставительном анализе, оценке качества перевода и корпусных исследованиях. В работе обсуждается, как параллельные тексты способствуют пониманию переводческих универсалий, развитию технологий перевода и совершенствованию подготовки переводчиков. Кроме того, анализируется влияние параллельных корпусов на компьютерную лингвистику и исследования межкультурной коммуникации.

Kalit so‘zlar: *parallel matnlar, tarjima nazariyasi, qiyosiy tahlil, korpus lingvistikasi, tarjima universallari, mashinali tarjima, tarjimon tayyorlash.*

Tarjima nazariyasi va amaliyotida parallel matnlarning o‘rni beqiyos ahamiyatga ega. Parallel matnlar - bu asl matn va uning bir yoki bir nechta tillardagi tarjimalaridan iborat bo‘lgan matnlar to‘plamidir. Ular tilshunoslik, tarjimashunoslik va madaniyatlararo muloqot sohalarida keng qo‘llaniladigan muhim vosita hisoblanadi. Ushbu maqolada parallel matnlarning tarjima nazariyasidagi o‘rni, ularning turlari, qo‘llanilish sohalari va ahamiyati batafsil ko‘rib chiqiladi. Parallel matnlar ikki yoki undan ortiq tildagi ekvivalent matnlardan iborat bo‘lib, ular odatda jumlama-jumla yoki paragrafma-paragraf moslashtirilgan

bo‘ladi. Bu matnlar tarjima jarayonini o‘rganish, tillararo farqlarni aniqlash va tarjima sifatini oshirish uchun boy material taqdim etadi.

Parallel matnlarning bir nechta asosiy turlarini ajratib ko‘rsatish mumkin:

1. Ikki tilli parallel matnlar: Bular eng keng tarqalgan tur bo‘lib, asl matn va uning boshqa tildagi tarjimasidan iborat.

2. Ko‘p tilli parallel matnlar: Bunday matnlar asl matn va uning bir nechta tillardagi tarjimalarini o‘z ichiga oladi. Ular turli tillar o‘rtasidagi umumiy va o‘ziga xos xususiyatlarni aniqlash uchun ayniqsa foydalidir.

3. Qayta tarjima qilingan matnlar: Bir asarning turli davrlarda yoki turli tarjimonlar tomonidan amalga oshirilgan tarjimalarini o‘z ichiga oladi. Bu tur "qayta tarjima gipotezasi"ni tekshirish va vaqt o‘tishi bilan tarjima sifatining o‘zgarishini kuzatish imkonini beradi.

4. Psevdo-parallel matnlar: Turli tillarda yozilgan, ammo bir xil mavzuga oid matnlardir. Ular turli madaniyatlarda o‘xshash mavzularning qanday yoritilishini o‘rganish uchun foydalidir [1].

Parallel matnlar tarjima nazariyasi va amaliyotida keng qo‘llaniladi. Ularning asosiy qo‘llanilish sohalariga quyidagilar kiradi:

1. Qiyosiy tahlil: Parallel matnlar turli tillarning grammatik, leksik va uslubiy xususiyatlarini qiyoslash imkonini beradi. Bu tillar o‘rtasidagi farqlarni aniqlash va tarjima jarayonida yuzaga keladigan muammolarni tushunish uchun muhimdir.

2. Tarjima sifatini baholash: Parallel matnlar tarjima sifatini baholash uchun muhim vosita hisoblanadi. Ular asl matn va tarjima o‘rtasidagi ekvivalentlikni, mazmun va uslubning saqlanishini tekshirish imkonini beradi.

3. Korpus lingvistikasi: Parallel matnlar katta hajmdagi lingvistik ma’lumotlarni tahlil qilish uchun foydalaniladigan parallel korpuslarni yaratishda asosiy manba hisoblanadi. Bu korpuslar tarjima universaliyalarini o‘rganish, so‘z va iboralarning tarjima ekvivalentlarini aniqlash uchun qo‘llaniladi.

4. Tarjimon tayyorlash: Tarjima o‘qitish jarayonida parallel matnlar qimmatli resurs hisoblanadi. Ular yordamida talabalar amaliy tarjima ko‘nikmalarini rivojlantirish, turli tarjima muammolarini hal qilish usullarini o‘rganish imkoniyatiga ega bo‘ladilar. Bundan tashqari, parallel matnlar tarjima nazariyasi bo‘yicha nazariy bilimlarni mustahkamlashda ham samarali vosita sifatida xizmat qiladi.

5. Mashinali tarjima: Parallel matnlar mashinali tarjima tizimlarini ishlab chiqish va takomillashtirishda muhim rol o‘ynaydi. Ular statistik va neyron tarjima modellarini o‘qitish uchun zarur ma’lumotlarni ta’minlaydi.

6. Leksikografiya: Parallel matnlar ikki tilli va ko‘p tilli lug‘atlarni tuzishda qo‘llaniladi. Ular so‘z va iboralarning kontekstdagi qo‘llanilishini va ularning turli tillardagi ekvivalentlarini aniqlash imkonini beradi.

Parallel matnlar tarjima nazariyasining rivojlanishiga sezilarli hissa qo‘shgan.

Parallel matnlar turli tillarda va turli tarjimonlar tomonidan amalga oshirilgan tarjimalarda kuzatiladigan umumiy tendensiyalarni aniqlash imkonini beradi. Bu tarjima universaliyalari nazariyasini rivojlantirishga yordam beradi [2]. Parallel matnlar tarjimonlar tomonidan qo‘llaniladigan turli strategiyalarni, jumladan, konkretlashtirish, umumlashtirish, kompensatsiya va boshqalarni o‘rganish imkonini beradi. Bu tarjima jarayonini chuqurroq tushunishga olib keladi. Parallel matnlar turli madaniyatlarga xos bo‘lgan tushunchalar, metaforalar va iboralarning tarjimada qanday uzatilishini kuzatish imkonini beradi. Bu madaniyatlararo muloqot nazariyasini rivojlantirishga hissa qo‘shadi.

Parallel matnlar turli darajadagi ekvivalentlikni (so‘zma-so‘z, grammatik, semantik, pragmatik) o‘rganish imkonini beradi. Bu E. Nida tomonidan taklif etilgan “formal ekvivalentlik” va “dinamik ekvivalentlik” tushunchalarini yanada chuqurroq tadqiq etishga yordam beradi [3]. Tarjima sifatini baholash mezonlarini ishlab chiqish: Parallel matnlar tarjima sifatini baholashning obyektiv mezonlarini ishlab chiqish uchun asos bo‘lib xizmat qiladi. J. House ta’kidlaganidek, “tarjima sifatini baholash uchun asl matn va tarjima matni o‘rtasidagi munosabatni tahlil qilish zarur” [4].Tarjima jarayonini modellashtirish: Parallel matnlar tarjima jarayonining turli bosqichlarini, jumladan matnni tushunish, qayta ifodalash va tahrir qilish bosqichlarini o‘rganish imkonini beradi. Bu D. Gile tomonidan taklif etilgan “tarjima harakatlari modeli”ni yanada takomillashtirish imkonini beradi [5]. Zamonaviy texnologiyalar rivojlanishi bilan parallel matnlarning ahamiyati yanada ortmoqda. Elektron parallel korpuslar yaratilishi tarjima nazariyasi sohasida yangi imkoniyatlar ochmoqda. Ular yordamida katta hajmdagi ma’lumotlarni tez va samarali tahlil qilish, statistik ma’lumotlar olish va turli gipotezalarni tekshirish mumkin. Parallel matnlar tarjimonlik ta’limida keng qo‘llaniladi. K. Nord ta’kidlaganidek, “parallel matnlarni tahlil qilish orqali talabalar turli tarjima muammolarini hal qilish yo‘llarini o‘rganadilar” [6]. Tarjima texnologiyalari: Parallel matnlar tarjima xotirasini yaratish va mashinali tarjima tizimlarini o‘qitish uchun asosiy manba hisoblanadi. F. Zanettin bu haqda shunday yozadi: “Parallel korpuslar zamonaviy tarjima texnologiyalarining asosiy tarkibiy qismi hisoblanadi”[7]. Parallel matnlar ikki tilli va ko‘p tilli lug‘atlarni tuzishda qo‘llaniladi. M. Rundell ta’kidlaganidek, “parallel korpuslar lug‘at tuzuvchilarga so‘z va iboralarning real qo‘llanilishini va ularning ekvivalentlarini aniqlash imkonini beradi” [8]. Parallel matnlar turli tillarning grammatik, leksik va uslubiy xususiyatlarini qiyosiy o‘rganish uchun boy material taqdim etadi. S. Johansson bu haqda shunday yozadi: “Parallel korpuslar tillarning o‘ziga xos xususiyatlarini

aniqlashda bebaho manba hisoblanadi”. Parallel matnlar tarjima va til o‘rganish dasturlarini yaratish va takomillashtirishda ham muhim ro‘l o‘ynaydi. Ular sun’iy intellekt va mashinali tarjima tizimlarini o‘qitish uchun zarur bo‘lgan ma’lumotlar bazasini tashkil etadi. Bu esa o‘z navbatida tarjima nazariyasi va amaliyotining yangi yo‘nalishlarini shakllantirishga olib kelmoqda. M. Baker ta’kidlaganidek, “Parallel matnlar tarjima jarayonini va natijasini tushunishda muhim vosita hisoblanadi”. Ular tarjima universaliyalarini aniqlash, tarjima strategiyalarini tahlil qilish va madaniyatlararo farqlarni o‘rganish imkonini beradi. Parallel matnlarning ahamiyati, ayniqsa, zamonaviy raqamli texnologiyalar rivojlanishi bilan yanada ortmoqda. S. Laviosa bu haqda shunday yozadi: “Parallel korpuslar va ularni qayta ishlash texnologiyalari tarjimashunoslik sohasida yangi tadqiqot imkoniyatlarini ochib bermoqda”. Shu bilan birga, parallel matnlardan foydalanishda ba’zi cheklovlar va muammolar ham mavjud:

1. Reprezentativlik muammosi: K. Malmkjær ta’kidlaganidek, “Har qanday korpus, shu jumladan parallel korpus ham, tilning cheklangan namunasi hisoblanadi va uning natijalarini butun tilga umumlashtirish xavfli bo‘lishi mumkin” [9]. Reprezentativlik muammosi parallel matnlar va tarjima nazariyasida muhim o‘rin tutadi. Bu masala parallel matnlarning qanchalik to‘g‘ri va to‘liq tarzda asl matnni va uning tarjimasini aks ettirishiga bog‘liq. Reprezentativlik tushunchasi parallel matnlarning ma’lum bir til, janr, davr yoki mavzu doirasidagi tarjimalarni qanchalik to‘g‘ri va to‘liq aks ettirishini bildiradi. Ideal holatda, parallel matnlar korpusi o‘rganilayotgan sohaning barcha jihatlarini qamrab olishi kerak. Reprezentativlik muammosining mohiyati bir necha jihatlarga ega. Birinchidan, tanlash xatoligi mavjud bo‘lib, parallel matnlar tanlashda ma’lum bir tur yoki janrga ortiqcha e’tibor berilishi mumkin, bu esa natijalarning noto‘g‘ri talqin qilinishiga olib keladi. Ikkinchidan, hajm chegaralari muammosi bor, chunki barcha mavjud tarjimalarni qamrab olish imkonsiz, shuning uchun tanlangan matnlar qanchalik reprezentativ ekanligini aniqlash qiyin. Uchinchidan, sifat masalasi muhim, chunki tanlangan parallel matnlarning sifati turlicha bo‘lishi mumkin, bu esa umumiy xulosalarga ta’sir qiladi. Reprezentativlikka ta’sir qiluvchi omillar orasida til juftligi, janr va mavzu, vaqt omili hamda tarjimonlar bor. Ayrim til juftliklari uchun ko‘proq parallel matnlar mavjud bo‘lishi mumkin. Ba’zi janrlar yoki mavzular boshqalariga nisbatan ko‘proq tarjima qilinadi. Turli davrlardagi tarjimalar o‘rtasida nomutanosiblik bo‘lishi mumkin. Shuningdek, ayrim mashhur tarjimonlarning ishlari ko‘proq e’tiborga olinishi mumkin. Reprezentativlikni ta’minlash uchun bir necha usullar qo‘llaniladi. Bular orasida tasodifiy tanlash, stratifikatsiyalangan tanlash, korpus hajmini kengaytirish va metama’lumotlardan foydalanish bor. Tasodifiy tanlash mavjud parallel matnlardan tasodifiy namunalar olishni nazarda tutadi. Stratifikatsiyalangan tanlash turli mezonlar bo‘yicha matnlarni proporsional tarzda tanlashni anglatadi. Korpus hajmini kengaytirish imkon qadar ko‘proq parallel

matnlarni qamrab olishga qaratilgan. Metama’lumotlardan foydalanish esa har bir parallel matn haqida batafsil ma’lumotlarni to‘plash va tahlil qilishni o‘z ichiga oladi.

2. Tarjima sifati masalasi: Parallel korpuslarda foydalaniladigan tarjimalar sifati har doim ham yuqori bo‘lavermaydi. Bu haqda D. Kenny shunday yozadi: “Parallel korpuslardagi tarjimalar sifatini baholash va nazorat qilish muhim masala hisoblanadi” [10]. Tarjima sifatini baholash murakkab va ko‘p qirrali jarayon hisoblanadi. Bunda bir qator omillar, jumladan, asl matnning mazmunini to‘g‘ri va to‘liq uzatish, grammatik va uslubiy xatolarning yo‘qligi, terminologiyaning to‘g‘ri qo‘llanilishi, matnning maqsadli auditoriyaga mosligini hisobga olish zarur. Tarjima sifatini ta’minlash uchun tarjimondan nafaqat chuqur lingvistik bilimlar, balki madaniyatlararo muloqot ko‘nikmalari, mavzu bo‘yicha bilimlar va professional etika tamoyillariga rioya qilish ham talab etiladi. Sifatli tarjima yaratish uchun tarjimon asl matnni sinchiklab tahlil qilishi, uning mazmuni va uslubini to‘liq tushunishi, so‘ngra esa bu xususiyatlarni maqsadli tilda qayta yaratishi kerak. Tarjima sifatini baholashning turli usul va mezonlari mavjud. Ba’zi yondashuvlar xatolarni sanash va tasniflashga asoslangan bo‘lsa, boshqalari tarjimaning kommunikativ samaradorligini baholashga qaratilgan. Zamonaviy yondashuvlar ko‘pincha miqdoriy va sifat ko‘rsatkichlarini birlashtirishga intiladi. Tarjima sifatini oshirish uchun turli vositalar va texnologiyalar qo‘llaniladi. Masalan, tarjima xotiralari tizimlari tarjimonlarga avvalgi tarjimalardan foydalanish imkonini beradi, bu esa izchillikni ta’minlaydi va xatolar sonini kamaytiradi. Terminologik ma’lumotlar bazalari esa atamalarning to‘g‘ri va bir xilda qo‘llanilishini ta’minlaydi. Tarjima sifatini nazorat qilish jarayoni odatda bir necha bosqichdan iborat bo‘ladi. Bular qatoriga tarjimonning o‘z-o‘zini tekshirishi, tahrirlash, professional muharririning ko‘rib chiqishi va zarur hollarda mutaxassislar tomonidan texnik ekspertiza o‘tkazilishi kiradi. Tarjima sifati masalasi globallashuv va xalqaro aloqalarning kuchayishi sharoitida yanada dolzarb ahamiyat kasb etmoqda. Yuqori sifatli tarjimalar madaniyatlararo muloqotni yaxshilaydi, xalqaro hamkorlikni rivojlantiradi va global bozorda raqobatbardoshlikni oshiradi.

3. Tillarning o‘ziga xos xususiyatlari: Ba’zi tillar o‘rtasidagi tizimli farqlar parallel matnlarni taqqoslashni qiyinlashtirishi mumkin. M. Olohan bu haqda shunday fikr bildiradi: “Tillarning tipologik farqlari parallel matnlarni moslashtirish va tahlil qilishda qo‘shimcha qiyinchiliklar tug‘diradi”. Tillarning o‘ziga xos xususiyatlari har bir tilning noyob va betakror qiyofasini shakllantiradigan lingvistik, madaniy va tarixiy jihatlarni o‘z ichiga oladi. Bu xususiyatlar tilning fonetik tizimi, grammatik tuzilishi, leksik boyligi va pragmatik qoidalarida namoyon bo‘ladi. Fonetik xususiyatlar tilning tovush tizimini belgilaydi. Har bir tilning o‘ziga xos fonema to‘plami, urg‘u va intonatsiya qoidalari mavjud. Ayrim tillarda murakkab undosh birikmalari keng tarqalgan bo‘lsa, boshqalarida bunday birikmalar

deyarli uchramaydi. Grammatik tuzilish tilning sintaktik va morfologik xususiyatlarini o‘z ichiga oladi. Ba’zi tillar flektiv (o‘zgaruvchan) bo‘lsa, boshqalari agglutinativ (qo‘shimchalar qo‘shish orqali so‘z yasash) xususiyatga ega. So‘z tartibi, kelishiklar tizimi, zamonlar tizimi kabi jihatlar ham tildan tilga farq qiladi. Masalan, yapon tilida fe’l odatda gapning oxirida keladi, ingliz tilida esa fe’l odatda egadan keyin joylashadi. Leksik boylik tilning so‘z boyligini, iboralarini va frazeologizmlarini o‘z ichiga oladi. Har bir tilning o‘ziga xos tushunchalarni ifodalovchi so‘zlari mavjud bo‘lib, ular ko‘pincha boshqa tillarga to‘liq tarjima qilinmaydi. Masalan, nemis tilidagi “Schadenfreude” so‘zi (boshqalarning omadsizligidan zavqlanish) ko‘p tillarda bir so‘z bilan ifodalanmaydi. Pragmatik qoidalar tilning ijtimoiy va madaniy kontekstda qo‘llanilishini belgilaydi. Bu murojaat shakllari, hurmat darajalari, so‘zlashuv odatlari va noverbal muloqot elementlarini o‘z ichiga oladi. Masalan, yapon tilida hurmat darajalari juda murakkab tizimni tashkil etadi, bu esa chet elliklar uchun o‘zlashtirishda qiyinchilik tug‘diradi. Tillarning o‘ziga xos xususiyatlari ularning tarixiy rivojlanishi va madaniy konteksti bilan chambarchas bog‘liq. Til evolyutsiyasi davomida boshqa tillar bilan aloqalar, tarixiy voqealar va jamiyatning rivojlanishi natijasida shakllanadi. Masalan, ingliz tili tarixan german va roman tillarining ta’sirida rivojlangan, bu esa uning lug‘at boyligida aks etgan. Metaforalar va idiomalar ham tilning o‘ziga xos xususiyatlarini aks ettiradi. Ular ko‘pincha ma’lum bir madaniyatning dunyoqarashi va hayot tajribasini aks ettiradi. Masalan, “qo‘y og‘zidan cho‘p olmagan” iborasi o‘zbek tiliga xos bo‘lib, boshqa tillarga so‘zma-so‘z tarjima qilinganda ma’nosi yo‘qolishi mumkin

4. Texnik muammolar: Katta hajmdagi parallel matnlarni qayta ishlash va tahlil qilish uchun maxsus dasturiy ta’minot va katta hisoblash quvvati talab etiladi. G. Corpas Pastor ta’kidlaganidek, “Parallel korpuslarni yaratish va ulardan foydalanish sezilarli texnik va moliyaviy resurslarni talab qiladi” [11]. Texnik muammolar zamonaviy hayotimizning ajralmas qismi bo‘lib, turli sohalarda va turli ko‘rinishlarda namoyon bo‘ladi. Ular oddiy maishiy texnikadan tortib murakkab sanoat tizimlarigacha bo‘lgan keng doiradagi qurilma va tizimlarda uchraydi. Kompyuter va axborot texnologiyalari sohasida texnik muammolar juda keng tarqalgan. Dasturiy ta’minot xatolari, viruslar va zararli dasturlar, tarmoq ulanishi muammolari, ma’lumotlar yo‘qolishi va qurilmalarning ishlamay qolishi kabi holatlar foydalanuvchilar va tashkilotlar uchun jiddiy qiyinchiliklar tug‘diradi. Bu muammolar ish unumdorligiga salbiy ta’sir ko‘rsatishi, moliyaviy yo‘qotishlarga olib kelishi va hatto ma’lumotlar xavfsizligini xavf ostiga qo‘yishi mumkin. Telekommunikatsiya sohasida signal kuchi pasayishi, aloqa uzilishlari, tarmoq sig‘imining yetishmasligi kabi texnik muammolar uchraydi. Bu muammolar aloqa sifatiga ta’sir qiladi, internet tezligini pasaytiradi va foydalanuvchilar uchun noqulayliklar yaratadi. Ayniqsa, uzoq masofali aloqa va katta hajmdagi ma’lumotlar

uzatishda bu muammolar yanada sezilarli bo‘ladi. Sanoat va ishlab chiqarish sohalarida texnik muammolar jiddiy oqibatlarga olib kelishi mumkin. Mashina va uskunalarning nosozligi, avtomatlashtirish tizimlaridagi xatolar, sifat nazorati muammolari ishlab chiqarish jarayonini to‘xtatib qo‘yishi, mahsulot sifatiga salbiy ta’sir ko‘rsatishi va hatto xavfsizlik xavfini keltirib chiqarishi mumkin. Transport sohasida texnik muammolar alohida e’tiborni talab qiladi. Avtomobillar, samolyotlar, poyezdlar va boshqa transport vositalaridagi nosozliklar nafaqat harakatlanishni qiyinlashtiradi, balki yo‘lovchilar xavfsizligiga ham jiddiy tahdid soladi. Shuning uchun bu sohada profilaktika va muntazam texnik xizmat ko‘rsatish alohida ahamiyatga ega.Energetika sohasidagi texnik muammolar butun bir hududlarning elektr ta’minotiga ta’sir ko‘rsatishi mumkin. Elektr tarmoqlaridagi uzilishlar, generatorlar va transformatorlardagi nosozliklar, energiya taqsimotidagi xatolar katta miqyosdagi qiyinchiliklarga olib keladi va iqtisodiy faoliyatga salbiy ta’sir ko‘rsatadi. Tibbiyot sohasida texnik muammolar o‘ta muhim ahamiyatga ega, chunki ular bevosita inson hayoti va sog‘lig‘i bilan bog‘liq. Tibbiy asbob-uskunalarning nosozligi, diagnostika qurilmalarining noto‘g‘ri ishlashi yoki dori-darmonlarni saqlash tizimlaridagi muammolar jiddiy oqibatlarga olib kelishi mumkin.

Yuqoridagi cheklovlarga qaramay, parallel matnlar tarjimashunoslik va boshqa aloqador sohalarda muhim o‘rin tutishda davom etmoqda. Ular nafaqat nazariy tadqiqotlar uchun, balki amaliy tarjima faoliyati, tarjimonlarni tayyorlash va tarjima texnologiyalarini rivojlantirish uchun ham qimmatli manba hisoblanadi.

Foydalanilgan adabiyotlar:

1. Olohan, M. (2004). Introducing Corpora in Translation Studies. Routledge.
2. Baker, M. (2018). In Other Words: A Coursebook on Translation. 3rd ed. Routledge.
3. Nida, E, (1964). Toward a science of translating. Netherlands.
4. House, J, (2015). Translation quality assessment: past and present. Routledge is an imprint of the Taylor & Francis Group.
5. Gile, D. (1999). Testing the Effort Models’ Tightrope Hypothesis in simultaneous interpreting – a Contribution. Hermes 23.153-172.
6. Nord, C. (2005). Text Analysis in Translation: Theory, Methodology, and Didactic Application of a Model for Translation-Oriented Text Analysis.
7. Zanettin, F. (2012). Translation-Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies. St. Jerome Publishing
8. Rundell, M. (2015). Corpora as game changers: The growing impact of corpus tools for dictionary markers and users. Cambridge University Press.

9. Malmkjær, K. (2005). Linguistics and the Language of Translation. Edinburgh University Press.

10. Kenny, D. (2001). Lexis and Creativity in Translation: A Corpus-based Study. St. Jerome Publishing

11. Corpas Pastor, G. (2008). Investigar con corpus en traducción: los retos de un nuevo paradigma. Peter Lang.

UDK: 811.512.133`42

MASHINA TARJIMASINING SIFATINI OSHIRISHDA PARALLEL KORPUSNING AHAMIYATI

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Xolmonova Iqbola Alisher qizi,
tayanch doktoranti
iqbolabintualisher@gmail.com
ToshDO‘TAU

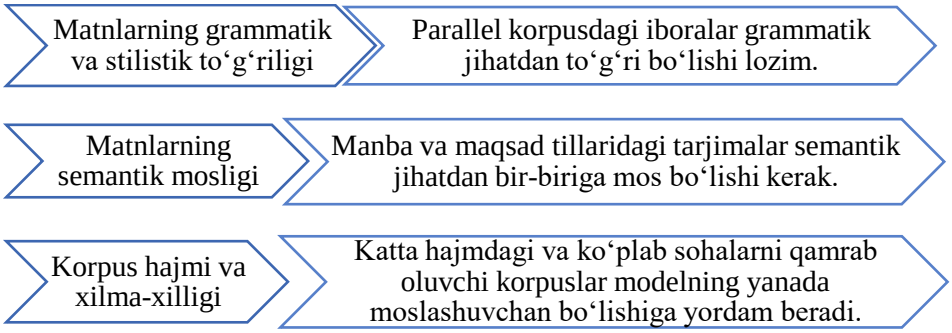
Annotatsiya: Ushbu maqolada mashina tarjimasi sifatiga ta’sir etuvchi asosiy omillar, parallel korpusning tarjima jarayonidagi roli va uning optimallashtirish usullari tahlil qilinadi. Maqolada parallel korpusni optimallashtirishning turli usullari, jumladan, ma’lumotlarni tozalash, balanslash va ko‘paytirish jarayonlari tahlil qilinadi. Korpusning sifatini yaxshilash orqali tarjima tizimlarining samaradorligini oshirish mumkin. Shuningdek, tilning grammatik va struktural farqlarini hisobga olish, ma’lumotlarni optimallashtirish va ko‘paytirishning ahamiyati ko‘rsatilgan.

Kalit so‘zlar: *Mashina tarjimasi, parallel korpus, optimallashtirish, ma’lumotlarni tozalash, ma’lumotlarni ko‘paytirish, tarjima tizimi, til farqlari, korpus sifati, korpusni balanslash*

Zamonaviy mashina tarjimasi texnologiyalari til to‘sig‘ini bartaraf etish va turli tillar o‘rtasida samarali kommunikatsiyani ta’minlashga xizmat qilmoqda. Mashina tarjimasining sifati ko‘plab omillarga bog‘liq bo‘lib, ularning eng asosiylaridan biri o‘quv ma’lumotlari sifatidir. Xususan, parallel korpuslar neyron mashina tarjimasi modellarining samaradorligini oshirishda muhim o‘rin tutadi. Parallel korpusning sifati, hajmi va xilma-xilligi tarjima natijalarining aniq va tabiiy chiqishini ta’minlaydi. Sifatli parallel korpusning yetishmovchiligi esa mashina tarjimasining aniqligi va grammatik to‘g‘riligiga salbiy ta’sir ko‘rsatishi mumkin. Shuning uchun parallel korpusni to‘plash, tozalash va optimallashtirish jarayonlari mashina tarjimasi sifatini oshirishning muhim bosqichlaridan biri hisoblanadi.

Mashina tarjimasi sun’iy intellekt va tabiiy tilni qayta ishlash sohasidagi muhim yo‘nalishlardan biri bo‘lib, turli tillar o‘rtasida avtomatik tarjima qilish imkonini beradi. Hozirgi kunda neyron mashina tarjimasi (NMT) modellarining rivojlanishi tufayli tarjima sifatini sezilarli darajada oshirishga erishilmoqda. Biroq, ushbu tizimlarning samaradorligi turli omillarga bog‘liq bo‘lib, ularning eng

muhimlaridan biri o‘quv ma’lumotlari sifatidir. Parallel korpus – neyron mashina tarjimasi uchun asosiy ma’lumot manbai bo‘lib, tarjimaning aniqligi, tabiiyligi va semantik mosligini belgilaydi. Mashina tarjimasi tizimlari turli bosqichlardan o‘tib ishlab chiqiladi va natijalarning sifati ko‘plab omillarga bog‘liq. Masalan, parallel korpus sifati va hajmi. Parallel korpus mashina tarjimasi modelini o‘rgatish uchun asosiy vosita hisoblanadi. Masalan, o‘zbek-turk neyron mashina tarjimasi uchun tuzilayotgan parallel korpus yetarlicha hajmga ega bo‘lishi va har ikkala til tizimiga mos kelishi lozim. Bunday korpusning sifati quyidagi jihatlarga bog‘liq bo‘ladi:

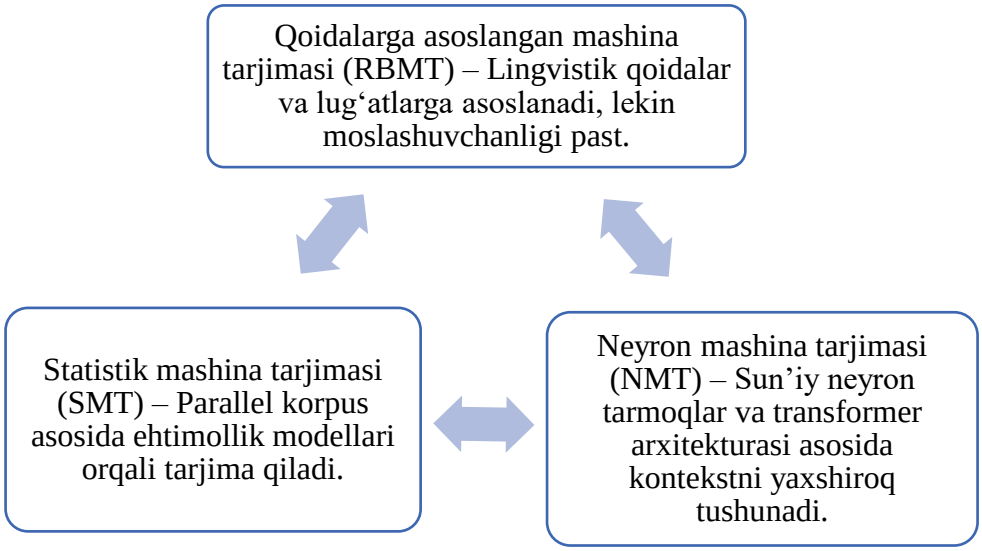


1-chizma. Sifatli korpusning asosiy jihatlari

Mashina tarjimasi tizimlarining samaradorligini oshirish uchun til juftligi o‘rtasidagi grammatik tafovutlarni to‘g‘ri aniqlash va ularni hisobga olish juda muhimdir. Har bir til o‘zining morfologik, sintaktik va semantik xususiyatlariga ega bo‘lib, bu xususiyatlar ikki til o‘rtasidagi tarjimani murakkablashtirishi yoki osonlashtirishi mumkin. Masalan, o‘zbek va turk tillari o‘rtasidagi so‘z tartibidagi grammatik tafovutlar. O‘zbek va turk tillari SOV (Subject-Object-Verb) tuzilmasiga asoslangan bo‘lsa-da, so‘z tartibida ba’zi farqlar mavjud. Misol uchun, o‘zbek tilida ba’zi hollarda so‘z tartibi sintaktik funktsiyalarni aniq ifodalash uchun o‘zgartirilishi mumkin, ammo turk tilida bu nisbatan qat’iyroq. Turk tilida ham so‘z tartibi asosan SOV, lekin ba’zi holatlarda SVO (Subject-Verb-Object) ham ishlatiladi, ayniqsa, urg‘u qayerga qo‘yilishi kerakligi bilan bog‘liq bo‘lgan kontekstlarda. Mashina tarjimasi tizimida bu farqlarni hisobga olish kerak, chunki so‘zlar o‘zaro joylashuvini noto‘g‘ri tarjima qilish tarjima sifatining pasayishiga olib kelishi mumkin. Bundan tashqari sinonimlar va idiomatik iboralarni ishlatishdagi tafovutlar ham mavjud. O‘zbek va turk tillarida ba’zi sinonimlar va idiomatik ifodalar bir-biriga o‘xshash bo‘lishi mumkin, ammo ularning ishlatilishi va konteksti farq qiladi. Masalan: o‘zbek tilida “Yurtimizning kelajagi” (The future of our country) turk tilida “Ülkemizin geleceği” (The future of our country) shaklida bo‘ladi. Ammo, ba’zi hollarda bu so‘zlarning kontekstual ishlatilishi bir-biridan farq qilishi mumkin, masalan, o‘zbek tilidagi “yurakdan” (from the heart) turk tilida “kalpten” sifatida tarjima qilinadi, lekin ba’zi hollarda “gönülden” ham ishlatiladi. Mashina tarjimasi tizimining sifatini oshirish uchun, sinonimlar va idiomatik iboralarni to‘g‘ri

tushunish va tarjima qilish zarur. Bu so‘zlarning kontekstual ma’nosi va ular ishlatilgan muhitni hisobga olish muhim. O‘zbek va turk tillari o‘rtasidagi grammatik tafovutlar mashina tarjimasi tizimining samaradorligini ta’minlashda katta ahamiyatga ega. So‘z tartibi, morfologiya, fe’l tizimi va idiomatik iboralar kabi grammatik xususiyatlarni tushunish va to‘g‘ri aks ettirish mashina tarjimasi tizimlarining sifatini oshiradi. Bunday tafovutlarni hisobga olgan holda, parallel korpuslar tuzish va mashina tarjimasi modellarini o‘rgatishda bu farqlarni inobatga olish zarur. Shuningdek, o‘zbek-turk til juftligi uchun o‘ziga xos grammatik strategiyalar ishlab chiqilishi kerak, bu esa tarjima sifatining yaxshilanishiga olib keladi.

Mashina tarjimasi tizimlarining sifatini oshirishda ikki asosiy omil muhim rol o‘ynaydi: parallel korpus sifati va model arxitekturasi. Ularning o‘zaro ta’siri va to‘g‘ri ishlashiga bog‘liq holda, tarjima tizimlarining aniqligi va samaradorligi sezilarli darajada yaxshilanadi. Shuni hisobga olgan holda mashina tarjimasi texnologiyalarining rivojlanishi natijasida turli yondashuvlar ishlab chiqilgan. Bular:



2-chizma. Mashina tarjimasi yondashuvlari

Hozirgi kunda NMT modellari, ayniqsa transformer arxitekturasi, eng samarali deb hisoblanadi[1]. Mashina tarjimasi tizimlarining samaradorligini oshirishda parallel korpuslardan to‘g‘ri foydalanish juda muhim. Korpusdan maksimal samaradorlik olish uchun ma’lumotlarni oldindan qayta ishlash va optimallashtirish zarur. Ushbu jarayonlar mashina o‘rganish modellari uchun sifatli va optimal trening ma’lumotlarini ta’minlashga yordam beradi.

Parallel korpusdan samarali foydalanish uchun ma’lumotlarni oldindan qayta ishlash va optimallashtirish zarur. Oldindan qayta ishlashda matnni tokenizatsiya qilish, stop-so‘zlarni olib tashlash, tozalash va lematizatsiya kabi jarayonlar amalga oshiriladi. Bu jarayonlar matnni modellar uchun to‘g‘ri formatda taqdim etishga yordam beradi. Optimallashtirish bosqichida, korpusning balanslanishi, ma’lumotlarni ko‘paytirish va modelning giperparametrlari ustida ishlash orqali modelning samaradorligini oshirish mumkin.

Mashina tarjimasi tizimlarining samaradorligini aniqlash va ularning sifatini baholash sohada muhim vazifa hisoblanadi. Tarjima sifatini baholashda parallel korpuslar katta ahamiyatga ega, chunki ular turli tillar o‘rtasidagi to‘g‘ri moslikni aniqlash va tahlil qilish imkonini beradi. Parallel korpusdan foydalanish, mashina tarjimasi tizimlarining samaradorligini o‘lchashda aniq va obyektiv mezonlarni taqdim etadi. Tarjima sifatini baholashda bir qator avtomatik baholash mezonlari mavjud bo‘lib, ulardan eng keng tarqalganlari BLEU, TER, METEOR, NIST va inson baholashidir. Har bir mezon tarjimaning turli jihatlarini – sintaktik moslik, semantik aniqlik, grammatik tuzilish va o‘zaro aloqadorlikni – baholaydi. Parallel korpus orqali bu mezonlar yordamida mashina tarjimasi tizimlarining samaradorligini doimiy ravishda baholash va yaxshilash imkoniyati mavjud. Bu jarayon, o‘z navbatida, til juftliklari orasidagi tafovutlarni hisobga olishga ham yordam beradi, bu esa tarjimaning aniqligi va sifatini oshiradi.

Parallel korpuslar mashina tarjimasi tizimlarining asosiy tayanchidir. Bu korpuslar bir nechta tillarda bir xil ma’lumotlarni o‘z ichiga oladi va mashina tarjimasi modellari uchun o‘rganish materialini taqdim etadi. Parallel korpusning hajmi va sifati mashina tarjimasi tizimlarining umumiy samaradorligi va sifatiga katta ta’sir ko‘rsatadi. Parallel korpus hajmi mashina tarjimasi tizimining samaradorligini belgilovchi asosiy omillardan biridir. Korpus hajmi katta bo‘lsa, model ko‘proq ma’lumotni o‘zlashtirib, o‘z tarjima qobiliyatini yaxshilaydi. Katta hajmdagi parallel korpus mashina tarjimasi modelining generalizatsiya qobiliyatini oshiradi, chunki model turli kontekstlarda so‘zlar va iboralarni qanday ishlatishni o‘rganadi. Bir qator tadqiqotlar shuni ko‘rsatadiki, parallel korpus hajmi ortgan sari, mashina tarjimasi tizimining umumiy sifatida yaxshilanish kuzatiladi. Masalan, Bengio va boshqalarning tadqiqotlarida katta hajmdagi parallel korpuslar yordamida o‘rgangan neyron tarmoq modellari kichik hajmdagi korpuslarga nisbatan ancha yuqori sifatga ega ekanligi aniqlangan[2]. Korpus hajmi ortgan sari modelning tarjima xatolari kamayadi, ayniqsa, grammatik va semantik jihatlardan. Biroq, hajmning ortishi bilan modellarning o‘rganish va hisoblash resurslari talablarining oshishi mumkin. Bu holatni optimallashtirish uchun ma’lumotlarni tozalash, filtrlash va boshqa texnikalardan foydalanish muhimdir.

Parallel korpusning sifati uning mazmuni va tuzilishi bilan bevosita bog‘liq. Sifatli parallel korpuslar o‘z ichiga aniq, to‘liq va grammatik jihatdan to‘g‘ri

tarjimalar kiradi. Agar korpusda ko‘p miqdordagi noto‘g‘ri yoki noaniq tarjimalar bo‘lsa, mashina tarjimasi tizimi o‘rganishda xato natijalar beradi. Shuning uchun korpusning sifati tarjima sifatining yuqori bo‘lishi uchun hal qiluvchi omil hisoblanadi. Quyida korpusning sifatini baholashda asosiy faktorlarni keltirib o‘tamiz:

Grammatik to‘g‘rilik:	Semantik aniqlik:	Kontekstual moslik:
Korpusdagi tarjimalarning grammatik jihatdan aniq va to‘g‘ri bo‘lishi talab etiladi. Grammatik xatolar, so‘zning noto‘g‘ri ishlatilishi yoki sintaktik xatolar mashina tarjimasi tizimining sifatiga salbiy ta’sir ko‘rsatadi.	Tarjimada semantik ma’no xatolari bo‘lmasligi kerak. Agar parallel korpusda semantik jihatdan noaniq yoki chalkash tarjimalar mavjud bo‘lsa, mashina tarjimasi tizimi bu xatolarni o‘rganib, noaniq natijalar beradi.	Tarjimalar kontekstga mos bo‘lishi kerak. Yaxshi sifatli parallel korpus kontekstual farqlarni to‘g‘ri aks ettiradi, bu esa tarjima tizimining aniq va mantiqiy natijalar berishiga yordam beradi.

3-chizma. Korpusning sifatini baholashdagi asosiy faktorlar

Tadqiqotlar shuni ko‘rsatadiki, parallel korpusda sifatli tarjimalar mavjud bo‘lsa, mashina tarjimasi tizimining sifat ko‘rsatkichlari sezilarli darajada yaxshilanadi. Koehn o‘z tadqiqotida parallel korpusning sifati yuqori bo‘lgan hollarda tarjima tizimlarining aniq va mantiqiy natijalar berishini ta’kidlagan[3]. Shuningdek, Liu va boshqalarning tadqiqotlari parallel korpusning sifati va mashina tarjimasi tizimi o‘rtasidagi to‘g‘ri moslikni ko‘rsatadi[4]. Bundan tashqari, parallel korpusning hajmi va sifati o‘rtasidagi muvozanat ham muhimdir. Katta hajmdagi korpuslarning sifati yomon bo‘lsa, mashina tarjimasi tizimi yomon natijalar beradi, chunki model xatoliklar bilan to‘ldirilgan ma’lumotlardan o‘rganadi. Aksincha, kichik hajmdagi yuqori sifatli korpuslar ham samarali bo‘lishi mumkin, lekin bu holatda modelning umumiy o‘rganish imkoniyatlari cheklangan bo‘ladi. Yuqori sifatli va katta hajmdagi parallel korpuslar mashina tarjimasi tizimlarining samaradorligini maksimal darajada oshiradi. Sennrich va boshqalar o‘z tadqiqotida katta hajmdagi yuqori sifatli parallel korpuslar bilan mashina tarjimasi tizimining aniq va samarali ishlashini ko‘rsatgan[5].

Parallel korpusning hajmi va sifati mashina tarjimasi tizimlarining samaradorligi va sifatiga bevosita ta’sir ko‘rsatadi. Korpus hajmi ortgan sari modelning o‘rganish imkoniyatlari kengayadi, lekin uning sifati yaxshilanishi kerak. Katta hajmdagi, lekin sifatsiz korpuslardan foydalanish natijada noaniq va xato tarjimalarga olib kelishi mumkin. Shuning uchun, mashina tarjimasi tizimlarining

muvaffaqiyatli ishlashi uchun sifatli va katta hajmdagi parallel korpuslarni yaratish zarur.

Parallel korpusda noto‘g‘ri, chala, shovqinli (noisy) yoki takroriy (redundant) ma’lumotlar bo‘lishi mumkin. Bu ma’lumotlar mashina tarjimasi tizimining sifatini pasaytiradi. Korpusni optimallashtirishning birinchi qadamlaridan biri – ma’lumotlarni tozalash va filtrlashdir. Ma’lumotlarni tozalash jarayonida, noto‘g‘ri tarjimalar, grammatik xatolar yoki noaniqliklarni bartaraf etish kerak. Ma’lumotlarni tozalash va filtrlashning bir necha asosiy usullari mavjud. Masalan, stop-so‘zlarni olib tashlash. Parallel korpusda har xil tillar o‘rtasidagi umumiy so‘zlar va iboralarning tarjimasi ko‘p bo‘lishi mumkin. Bunday so‘zlar ko‘pincha tarjima sifatiga salbiy ta’sir ko‘rsatadi. Stop-so‘zlarni olib tashlash orqali modelni faqat muhim va kontekstual so‘zlarga yo‘naltirish mumkin. Bundan tashqari, grammatik tozalash ya’ni korpusdagi grammatik jihatdan xato yoki noto‘g‘ri tarjimalarni aniqlash va ularni olib tashlash mashina tarjimasi tizimining sifatini yaxshilaydi. Noto‘g‘ri tarjimalar modelga noto‘g‘ri ma’lumotlar kiritilishiga olib kelishi mumkin, bu esa xatoliklarni oshiradi. Shu bilan birga, parallel korpusda takroriy tarjimalar yoki o‘xshash ma’lumotlar bo‘lishi mumkin. Bu takroriy ma’lumotlar mashina tarjimasi tizimining o‘rganish jarayonini zaiflashtiradi. Shuning uchun redundant ma’lumotlarni olib tashlash korpusni optimallashtirishga yordam beradi[6].

Parallel korpusda tillar o‘rtasidagi so‘zlarning chastotalari va iboralarning turlari o‘rtasidagi farqlar ko‘p uchraydi. Ba’zi tillarda ba’zi so‘zlar yoki iboralar ko‘proq ishlatiladi, boshqalari esa kamroq ishlatiladi. Korpusni balanslash usuli, tillar o‘rtasidagi nisbatlarni muvozanatlash, tarjima tizimining umumiy ishlashini yaxshilaydi. Korpusni balanslash quyidagi usullarni o‘z ichiga oladi:

Sof ma’lumotlar olish:

Parallel korpusda ba’zi tillar juda yuqori chastotali so‘zlarga ega bo‘lishi mumkin. Korpusdagi bu yuqori chastotali so‘zlarni olib tashlash, ular bilan bo‘lgan yuqori chastotali takrorlanishlarni kamaytirishga yordam beradi.

So‘zlarning hamkorligini oshirish:

Korpusda biron-bir so‘z boshqa tilga tarjima qilinganda kamdan-kam uchrasa, u holda bunday ma’lumotlarni kiritishni ko‘paytirish mumkin. Korpusda so‘zlar va iboralar o‘rtasidagi hamkorlikni oshirish mashina tarjimasi tizimining xatoliklarini kamaytiradi.

O‘xshash ma’lumotlarni qo‘shish:

Ba’zi tillarda so‘zlarning chastotalari juda past bo‘lishi mumkin. Bunday hollarda, o‘xshash ma’lumotlardan foydalanib, korpusni kengaytirish mumkin.

4-chizma. Korpusni balanslash usullari

Korpusni balanslash orqali mashina tarjimasi tizimi tilning barcha sohalarini o‘rganib, ko‘proq kontekstual ma’lumotlarni o‘zlashtiradi[7].

Parallel korpusni optimallashtirishning yana bir muhim usuli – bu ma’lumotlarni ko‘paytirishdir. Korpus hajmi kichik bo‘lganda, mashina tarjimasi tizimi o‘rganishda yetarli ma’lumotga ega bo‘lmasligi mumkin. Ma’lumotlarni ko‘paytirish yordamida bu muammoni hal qilish imkoniyati mavjud. Quyida ma’lumotlarni ko‘paytirishning ba’zi usullarini ko‘rib chiqamiz:

Parafrazlash (Paraphrasing):	Sintetik ma’lumotlar yaratish:	Back-translation:
Parafrazlash usulida, tarjima matnlarini qayta yozish orqali yangi tarjimalar hosil qilinadi. Bu usul yordamida yangi tarjimalar yaratiladi, bu esa modelga ko‘proq o‘rganish materialini taqdim etadi.	Sintetik tarjimalar yaratish orqali parallel korpusni sun’iy ravishda kengaytirish mumkin. Sintetik ma’lumotlarni yaratish mashina tarjimasi tizimining o‘rganish imkoniyatlarini kengaytiradi.	Back-translation usulida birinchi tildan ikkinchi tilga tarjima qilingan matn qaytadan birinchi tilga tarjima qilinadi. Bu usul yordamida yangi tarjimalar hosil qilinadi va ularni modelga qo‘shish orqali o‘rganishni yanada mustahkamlash mumkin

5-chizma. Korpus ma’lumotlarini ko‘paytirishning usullari

Bu usullar yordamida, parallel korpusning hajmi oshiriladi va mashina tarjimasi tizimining umumiy samaradorligi yaxshilanadi[5].

Parallel korpusda turli xil matnlar bo‘lishi mumkin, masalan, ilmiy, texnik, adabiy yoki so‘zlashuv uslubiga oid matnlar. Korpusni optimallashtirishda matnlarni kategoriyalash va tasniflash ham juda muhim rol o‘ynaydi. Bu usul, ma’lumotlarni til turiga ko‘ra ajratib, har bir turdagi matn uchun maxsus modellar yaratishga yordam beradi. Korpusni kategoriyalashning asosiy afzalliklariga matn turlari bo‘yicha optimallashtirish kiradi. Ya’ni har bir matn turi uchun alohida model yaratish, mashina tarjimasi tizimining o‘rganish jarayonini yanada optimallashtiradi. Masalan, ilmiy matnlar uchun maxsus model, adabiy matnlar uchun esa alohida model yaratilishi mumkin. Bundan tashqari, tizimni tezlashtirish mumkin. Bunda kategoriyalash yordamida, mashina tarjimasi tizimining o‘rganish jarayoni tezlashadi, chunki model faqat o‘sha turdagi matnlarni o‘rganadi va noaniqliklar kamayadi [1]. Parallel korpusni optimallashtirishda o‘quv algoritmlarini takomillashtirish ham muhimdir. Mashina tarjimasi tizimlarining samaradorligini oshirish uchun o‘rganish algoritmlarini optimallashtirish zarur. Yangi arxitekturalar, masalan, neyral tarmoq va attention mechanism yordamida, model ko‘proq kontekstual ma’lumotlarni o‘zlashtirishi mumkin. Bu esa mashina tarjimasi tizimining to‘g‘ri ishlashiga yordam beradi[1]. Parallel korpusni optimallashtirish, mashina tarjimasi tizimlarining samaradorligini oshirish uchun juda muhim jarayon hisoblanadi. Korpusni tozalash, filtrlash, balanslash, ko‘paytirish va kategoriyalash kabi usullar yordamida modelning sifatini yaxshilash mumkin. Shuningdek, o‘quv algoritmlarini optimallashtirish va yangi texnologiyalarni qo‘llash orqali mashina tarjimasi tizimlari yanada samarali ishlashi mumkin. Bu optimallashtirish usullari parallel korpusdan maksimal foydalanish imkoniyatini yaratadi va mashina tarjimasi tizimining sifatini sezilarli darajada oshiradi.

Mashina tarjimasi tizimlarining samaradorligini oshirishda parallel korpuslarning ahamiyati juda katta. Parallel korpus, ikki yoki undan ortiq tillarda tarjima qilingan matnlar to‘plamidir, va bu korpuslar mashina tarjimasi tizimlarini o‘rgatishda asosiy ma’lumot manbai hisoblanadi. Korpusning hajmi, sifati va tarkibi tarjima tizimlarining natijalariga bevosita ta’sir qiladi.

Parallel korpusni optimallashtirish orqali mashina tarjimasi tizimining sifatini sezilarli darajada oshirish mumkin. Korpusni tozalash, filtrlash va balanslash kabi usullar yordamida sifatni yaxshilashga erishish mumkin. Korpusdagi redundant yoki noaniq ma’lumotlarni olib tashlash, grammatik xatolarni tuzatish va o‘xshash ma’lumotlarni kiritish orqali modelga aniq va ishonchli ma’lumotlar taqdim etiladi. Bundan tashqari, ma’lumotlarni ko‘paytirish va back-translation kabi metodlar yordamida korpusning hajmi kengaytirilishi mumkin, bu esa o‘rganish jarayonini kuchaytiradi. Korpusni kategoriyalash va turli matn turlariga mos ravishda modellarni yaratish, mashina tarjimasi tizimining muvaffaqiyatini oshirishda muhim

o‘rin tutadi. Shuningdek, o‘quv algoritmlarini optimallashtirish va neyron tarmoqlarni qo‘llash orqali tarjima tizimlarining samaradorligi oshirilishi mumkin. Bundan tashqari, parallel korpusning lingvistik xususiyatlarini, grammatika, so‘z tartibi va til strukturasidagi farqlarni hisobga olish, modelni yanada aniqroq va moslashtirilgan tarjimalar yaratishga yo‘naltiradi. Korpusning xilma-xilligi va sifatli tarkibi mashina tarjimasi tizimining tillararo o‘rganishni yaxshilaydi. Shu bilan birga, parallel korpusdan foydalanish nafaqat mashina tarjimasi tizimlarini rivojlantirishga, balki boshqa tilni o‘rganish va lingvistik tadqiqotlar uchun ham katta imkoniyatlar yaratadi. Korpusning sifatini oshirish va optimallashtirish, kelajakda mashina tarjimasi tizimlarining yanada mukammal va ishonchli bo‘lishiga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Vasvani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L., Polosukhin I. Attention Is All You Need. Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017−P. 5998–6008
2. Bengio Y., Ducharme R., Vincent P., Jauvin C. A Neural Probabilistic Language Model. Journal of Machine Learning Research, 3. 2003−P. 1137–1155
3. Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation. Proceedings of the 10th Machine Translation Summit. 2005−P. 79–86.
4. Liu Y. Automatic Evaluation of Machine Translation Quality: A Survey. 2016
5. Sennrich R., Haddow B., Birch A. Neural Machine Translation of Rare Words with Subword Units. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016−P. 1715–1725.
6. Koehn P. Statistical Machine Translation. Cambridge University Press. 2016

Tiedemann J. Parallel Data, Tools and Interfaces in OPUS. Proceedings of the 8th International Conference on Language Resources and Evaluation. 2012−P. 2214–2218

EXPLORING THE STYLISTIC RICHNESS OF FIGURATIVE EXPRESSIONS IN THE UZBEK LANGUAGES

Sulaymonova Mahfuza Shavkat qizi,
PhD Student, I grade
TSUULL

Annotation. This article analyzes the characteristics of figurative devices in the Uzbek language, particularly metaphor, simile, epithet, hyperbole, and litotes. The linguistic structure of these devices, their role in poetic speech, and their functional features are examined based on scientific approaches. The differences between similes and metaphors, their components, and their role in literary texts are illustrated with examples. This article serves as a valuable resource for those engaged in literary analysis, linguistic research, and stylistics.

Аннотация. В данной статье анализируются особенности изобразительных средств в узбекском языке, в частности метафоры, сравнения, эпитеты, гипербола и литота. Рассматриваются их лингвистическая структура, роль в поэтической речи и функциональные особенности на основе научных подходов. Выявлены различия между сравнением и метафорой, их компоненты, а также роль в художественной литературе, что подтверждается примерами. Данная статья является ценным источником для специалистов, занимающихся литературным анализом, лингвистическими исследованиями и стилистикой.

Keywords: *Uzbek language, figurative devices, metaphor, simile, epithet, hyperbole, litotes, poetic speech, literary imagery, stylistics, emotional-expressive language.*

The components of simile that we study – the object being compared, the comparative image, and the comparative attribute – have been referred to by various scholars using different terminologies: B. V. Tomashevsky named them object, image, and attribute[1] ; A. I. Efimov used subject of comparison, object of comparison, and basis of comparison[2]; M. I. Sharipo referred to them as object of comparison, image of comparison, and attribute of comparison[3]; A. Sa’diy named them the compared element, the element it is compared to, and the attribute of similarity[4].

Depending on whether the comparative attribute in the simile is expressed through lexical units or not, similes can be divided into complete and incomplete similes. If all three components are expressed through lexical units, it is referred to as a complete simile. In an incomplete simile, the comparative attribute is not

explicitly expressed. This allows us to classify similes as three-component and two-component structures.

For example:

a) Three-component simile:

Nizomjon, unsure of what to say, glanced at the golden-crumpled velvet-like shimmering sands. (Said Ahmad)

In this example:

"golden-crumpled velvet-like" is the comparative image,

"shimmering" is the comparative attribute,

"sands" is the object being compared.

b) Two-component simile:

The woman stood leaning against the wall near the door, her face pale as a corpse, trembling. (A. Qahhor)

Here, the woman's paleness is compared to that of a corpse.

"face" is the object being compared,

"as a corpse" is the comparative image,

and the comparative attribute is not explicitly mentioned.

In speech, qualification (attribution) and simile are closely interconnected in terms of their function of conveying concreteness and emotional-expressive evaluation. They transition into one another in artistic imagery and assist each other in characterizing objects. The main aesthetic function of simile is the function of concreteness. If the attributive relation is removed from a simile, its meaning becomes less clear and less emotionally expressive. Therefore, similes often appear alongside qualifiers (adjectives or attributes).

Hyperbole is the art of exaggerated description of objects and events. Based on its nature, hyperbole is categorized into simple, igroq (intensified hyperbole), and gulu (unrealistic, implausible exaggeration that goes beyond possibility).Even though igroq and gulu types of hyperbole are often used without explicit simile markers, a simile can be understood from the meaning. In the example from Navoi (poetic part omitted here), the beloved's face being covered is likened to the setting of the sun, and the tears shed due to separation are likened to the appearance of stars after sunset. Hyperbole is more commonly found in folk epics and songs.

Litotes is a stylistic device that depicts objects and events by understating or diminishing them. Like hyperbole, litotes is also used across all types of literary works for specific artistic purposes. For example, in the sentence: "…hoping for

Zuhrakhon’s sharp mind to grasp knowledge faster, he didn’t notice her elephant-sized sin at home because of her thread-sized success at school.” (A. Qahhor)

The use of the suffix -day in "qilday" (thread-like) expresses litotes, as the success being that small seems unbelievable. Simultaneously, the phrase "filday gunoh" (elephant-sized sin) employs hyperbole, as the notion of a sin being that large is also unrealistic.

Metaphor. In a metaphor, two objects or phenomena are compared based on certain features. In this case, the described object is portrayed as something else, and another object is used in its place based on similarity or relation. Unlike similes, metaphors consist only of the compared image and convey both the quality and the object in a figurative sense. Considering these features, scholars have provided various definitions for metaphor. For example, Quintilian explains that, unlike simile, which compares the depicted object to something else, metaphor directly takes the name of that object. Aristotle considered simile to be metaphor with only a partial difference between them[5]. A. Bely interpreted simile as the comparison of qualities, and metaphor as the completed form of that comparison[6]. G. L. Abramovich described metaphor as a hidden form of simile, while A. T. Rubaylo stated that simile underlies metaphor. He asserted that simile is the initial form of metaphor, and metaphor emerges from simile; moreover, some metaphors can easily be converted into similes[7]. S. Usmonov, in his Uzbek language textbook, defines it as follows: “Metaphor is the use of a word that expresses one object or phenomenon to represent another, based on relative similarity between them”. Sh. Rahmatullayev further explains: “Metaphor is the transfer of the name of one object, quality, or action to another based on mutual similarity”[8].

According to the Dictionary of Literary Terms, metaphor is described as one of the main poetic transfers (tropes), which are of two types: "transfer based on similarity – metaphor, and transfer without similarity – metonymy." It is evident that in both literary and linguistic studies, metaphor is referred to as a “shortened simile,” “simile itself,” “completed form of comparison of features,” “hidden simile,” “derived from simile,” “transfer based on similarity,” and a “transfer with similarity.” All of these definitions affirm that simile underlies metaphor. However, the main point to understand is that the difference between metaphor and simile lies in their components, specifically in the presence or absence of connecting elements. For instance:

If we say Karimjon is as strong as a lion, it is a three-component full simile.

If we say Karimjon is like a lion, it becomes a two-component incomplete simile.

If we say Karimjon is a lion, we create a metaphor.

It becomes clear that a metaphor does not include all the components (object being compared, comparative quality, and image), but rather implies them within a figurative expression. It is important to emphasize that metaphor is not always expressed through a single word; it can also appear in phrases. For example: “The captains of the blue ship prepared their steel horses for the harvesters of white gold.” In this sentence:

Captains of the blue ship refers to mechanics,

White gold refers to cotton,

Steel horses refers to machines.

It is evident that each metaphor is based on a simile and is formed through abstract thinking.

Therefore, studying the linguistic features of simile as one of the stylistic devices in literary language allows us to better understand concepts such as emotional-expressive tone, figurative imagery, and tropes in literary discourse. Metaphor is one of the most powerful and widely used figurative devices in language. It not only serves as a linguistic ornament but also functions as a cognitive tool that reflects how humans understand and conceptualize the world. By comparing one thing to another, metaphors provide insights into abstract concepts and enhance the emotional depth of communication. The term metaphor comes from the Greek word metaphora, meaning "transfer" or "carrying over." This reflects the fundamental nature of metaphor, which involves transferring meaning from one domain to another. For example, when we say "the classroom was a zoo," the classroom is metaphorically compared to a zoo, evoking an image of chaos and noise.

Metaphors are integral to language, appearing in both everyday speech and literary texts. They allow speakers to describe abstract or complex ideas in concrete terms. For instance, saying "he's in the driver’s seat of his life" uses the metaphor of driving to convey the idea that someone is in control of their life. In contrast to similes, which make comparisons explicit with words like "like" or "as," metaphors imply a more direct and sometimes more profound comparison. Metaphors can be divided into different types. Conventional metaphors are common expressions that have become so familiar that we no longer notice their figurative nature. For example, phrases like "a flood of emotions" or "a heavy heart" are conventional metaphors that evoke powerful imagery. Literary metaphors, on the other hand, are often more creative and novel, used to add depth and beauty to language, as seen in poetry and novels. "The curtain of night" or "a sea of stars" are examples of literary metaphors that invoke vivid and artistic imagery. Dead metaphors are those that have

been used so frequently that their original figurative meaning has been lost, such as "the arm of a chair" or "the foot of a mountain."

The stylistic richness of metaphors lies in their ability to evoke powerful emotional responses and provide clarity to abstract concepts. They can convey complex ideas in a more relatable or tangible way. For example, in Uzbek, a metaphor like "yuragi tog‘dek" (his heart is like a mountain) is used to describe someone with immense bravery or strength. This cultural metaphor relies on the imagery of a mountain, a symbol of stability and resilience, to communicate a deep personal quality. In addition to their stylistic value, metaphors are also significant from a cognitive perspective. Scholars such as George Lakoff and Mark Johnson in their seminal work Metaphors We Live By (1980) argue that metaphors shape our understanding of the world. They are not just rhetorical tools but fundamental to how we think, process, and conceptualize our experiences. For instance, the metaphor of "time is money" reflects a Western cultural view of time as a valuable resource that should be managed efficiently. Across languages and cultures, metaphors vary in their expressions and imagery, but they often share similar conceptual structures. In English, the metaphor "time flies" suggests that time passes quickly. In Uzbek, a similar expression might be "vag‘z uzildi", meaning "time is broken," which evokes the idea of time slipping away or being fragmented.

In conclusion, metaphors are a central element of both language and thought. They enrich communication by adding vividness and emotional depth, allowing speakers to articulate complex ideas in accessible and impactful ways. Studying metaphors offers valuable insights into how different cultures and languages perceive and conceptualize the world.

List of References:

1. B.V. Tomashevsky, Stylistics and Versification, Leningrad, 1952, pp. 208–209.
2. A.I. Efimov, Stylistics of Artistic Speech, Moscow, 1961, p. 491.
3. M.A. Sharipo, Structure of Comparisons, “Scientific Notes of the Tiraspol Pedagogical Institute”, Issue 5, Chisinau, 1958.
4. A. Sa’diy, Practical and Theoretical Literature Lessons (Textbook for Secondary Schools), Tashkent, 1925, pp. 21–22.
5. T.A. Timofeyeva, Metaphor as an Artistic Means in Popular Science Literature, pp. 214–215;
6. A. Bely, Symbolism, Moscow, “Musaget”, 1910, p. 78.
7. A.T. Rubaylo, Artistic Means of Language, Moscow, 1961, p. 37.
8. U. Tursunov, J. Mukhtorov, Sh. Rahmatullayev, Modern Uzbek Literary Language, Tashkent, 1965, p. 125.

9. I. Rybakova, On the Lexical-Stylistic Description of the Epithet, 1964, Issue 52, and others.
10. H. Hamidiy, Sh. Abdullayeva, S. Abdurahimova, Dictionary of Literary Terms, Tashkent, 1967, p. 228.
11. G.L. Abramovich, Introduction to Literary Studies, Moscow, 1953, p. 119.
12. M. Mirzayev, S. Usmonov, I. Rasulov, Uzbek Language, Tashkent, 1962, p. 424.

UDK:811.512.133’322+821.512.133.09-1

ABDULLA ORIPOV, MUHAMMAD YUSUF, SHAVKAT RAHMON SHE’RIYATI KORPUSINI YARATISHNING AHAMIYATI

Maxammadiyeva Diyora Raxmatovna
Tayanch doktarant
diyoramaxammadiyeva@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada dunyo tilshunosligida korpus lingvistikasi sohasining taraqqiyoti, rivojlangan mamlakatlar korpuslari va undan foydalanish samaradorligi haqida mulohazalar bayon etilgan. Xususan, Rus tili milliy korpusi tarkibida yaratilgan poetik asarlar korpusiga batafsil to‘xtalib, o‘tilgan. Xuddi shunday tajribalar asosida o‘zbek tilshunosligida ham xalqimizning sevimli shoirlari Abdulla Oripov, Muhammad Yusuf, Shavkat Rahmon she’riy asarlari korpusini yaratish masalasi kun tartibiga qo‘yilgan. Poetik asarlar korpusiga bo‘lgan ehtiyoj, nafaqat amaliy tilshunoslik va adabiyotshunoslik uchun ahamiyatlidir, balki leksikografiya, lingvostatistika, lingvokulturologiya sohalarining rivojlanishi uchun ham foydalidir. Ushbu korpusni yaratish yo‘lida amalga oshirilgan ishlar yuzasidan daslabki xulosalar va tahlillar taqdim etilmoqda.

Abstract. This article presents considerations on the development of the field of corpus linguistics in world linguistics, corpora of developed countries and the effectiveness of its use. In particular, the corpus of poetic works created as part of the Russian language national corpus is discussed in detail. Based on similar experiences, the issue of creating a corpus of poetic works of our people's favorite poets Abdulla Oripov, Muhammad Yusuf, Shavkat Rahmon has been put on the agenda in Uzbek linguistics. The need for a corpus of poetic works is not only important for applied linguistics and literary studies, but is also useful for the development of the fields of lexicography, linguostatistics, and linguoculturology. Preliminary conclusions and analyses of the work carried out to create this corpus are presented.

Абстрактный. В статье излагаются соображения о развитии направления корпусной лингвистики в мировой лингвистике, корпусах развитых стран и эффективности их использования. В частности, подробно рассматривается корпус поэтических произведений, созданных в рамках национального корпуса русского языка. На основе подобного опыта в узбекской лингвистике на повестку дня выдвинут вопрос о создании корпуса поэтических произведений любимых поэтов нашего народа Абдуллы Орипова, Мухаммада Юсуфа и Шавката Рахмона. Потребность в корпусе поэтических произведений важна не только для прикладной лингвистики и

литературоведения, но и полезна для развития таких направлений, как лексикография, лингвостатистика и лингвокультурология. Представлены предварительные выводы и анализы проделанной работы по созданию данного корпуса.

Kalit so‘zlar. *Korpus lingvistikasi, Rus tili milliy korpusi, poetik korpus, Korpus yaratishning texnologik bosqichlari, asarlarni korpusga tayyorlash jarayoni.*

XXI asrda texnologiyalar shu qadar tez sur’atlar bilan rivojlanib, takomillashib bormoqdaki, barcha tarmoqlar singari biz tilshunoslik sohasi vakillari ham lingvistik tadqiqotlar qilishning texnologik usullarini o‘rganish va yaratishga doimo tayyor turishimizni zamonning o‘zi taqozo qilmoqda. Korpus lingvistikasi o‘zbek tilshunosligida yangi paradigma hisoblanib, oxirgi yillarda ushbu yo‘nalish faol sohalardan biriga aylandi. Til korpuslarining paydo bo‘lishi tilshunoslarga til faktlarini tahlil qilish uchun yangi kuchli vositani berdi. Amerika, Yevropa, Rossiya va boshqa rivojlangan mamlakatlar amaliy tilshunosligida til korpuslari o‘tgan asrning 70-80-yillarida yaratila boshlangan edi. Oradan bor-yo‘g‘i yarim asr vaqt o‘tishiga qaramay, bugun dunyo tillarining ham og‘zaki, ham yozma matnlarini o‘zida jamlagan yirik til korpuslari paydo bo‘ldi. Xususan, Amerika milliy korpusi, Britaniya milliy korpusi, Lankaster Oslo/Bergen korpusi, London-Lund korpusi, Og‘zaki amerikacha ingliz tili korpusi, Rus tili milliy korpusi, Rus tilining Upsala korpusi shular jumlasidandir. Katta hajmdagi matnlarni o‘zida jamlagan, qo‘shimcha leksik, morfologik, grammatik, semantik jihatdan ishlov berilgan til korpuslari yordamida tadqiqotlarni amalga oshirish qulay va ishonchlidir. Raqamli texnologiyalar rivojidagi inqilobiy o‘zgarishlar har bir sohaga, xususan, tilshunoslik va unga yondosh sohalarga ham sezilarli ta’sir ko‘rsatdi. Xususan, bugungi kun tilshunosligining muhim sohalaridan biri korpus lingvistikasi kompyuter tilshunosligi va amaliy tilshunoslik sohalarining kesishgan nuqtasida paydo bo‘ldi.

O‘zbek tilshunosligida ham keyingi yillarda Til birliklarini lingvostatiktik tadqiq etishga qaratilgan ishlar[1,2,3,4,5,6], til korpuslarini tuzish metodikasi va texnikasi tadqiqiga bag‘ishlangan tadqiqotlar ko‘paydi[7,8,9,10,11,12,13,14,15]. Ular asosida bir necha korpuslar ishlab chiqildi. Ammo turli o‘zbek shoirlari asarlari korpusini tuzish va unga asoslangan lingvostatistik tadqiqot olib borish dolzarbligicha qolmoqda. Abdulla Oripov, Muhammad Yusuf va Shavkat Rahmon she’riyati korpusini tuzish mazkur ijodkorlar madaniy merosini kelajak avlodga yetkazish, internet tizimi yoki axborot qidiruv tizimlarida sifatli saqlash, qayta ishlash, ular ustida turli tadqiqotlar olib borishda dolzarblik kasb etadi.

Hozirgi vaqtda korpus – texnik taraqqiyotning bir qismi, zarur til birliklariga oid namunalarni topish uchun qulay vosita, shuningdek, til o‘rganish va o‘rgatishning yangi shakli hamdir.

Jahondagi eng yirik korpuslardan biri Rus tili milliy korpusi ajoyib korpus interfeysi va mukammal korpus annotatsiyasi bilan ahamiyatlidir. Rus tilining Milliy korpusi 2004-yil aprelidan boshlab, Internetda www.ruscorpora.ru sayti [15]da mavjud bo‘lib, hozirgi vaqtda u 133554 ta matn, 389 milliondan ortiq so‘zlarni o‘z ichiga oladi. Korpusga XVIII asr boshidan XXI asr boshlarigacha bo‘lgan tarixiy, publisistik, ijtimoiy, og‘zaki ko‘rinishdagi matnlar kiritilgan. Ushbu korpusda she’riy matnlar korpusi alohida bo‘limni tashkil etib, 103626 matn, 14 milliondan ortiq so‘zlardan iborat. She’riy matnlarni rus tili milliy korpusi tarkibiga joylashtirish ishlari Rossiya Fanlar akademiyasining Rus tillari Instituti jamoasi tomonidan 2006 yildan boshlab amalga oshirilgan. V.A.Plungyan rahbarligidagi Rossiya Fanlar akademiyasi Falsafa instituti ko‘magida “Rus tilining milliy korpusi doirasida “Rus she'riyati korpusi” axborot tizimini yaratish” loyihasi doirasida she’riy matnlarni to‘plash va korpus yaratish ishlari boshlangan. She’riy matnlarni belgilash tamoyillari V. A. Plungyan tomonidan E. A. Grishina, K. M. Korchagin, N. V. Pertsov, A. E. Polyakov va D. V. Sichinava ishtirokida ishlab chiqilgan. She'riy matnlarni to‘plash va belgilash bo‘yicha eng asosiy ishlar T. A. Arxangelskiy, D. A. Ivanova, K. M. Korchagin, B. V. Orexov, D. V. Sichinava va K. A. Florinskiylar tomonidan amalga oshirilgan. Korpusni boyitishda L. A. Barkova, N. K. Bogomolova, A. V. Bulannikov, I. S. Bulkina, A. A. Vasyatkina, E. A. Grishina, E. N. Lovlya, D. A. Paramonova, A. V. Praxt, D. E. Smirnov, D. A. Ershler kabi tadqiqotchi-olimlar ham faol ishtirok etishgan. Ishni muvofiqlashtirish va she’riy subkorpusning belgilarini tahrirlash D. V. Sichinava ishtirokida V. A. Plungyan va K. M. Korchagin tomonidan amalga oshirilgan [16]. 2006- yildan bugungi kungacha rus tili milliy korpusi tarkibida She’riy matnlar korpusi rivojlanib, boyitilmoqda. Ushbu korpus XVIII boshlab bugungi rus zamonaviy she’riyati namunalari va ijodkorlarning she’riy tarjimalari bilan boyitilgan. Manbalar mualliflarning eng to‘liq va mukammal nashrlari (shoir shaxsiy kutubxonasi, mukammal asarlari) materiallariga asoslangan. Korpusga yosh avlod shoirlarining jamoaviy antologiyalari ham kiritilgan. Odatiy semantik va morfologik belgilarga qo‘shimcha, meta-belgilashning bir qismi sifatida maxsus versifikatsion belgi taqdim etiladi.

O‘zbek tilshunosligida esa she’riy matnlar korpusini tuzish ishlari kun tartibidagi birinchi masalalardan biridir. Bunda korpus tuzuvchisi oldidagi dastlabki qadam – davri, maqsadi, izohlanishi, matn hajmi, izoh xususiyatiga ko‘ra yaratilajak korpus turini va korpus materialini aniqlashtirib olishdan iborat. Professor Sh.Hamroyevaga ko‘ra[17] Korpus tuzishning texnologik jarayoni bir necha bosqichlarni qamrab oladi. 1. Belgilangan manbaga muvofiq holda matnning korpusga kirishini ta’minlash. 2. Matnni avtomatik o‘qilish shaklida qayta ishlash, ya’ni korpusga kiritiladigan elektron shakldagi matn turli usul bilan olingan bo‘lishi mumkin: qo‘lda terilgan, skanerlangan, mualliflik nusxasi, hadya, ayirboshlash, internet, nashriyotlar tomonidan korpus tuzuvchisiga beriladigan orginal maketlar.

3. Tahlil, matnga dastlabki ishlov berish bosqichida turli manbalardan qabul qilingan matnlar filologik tekshiruv, tahrirdan o‘tadi. 4. Konversiyalash, grafematik tahlil, ba’zi matnlar qayta kodlashtirish jarayoni amalga oshadigan ilk mashina ishlovidan qayta-qayta o‘tadi, nomatniy qismlar, jadvallar, rasmlar o‘chiriladi yoki o‘zgartiriladi. Matndagi bo‘g‘in ko‘chirish, chegaralar bekor qilinadi, tire va boshqa belgilar bir xilligiga erishiladi. 5. Grafematik tahlil korpusga kiruvchi matnni qismga (so‘z, bog‘lovchi) ajratish, nomatniy elementni o‘chirish kabi amallarni bajaridan iborat. 6. Nostandart (noleksik) elementni belgilash, rasmiylashtirish, maxsus matniy elementni (qisqartma asosida yozilgan nom (ism, familiya), boshqa alifboda yozilgan o‘zlashma leksema, rasmga berilgan izoh, nom, zarvaraq, adabiyotlar ro‘yxati va b.) bir xil me’zon asosida qayta ko‘rib chiqish. Ammo she’riy asarlar uchun korpus yaratish jarayinida yuqorida sanab o‘tilgan bosqichlardan tashqari yana ayrim qo‘shimcha ishlar amalga oshirilishi maqsadga muvofiq. Hozirda bizning tadqiqotimiz korpus tuzishning 1-2-3-texnologik jarayonlari yakunlangan bo‘lib, dastlab O‘zbek xalqining buyuk shoirlari Abdulla Oripov, Muhammad Yusuf, Shavkat Rahmon ijodiga mansub asarlar Ziyo.net kutubxonasidan[18] elektron shaklda pdf formatda kompyuter xotirasiga saqlab olindi. Abdulla Oripovning 26 ta (shoirning dastlabki to‘plamlari tarkibidagi she’rlar mukammal asarlari to‘plamida takrorlanganligi sababli tadqiqot obyekti sifatida 8 jildli mukammal asarlari to‘plamini oldik). Muhammad Yusufning 13 ta, Shavkat Rahmonning 5 ta she’riy to‘plami mavjud ekanligi aniqlandi. Pdf formatidagi barcha kitoblar PDF to Word Converter dasturi orqali docx format holatiga keltirildi, Bexato telegram boti orqali kitoblar kril alifbosidan lotin alifbosiga o‘girildi, ushbu jarayonda yuzaga kelgan texnik va imloviy kamchiliklar filologik tekshiruv va tahrirdan o‘tkazilib, she’riy matnlar nashr holati bilan bir xillikka erishildi. Asarlardagi nashriyot nomi, kirish so‘zlar, mundarija kabi asasiy matn tarkibiga kirmaydigan qismlar olib tashlandi. Abdulla Oripov, Muhammad Yusuf, Shavkat Rahmon asarlari va ularning umumiy hajmini (sahifada) quyidagi jadvalda keltirdik.

T/R	Abdulla Oripov		Muhammad Yusuf		Shavkat Rahmon	
	Asar	Sahifalari	Asar	Sahifalari	Asar	Sahifalari
1.	Mukammal asarlari to‘plami 1-jild	428	Ko‘nglimda bir yor	32	Gullayotgan tosh	62
2.	Mukammal asarlari to‘plami 2-jild	492	Bulbulga bir gapim bor	54	Hulvo	63
3.	Mukammal asarlari to‘plami 3-jild	382	Erka kiyik	70	Saylanma	215
4.	Mukammal asarlari to‘plami 5-jild	399	Halima enam allalari	36	Uyg‘oq tog‘lar	190

5.	Mukammal asarlari to‘plami 6-jild	382	Iltijo	60	Yurak qirralari	107
6.	Mukammal asarlari to‘plami 7-jild	299	Ishq kemasi	220		
7.	Mukammal asarlari to‘plami 8-jild	322	Iymon qo‘shig‘i	75		
8.			O‘zingdan qo‘ymasin xalqim	102		
9.			Saylanma	366		
10.			Uyqudagi qiz	206		
11.			Tanish teraklar	66		
12.			Xalq bo‘l elim	189		
13.			Yolg‘onchi yor	151		
jami	7 ta to‘plam	2704 sahifa	13 ta to‘plam	1627 sahifa	5 ta to‘plam	637 sahifa

Korpus yaratishda amalga oshirilishi kerak bo‘lgan dastlabki matn to‘plash va tahrir bilan bog‘liq texnik ishlar yuqoridagi ketma-ketlikda amalga oshiriladi.

Xulosa o‘rnida shuni aytish mumkinki, Dunyo tajribasida korpus tildan foydalanishda uning statistik analizi, tabiiy tilni qayta ishlash dasturiy ta’minoti, leksik resurslarni yaratish, til o‘qitish va boshqa qator maqsadlarda qo‘l kelmoqda. O‘zbek adabiyotining buyuk shoirlari Abdulla Oripov, Muhammad Yusuf, Shavkat Rahmon o‘zining betakror she’riyati va uslubiga ega, shuning uchun ularning asarlari asosida yaratilgan she’riy korpus nafaqat o‘zbek adabiyoti, she’riyatini chuqurroq tushunishga, balki zamonaviy poetikaning rivojlanishini o‘rganishga ham katta imkoniyat yaratadi va o‘zbek tilshunosligida korpus lingvistikasi yo‘nalishida amalga oshirilayotgan tadqiqotlarning davomi sifatida ularni to‘ldiradi.

Foydalanilgan adabiyotlar:

1. Бабанаров А. Разработка принципов построения словарного обеспечения турецко-русского машинного перевода: Автореф. дисс… канд. фил. наук. – Ленинград, 1981.;

2. Айымбетов Н.К. Опыт лингвостатистического анализа лексики и морфологии каракалпакского публицистического текста: Автореф. дисс… канд. фил. наук.– Ташкент, 1987.;

3. Каримов С., Қаршиев А., Исроилова Г. Абдулла Қаҳҳор асарлари тилининг луғати. Алфавитли луғат. Частотали луғат. Терс луғат. – Тошкент, 2007. – 420 б.;

4. Ризаев С. Ўзбек тилининг лингвостатистик тадқиқи: Фил.фан.док.дис…автореф. – Тошкент, 2008.;

5. Ўринбоева Д.Б. Ўзбек фольклори матнларининг лингвостатистик тадқиқи. – Тошкент: Фан, 2010.

6. Муҳаммедова С. Ҳаракат феъллари асосида компьютер дастурлари учун лингвистик таъмин яратиш. Методик қўлланма. – Тошкент, 2006. – 45 б.

7. Хамроева Ш.М. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Фил.фан.бўйича фалсафа доктори (PhD)…дис. – Бухоро, 2018. – 250 б.;

8. Тоирова Г.И. Ўзбек тили миллий корпусини яратишнинг назарий ва амалий масалалари. Филология фанлари доктори (DSc) диссертацияси автореферати. – Бухоро, 2021.

9. Ахмедова Д. Атов бирликларини ўзбек тили корпуслари учун лексик-семантик теглашнинг лингвистик асос ва моделлари Фил.фан.бўйича фалсафа доктори (PhD) дис. – Бухоро, 2020. – 40 б.;

10. Эшмуминов А.А. Ўзбек тили миллий корпуси синоним сўзлар базасини яратиш асослари: Фил.фан.бўйича фалсафа доктори (PhD) дис. – Термиз, 2019. – 140 б.

11. Атабоев Н. Инглиз тили корпусининг функционал хусусиятлари (COCA мисолида) . Дис. Автореферат. 15- бет.

12. Begmatova G. O‘zbek milliy korpusida idiomalar bazasini yaratish. PhD dissertatsiya. 2021

13. Xidirov O. Milliy korpus uchun parsing dasturi yaratishning lingvistik asoslari. .PhD dissertatsiya 2021.

14. Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (Содда гаплар мисолида). Филол.фан.бўйича фалсафа доктори (PhD)…дис. афтореф. – Тошкент, 2018.

15. www.ruscorpora.ru sayti

16. https://ruscorpora.ru/corpus/poetic

17. Хамроева Ш. Корпус лингвистикаси атамаларининг қисқача изоҳли луғати”. –Тошкент: Камалак, 2018.

Ziyo.net kutubxonasi.

UDK: 81'33

O‘ZBEK TILI KORPUSINI ANNOTATSIYALASH VA TEGLASH (UMUMIY QOIDA VA TIZIMLAR)

Samatboyeva Madina To‘lqinjon qizi,
tayanch doktorant
samatboyevamadina@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya: Ushbu maqolada O‘zbek tili korpuslarini tuzish jarayonida amalga oshiriladigan **annotatsiyalash** va **teglash** ishlari, ularning umumiy qoidalari hamda qo‘llanilayotgan tizimlar haqida so‘z boradi. Maqolada avvalo korpuslar uchun zarur bo‘lgan morfologik, sintaktik va semantik darajadagi teglash turlari bayon etiladi. O‘zbek tilining agglutinativ xususiyati, morfologik murakkabligi va erkin so‘z tartibi kabi xususiyatlari tufayli annotatsiyalash jarayonida yuzaga keladigan muammolar va ularni hal etish yo‘llari muhokama qilinadi.

Abstract. This article discusses the annotation and tagging processes carried out during the construction of Uzbek language corpora, their general principles, and the systems being used. The paper outlines various types of tagging required at the morphological, syntactic, and semantic levels. Due to the agglutinative nature, morphological complexity, and free word order of the Uzbek language, certain challenges arise during annotation, and the article explores potential solutions to these issues.

Аннотация: В данной статье рассматриваются процессы аннотирования и теггирования, проводимые при создании корпусов узбекского языка, их общие принципы и используемые системы. В статье описываются различные уровни теггирования: морфологический, синтаксический и семантический. Агглютинативная природа узбекского языка, его морфологическая сложность и свободный порядок слов создают определённые трудности в процессе аннотирования, и в статье обсуждаются пути их решения.

Kalit so‘zlar: *korpus, annotatsiya, teg(lash), NLP, razmetka, POS-tagging.*

KIRISH.

Korpusli lingvistika zamonaviy tilshunoslikning muhim yo‘nalishlaridan biri bo‘lib, unda tabiiy til materiallari to‘plami – korpuslar asosida lingvistik tadqiqotlar olib boriladi. Annotatsiya esa korpusdagi til birliklarini maxsus belgilar orqali tavsiflash jarayonidir. Annotatsiyalangan korpuslar tilning turli darajadagi xususiyatlarini o‘rganishda asosiy vosita bo‘lib xizmat qiladi. Ushbu maqolada

korpusli annotatsiya qilish tamoyillari, ularning nazariy asoslari, yetakchi olimlar ishlari hamda xalqaro amaliyotdagi annotatsiya tizimlari tahlil qilinadi.

So‘nggi yillarda tabiiy tilni qayta ishlash (Natural Language Processing — NLP) sohasining jadal rivojlanishi korpus lingvistikasiga bo‘lgan ehtiyojni keskin oshirdi. Ayniqsa, kompyuter lingvistikasi va sun’iy intellekt tizimlarida til birliklarini chuqur tahlil qilish zarurati korpusli annotatsiya qilish tamoyillarini ishlab chiqishga asos yaratdi. Korpusli annotatsiya — bu til birliklariga ma’lum lingvistik yoki statistik ma’lumotlarni bog‘lash, ya’ni ularni tegishli kategoriyalar bilan belgilash jarayonidir. Ushbu maqolada korpusli annotatsiya qilish tamoyillari, shu yo‘nalishda olib borilgan tadqiqot ishlari yoritiladi.

ADABIYOTLAR SHARHI.

Leech korpus annotatsiyasining umumiy tamoyillari, POS-tagging (so‘z turkumlarini belgilash) bo‘yicha asosiy standartlarni ishlab chiqqan **Geoffrey Leech** o‘zinig *“Corpus Annotation Schemes”* asarida korpus va uni tuzish tamoyillariga atroflicha to‘xtalib o‘tadi[1].

Nancy Ide va Jean Veronis o‘zining “*Introduction to the Special Issue on Corpus Annotation” (1994) asari orqali a*nnotatsiya turlari (morfologik, sintaktik va semantik) va ularning korpuslar bilan integratsiyasi haqida muhim izlanishlar olib borgan[2].

“Foundations of Statistical Natural Language Processing” asari muallifi **Christopher D. Mann** o‘z ilmiy tadqiqot ishlarida annotatsiya va statistik modellar asosida tilni avtomatik tahlil qilish metodologiyasini ishlab chiqqan[3].

ASOSIY QISM

Annotatsiya

Annotatsiyalash (lotincha *annotatio* — izoh, tushuntirish) — bu korpusdagi matnlarga lingvistik axborot qo‘shish jarayoni. Bunda matndagi har bir so‘z yoki fraza:

- so‘z turkumi (masalan, ot, fe’l, sifat – POS-tagging),
- morfologik shakllari (yig‘iq, ko‘plik, zamon),
- sintaktik roli (gapda qanday funksiya bajaradi),
- yoki hatto semantik ma’no (mazmuni, mavzusi)

bilan belgilab chiqiladi. Matnni annotatsiyalashdan maqsad – matnni kompyuterlar uchun tushunarli va tahlilga yaroqli holga keltirish. *Annotatsiya* atamasi bir vaqtning o‘zida *razmetka* atamasi bilan yonma-yon qo‘llaniladi.

Razmetka – bu matnga belgilovchi teglar (markup tags) qo‘shish jarayoni. Bu teglar lingvistik bo‘lishi ham, bo‘lmasligi ham mumkin. Misol uchun:

- HTML razmetkasi <title>, <p> kabi teglar bilan tuziladi.
- Lingvistik korpusda esa razmetka:

  [NOUN]kitob[/NOUN] [VERB]o‘qidi[/VERB]

Ko‘rinib turibdiki, razmetka *texnik belgilar orqali* ma’lumotni ko‘rsatadi.

Annotatsiya va razmetkaning **o‘xshash** tomonlari: Ikkalasi ham *matnga qo‘shimcha ma’lumot* qo‘shadi, ikkalasi ham *korpuslar bilan ishlashda* qo‘llaniladi, ikkovining maqsadi ham – *matnni strukturaviy yoki semantik jihatdan boyitish.*

Farqlari:

№	Annotatsiya	Razmetka
1	Lingvistik tahlilga qaratilgan	Ko‘proq texnik ko‘rinishdagi belgilash
2	So‘z turkumi, morfologiya, sintaksis va semantika bilan bog‘liq	Strukturaviy yoki vizual tashkil etish uchun ishlatiladi
3	Masalan: kitob → NOUN, o‘qidi → VERB	Masalan: <bold>kitob</bold>
4	Ma’lumot ko‘pincha ko‘rinmas holatda fayl tuzilmasida saqlanadi	Ma’lumot ochiq ko‘rinishda, HTML/XML kabi formatlarda bo‘ladi

Agar siz matnning tilshunoslik jihatlariga e’tibor qaratayotgan bo‘lsangiz – bu *annotatsiyalash.* Agar siz matnning strukturaviy tashkiliga yoki formatlanishiga e’tibor qaratayotgan bo‘lsangiz – bu *razmetka.*

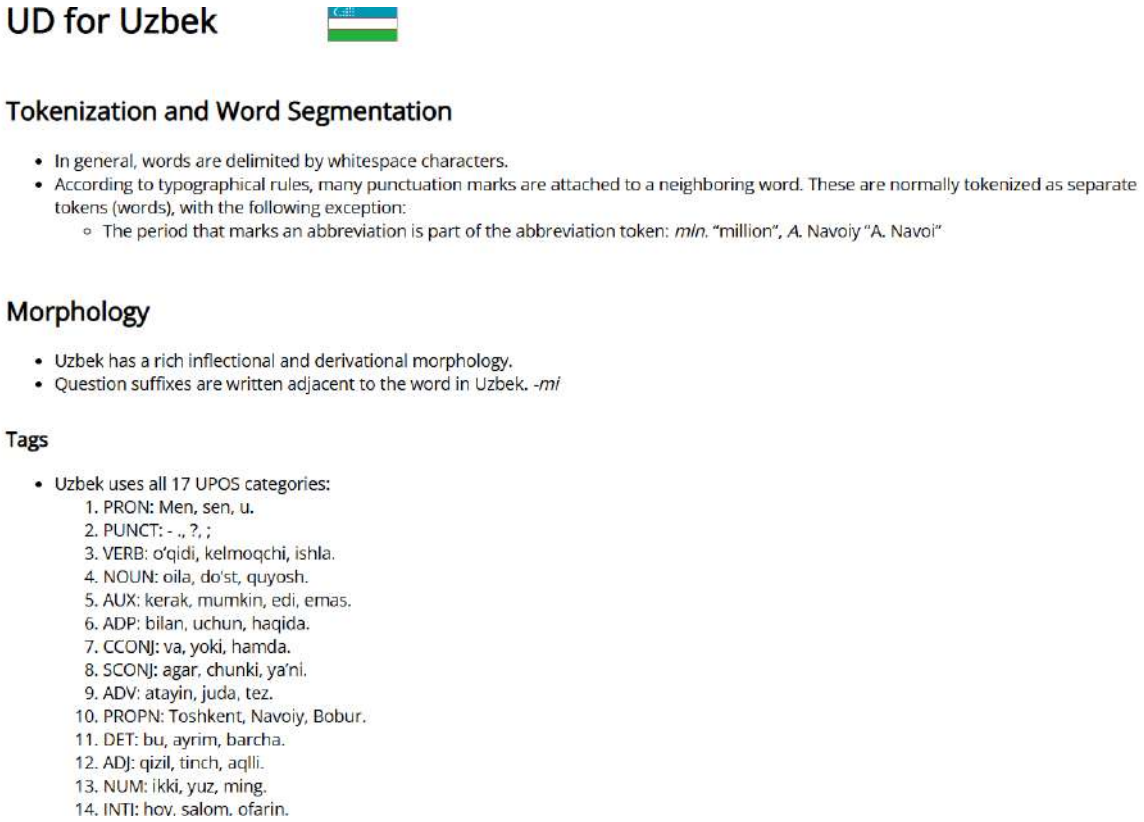
Korpus annotatsiyasi bu korpusdagi til birliklariga (so‘z, so‘z birikmalari, gap va h.k.) morfologik, sintaktik, semantik yoki pragmatik axborotlarni qo‘shish jarayonidir.

Shunday xalqaro annotatsiya korpus tizimlari mavjudki, ular juda katta hajmdagi ma’lumotlar bazasini o‘zida saqlaydi. Masalan, **“Penn Treebank (AQSh)”** (1993-yilda yaratilgan) POS-tagging va Constituency parsing (bu- gapni grammatik tuzilmasiga ko‘ra tahlil qilish usuli bo‘lib, u gapdagi so‘zlar va ularning guruhlari (ya’ni frazalar) qanday qilib bir-biriga bog‘langanini ko‘rsatadi. O‘zbekchaga tarjima qilganda**, “fraza tahlili”** yoki **“ibora tarkibiy tuzilmasini tahlil qilish”**) tizimida annotatsiya qiladi. Bu korpusda har bir so‘zga so‘z turkumi, har bir gapga esa sintaktik daraxt strukturasi biriktirilgan. Bu korpus, shuningdek, belgi (character-level) va so‘z darajasidagi (word-level) til modellari (Language Modelling) uchun ham keng qo‘llaniladi[4].

British National Corpus (BNC) - Bu ingliz tilidagi katta hajmdagi matnlar to‘plami bo‘lib, ingliz tilini o‘rganish, tadqiq qilish va tahlil qilish uchun yaratilgan. Ushbu korpus o‘z tarkibi 100 milliondan ortiq so‘zni jamlagan. POS-tagging asosida annotatsiya qilingan. Leech tomonidan ishlab chiqilgan CLAWS (Constituent Likelihood Automatic Word-tagging System) tizimi asosida annotatsiyalangan.

Yozma matnlariga: kitoblar, gazeta maqolalari, fan, san’at, din, qonun, hujjatlar va hokazolar (taxminan 90%) va og‘zaki nutq yozuvlari (spoken data): intervyular, suhbatlar, radio eshittirishlari va hokazolar (taxminan 10%) kiritilgan[5].

Shuningdek, **Universal Dependencies (UD)** (2014- hozirgi kunga qadar yangilanmoqda) – Leksik morfologik xususiyatlar, sintaktik bog‘liqliklarni tahlil qiluvchi har xil tillar uchun yagona sintaktik annotatsiya tizimidir. Ushbu tizim tarkibida 50ga yaqin tillar bazasi mavjud bo‘lib, hatto unda o‘zbek tili treebank bazasi ham kiritilgan. (Qarang: 1-rasm). O‘zbek tili treebankida 17 ta teglar ro‘yxati, morfologik tahlillar, segmentatsiya, abbreviatura, tokenlar tog‘risioda ma’lumotlar keltirilgan[6].



UD for Uzbek

Tokenization and Word Segmentation

- In general, words are delimited by whitespace characters.
- According to typographical rules, many punctuation marks are attached to a neighboring word. These are normally tokenized as separate tokens (words), with the following exception:
  - The period that marks an abbreviation is part of the abbreviation token: *mln.* “million”, *A.* Navoiy “A. Navoi”

Morphology

- Uzbek has a rich inflectional and derivational morphology.
- Question suffixes are written adjacent to the word in Uzbek. *-mi*

Tags

- Uzbek uses all 17 UPOS categories:
  1. PRON: Men, sen, u.
  2. PUNCT: - ., ?, ;
  3. VERB: o‘qidi, kelmoqchi, ishla.
  4. NOUN: oila, do'st, quyosh.
  5. AUX: kerak, mumkin, edi, emas.
  6. ADP: bilan, uchun, haqida.
  7. CCONJ: va, yoki, hamda.
  8. SCONJ: agar, chunki, ya'ni.
  9. ADV: atayin, juda, tez.
  10. PROPN: Toshkent, Navoiy, Bobur.
  11. DET: bu, ayrim, barcha.
  12. ADJ: qizil, tinch, aqlli.
  13. NUM: ikki, yuz, ming.
  14. INTI: hov, salom, ofarin.

1-rasm. Universal Dependencies (UD)da “Uzbek treebank”i

Annotatsiya qilish quyidagi tamoyillar asosida amalga oshiriladi:

1. Aniqlik va izchillik (consistency) – har bir birlik bir xil me’yorda belgilanishi kerak.
2. Qayta foydalanish (reusability) – korpus boshqa tadqiqotlar uchun ham foydalanish imkoniyatiga ega bo‘lishi lozim.
3. Standartlashtirish (standardization) – xalqaro annotatsiya standartlariga rioya qilish kerak.

4. Ko‘p darajalilik (multi-level annotation) – korpusda morfologik, sintaktik, semantik va pragmatik annotatsiyalar mavjud bo‘lishi mumkin.

Teg(lash)

Teglar barcha turdagi transkripsiyalarni (fonetik xususiyatdan nutqiy qurilmagacha), POS teglash, ma’no belgilari, sintaktik tahlil, NER obyektlar, semantik rol belgilari, vaqt va hodisalarni aniqlash, so‘zlarning sintaktik zanjirlarini, nutq darajasini o‘z ichiga olishi mumkin [7,8,9]. Lingvistik teglashning eng muhim komponenti bu tegishli teg birligi (masalan, tovush, token yoki so‘z, birikma, fragment, hujjat) bilan bog‘lanishi kerak bo‘lgan teglar va tegishli xususiyatlarni belgilaydigan teglash sxemasi hisoblanadi. Bugungi kunda korpus matnlarini lingvistik teglashni qo‘lda yoki yarim avtomatik tarzda amalga oshirish mumkin. NLP bilan bog‘liq lingvistik bilimlarning jihatlari haqida turli xil qarashlar mavjud. Umuman olganda, NLP tadqiqotchilarining aksariyati NLP bilan bog‘liq lingvistik bilimlar, hech bo‘lmaganda, leksik, sintaktik, semantikva pragmatik xususiyatlarni o‘z ichiga olishi kerak, deb hisoblashadi [10]

Korpus lingvistikasida teglardan foydalanish vaqt o‘tishi bilan rivojlanib, texnologiya, lingvistik nazariyalar va tadqiqot metodologiyasidagi yutuqlarni aks ettirdi. Korpus lingvistikasida teglar tarixini hisoblash tilshunosligining dastlabki rivojlanishi va matn korpusida tizimli va tuzilgan lingvistik annotatsiyaga bo‘lgan ehtiyoj bilan bog‘lash mumkin[11]

Shu o‘rinda annotatsiya va teglash atamalarining o‘xshash va farqli tomonlarini ko‘rib chiqish muhimdir. **Ikki atamaning o‘xshash tomonlari: H**ar ikkalasi ham **matnga qo‘shimcha axborot qo‘shish** bilan bog‘liq, har ikkalasi ham **korpuslarni boyitish,** ya’ni kompyuterga tilni "tushuntirish" uchun xizmat qiladi va har ikkalasi NLP (Natural Language Processing) va korpus lingvistikasi sohalarida keng qo‘llaniladi.

Farqlari:

№	Annotatsiya	Teglash
1	Kengroq tushuncha: har qanday qo‘shimcha axborotni matnga qo‘shish (POS, sintaksis, semantika, diskurs, his-tuyg‘u va h.k.)	Matndagi so‘zlarga maxsus teglar (labels) berish — ayniqsa grammatik yoki morfologik teglar (masalan: NOUN, VERB, ADJ)
2	Har xil darajadagi (morfologik, sintaktik, semantik) ma’lumotni qo‘shadi	Odatda faqat bir qatlamli axborot (ayniqsa POS-tagging) bilan chegaralanadi
3	Misol: matnda shaxs nomlarini, joy nomlarini, hissiyotlarni yoki referenslarni ajratish ham annotatsiyaga kiradi	Misol: “kitob” → NOUN, “o‘qidi” → VERB

4	Annotatsiyalashda razmetka texnologiyasi ishlatilishi mumkin (lekin bu shart emas)	Teglash — annotatsiyaning bir turi hisoblanadi

Teglash — annotatsiyaning bir ko‘rinishi. Ya’ni, har bir teglash annotatsiyadir, lekin har bir annotatsiya teglash emas.

Annotatsiya kengroq tushuncha bo‘lib, unga *teglash, semantik izoh, diskurs belgilarini qo‘shish, entitet aniqlash* kabi jarayonlar kiradi.

XULOSA

Korpus annotatsiyasi tilshunoslikda chuqur tahlil imkoniyatlarini yaratadi. Annotatsiyaning yuqorida ko‘rsatilgan tamoyillar asosida olib borilishi ilmiy tadqiqotlarning aniqligini ta’minlaydi. Yetakchi olimlar tomonidan ishlab chiqilgan tizimlar va korpuslar bu boradagi metodologik yondashuvlarni yanada takomillashtirishga xizmat qilmoqda. O‘zbek tilida ham shunday tizimli va standartlashtirilgan annotatsiyalangan korpuslar yaratish orqali milliy tilshunoslikni jahon miqyosida yangi bosqichga olib chiqish mumkin. Xulosa qilib aytganda, O‘zbek tili korpusini sifatli annotatsiyalash va teglash – tabiiy tilni qayta ishlash, avtomatik tarjima, matnni tahlil qilish kabi texnologiyalarni rivojlantirish uchun muhim bosqich hisoblanadi. O‘zbek tilining o‘ziga xos morfologik va sintaktik tuzilmalari ushbu jarayonni murakkablashtiradi, ammo moslashtirilgan teglash tizimlari va xalqaro standartlarga asoslangan yondashuvlar orqali bu muammolarni kamaytirish mumkin. Annotatsiya ishlari davomida yaratilgan korpuslar kelajakda o‘zbek tilini chuqur o‘rganish, raqamli texnologiyalarni rivojlantirish va sun’iy intellekt tizimlarida keng foydalanishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Leech, G. (1997). New Standards for Tagging and Parsing. In: Garside, R., Leech, G., McEnery, T. (Eds.) Corpus Annotation.
2. Ide, N., Veronis, J. (1994). Computational Linguistics, 19(2).
3. Manning, C., Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.
4. https://paperswithcode.com/dataset/penn-treebank
5. https://www.english-corpora.org/bnc/
6. https://universaldependencies.org/uz/index.html
7. Xusainova Z.Y. NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash // “O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi –Toshkent, 2022. No.1. –B.154-163.
8. Bi, P. (2018). Handbook of Linguistic Annotation. Journal of Quantitative Linguistics. https://doi.org/10.1080/09296174.2018.14244955.

9. B.Boltayevich, E.B., Mirdjonovna, H.S., & Ilxomovna, A. X. (2023). Methods for Creating a Morphological Analyzer. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 13741 LNCS. https://doi.org/10.1007/978-3-031-27199-1_4

10. Eugene. 1997. “Statistical Techniques for Natural Language Parsing”. AI Magazine 18(4):33–44

UDK: 811.512.133`42

O‘ZBEK-TURK PARALLEL KORPUSI UCHUN MATNLARNI TANLASH MEZONLARI

Xolmonova Iqbola Alisher qizi
tayanch doktoranti
iqbolabintualisher@gmail.com
ToshDO‘TAU

Annotatsiya: Ushbu maqolada O‘zbek-turk parallel korpusini yaratishda matn tanlashning lingvistik, tarjima sifati, texnik va huquqiy mezonlari tahlil qilinadi. Lingvistik va uslubiy mezonlar grammatik va semantik moslikni ta’minlasa, tarjima sifati va ekvivalentlik matn mazmunining aniq yetkazilishiga xizmat qiladi. Shuningdek, texnik va huquqiy jihatlar raqamli formatlarga moslik, mualliflik huquqlari va qonuniylik talablarini qamrab oladi. Ushbu mezonlarga asoslangan yondashuv parallel korpus sifatini oshirish va avtomatik tarjima tizimlarini takomillashtirishga xizmat qiladi.

Kalit so‘zlar: O‘zbek-turk parallel korpusi, matn tanlash mezonlari, lingvistik moslik, tarjima sifati, ekvivalentlik, texnik talablar, huquqiy normalar, avtomatik tarjima, neyron mashina tarjimasi (NMT), korpus lingvistikasi.

Til korpuslari zamonaviy lingvistika, kompyuter lingvistikasi va sun’iy intellekt sohalarida muhim tadqiqot obyekti bo‘lib, tabiiy tilni chuqur o‘rganish va avtomatlashtirilgan til tizimlarini yaratishda asosiy manba sifatida xizmat qiladi. Korpuslar yordamida turli tillarning grammatik va leksik xususiyatlarini o‘rganish, tarjima tizimlarini shakllantirish, matnlarni lingvistik jihatdan tahlil qilish imkoniyati yuzaga keladi. Ayniqsa, parallel korpuslar neyron mashina tarjima, avtomatik tarjima va ko‘p tilli ma’lumotlarni qayta ishlash jarayonlarida muhim ahamiyat kasb etadi.

O‘zbek-turk parallel korpusini shakllantirish, o‘zbek va turk tillari o‘rtasidagi lingvistik o‘xshashliklar va tafovutlarni o‘rganish, hamda ikki tilli tarjima tizimlarini rivojlantirish uchun zarur bo‘lgan muhim vositalardan biridir. Ushbu korpusning samaradorligi va ilmiy-texnikaviy qiymati ko‘p jihatdan unda foydalaniladigan matnlarning sifatiga bog‘liqdir. Shu sababli, korpus tarkibiga kiritiladigan matnlarni tanlashda aniq ilmiy mezonlarga asoslangan yondashuv talab etiladi. Matnlarni tanlash mezonlari uch asosiy tamoyilga asoslanadi. Birinchidan, lingvistik va tarjima sifati mezonlari matnlarning grammatik, leksik va sintaktik jihatdan to‘g‘ri bo‘lishini, tarjima sifatining yuqori darajada saqlanganligini ta’minlaydi. Ikkinchidan, janr va uslubiy muvozanat mezonlari turli janrlardagi matnlarning korpusda teng taqsimlanishini hamda real til materiallarini to‘liq qamrab olishni ta’minlaydi. Uchinchidan, texnik va huquqiy talablar mezonlari

matnlarning raqamli formatga mosligi, qayta ishlashga qulayligi hamda huquqiy jihatdan ruxsat etilgan materiallardan foydalanilishini kafolatlaydi[1].

Lingvistik va uslubiy mezonlar. Parallel korpuslar tabiiy tilni qayta ishlash, mashina tarjimasi va lingvistik tadqiqotlar uchun muhim ahamiyatga ega. O‘zbek-turk parallel korpusi uchun matnlarni tanlashda lingvistik va uslubiy mezonlarni hisobga olish korpus sifati va funksionalligini oshirishda muhim rol o‘ynaydi. Ushbu mezonlar matnlarning grammatik, leksik va sintaktik to‘g‘riligini ta’minlash bilan birga, ularning uslubiy xususiyatlarini ham inobatga oladi. O‘zbek va turk tillari agglutinativ tuzilishga ega bo‘lib, so‘z yasash va grammatik kategoriyalarni ifodalashda qo‘shimchalardan keng foydalaniladi. Shu sababli, korpus tarkibiga kiritiladigan matnlar grammatik jihatdan to‘g‘ri bo‘lishi talab qilinadi. Masalan: “*Men kitob o‘qiyapman”* gapi “*Ben kitap okuyorum”* tarzida bo‘lsa demak bu gap grammatik jihatdan to‘g‘ri gap hisoblanadi. Yoki aksincha “*Men kitob o‘qiydi”* va yoki *“Ben kitap okur”* kabi gaplar grammatik jihatdan noto‘g‘ri hisoblanadi. Ushbu grammatik jihatdan noto‘g‘ri gaplar yoki tarjimalar grammatik moslikni buzadi, chunki o‘zbek tilidagi *o‘qiydi* shakli umumiy hozirgi zamonni bildirsa, turk tilidagi *okur* odatiy harakatni ifodalaydi. Shuning uchun matnlarni tanlashda grammatik tuzilmalarning mos kelishiga e’tibor qaratish kerak.

So‘z va iboralarning aniq ekvivalentga ega bo‘lishi parallel korpus sifatini oshirishda muhim omildir. O‘zbek va turk tillari turkiy tillar oilasiga mansub bo‘lsa ham, ularda arabcha, forscha va yevropa tillaridan olingan so‘zlar mavjud. Shu sababli, korpus uchun tanlanadigan matnlar leksik jihatdan bir-biriga mos kelishi lozim. Maslan, “*Talaba ilm izlaydi”* gapi turkchaga “*Öğrenci ilim arar”*deb tarjima qilingan bo‘lsa, demak bu tog’ri tanlangan ekvivalent hisoblanadi. Agar mazkur gap

“*Öğrenci bilim arıyor”*deb Tarjima qilingan bo‘lsa, demak bu ekvivalent noto‘g‘ri tanlangan bo‘ladi. Chunki, bu holatda *bilim* so‘zi turk tilida *science* (ilm-fan) ma’nosida keladi, holbuki o‘zbek tilidagi *ilm* so‘zi turk tilida *ilim* shaklida qo‘llaniladi.

O‘zbek va turk tillari so‘z tartibi jihatidan o‘xshash bo‘lsa ham, ayrim hollarda sintaktik strukturalar farq qiladi. Masalan, o‘zbek tilida fe’l har doim gapning oxirida keladi, turk tilida esa ba’zan oldinga surilishi mumkin. Maslan, *“U kitob o‘qiyapti”* gapi “*O kitap okuyor”* deb qilingan bo‘lsa, demak mazkur parallel gaplar so‘z tartibi jihatidan to‘g‘ri tanlangan hisoblanadi. Agar *“U kitob o‘qiyapti”* gapi “*Okuyor o kitap”* deb tarjima qilingan bo‘lsa, korpus uchun bunday matnlarni tanlamagan ma’qul, chunki noto‘g‘ri sintaktik moslik tarjima sifatiga ham, parallel korpus sifatiga ham salbiy ta’sir ko‘rsatadi[2].

O‘zbek-turk parallel korpusi uchun turli janrdagi matnlar tanlanishi lozim. Har bir janr o‘ziga xos lingvistik va stilistik xususiyatlarga ega bo‘lib, bu tarjima modelining har xil kontekstlarga moslashishiga yordam beradi.

1-jadval. Parallel korpus uchun tanlanadigan turli til uslublaridagi matnlarga misollar

Janr	O‘zbek tilidagi matn	Turk tilidagi matn
Ilmiy matn	“Bu hodisa fizika qonunlariga asoslangan.”	“Bu olay fizik kanunlarına dayanmaktadır.”
Badiiy matn	“Quyosh ufq ortiga yashiringan edi.”	“Güneş ufkun ardına saklanmıştı.”
Rasmiy matn	“Mazkur qaror qabul qilindi.”	“Bu karar kabul edilmiştir.”
So‘zlashuv uslubidagi matn	“Nima gap?”	“Ne haber?”
Publitsistik matn	“Hozirgi zamonning eng katta muammosi – ekologiya.”	“Günümüzün en büyük sorunu - ekoloji.”

Matnlarni tanlashda janrlarning bir xil muvozanatda taqsimlanishiga e’tibor berish kerak, aks holda korpus faqat bitta uslubga xos xususiyatlarni o‘rganish imkonini beradi.

O‘zbek-turk parallel korpusi uchun matnlarni tanlashda lingvistik va uslubiy mezonlarni hisobga olish muhim ahamiyatga ega. Grammatik to‘g‘rilik, leksik-semantik moslik va sintaktik strukturalarning mos kelishi tarjima sifatini oshirishga xizmat qiladi. Janr va uslubiy tafovutlarni e’tiborga olish esa korpusdan olingan ma’lumotlarning turli kontekstlarda qo‘llanilish imkoniyatini kengaytiradi. Shunday qilib, lingvistik va uslubiy jihatlarga asoslangan matn tanlash metodologiyasi O‘zbek-turk parallel korpusining sifati va samaradorligini oshirishga xizmat qiladi.

Tarjima sifati va ekvivalentlik mezonlari. Parallel korpuslar neyron mashina tarjima, qoida asosidagi tarjima va statistik mashina tarjimasi kabi texnologiyalarni rivojlantirishda asosiy manba hisoblanadi. Korpus tarkibidagi matnlarning sifati tarjima tizimlarining samaradorligiga bevosita ta’sir ko‘rsatadi. Shu sababli, parallel korpus yaratishda matnlarni tanlashda tarjima sifati va ekvivalentlik mezonlarini hisobga olish muhim ahamiyatga ega. Ushbu mezonlar tarjimalarning aniq va tabiiy bo‘lishini, tarjima birliklarining semantik va pragmatik jihatdan to‘g‘ri kelishini hamda tarjima jarayonida ma’lumot yo‘qolishining oldini olishni ta’minlaydi. O‘zbek-turk parallel korpusidagi matnlar har xil tarjima usullaridan foydalangan holda yaratilgan bo‘lishi mumkin:

- To‘g‘ridan-to‘g‘ri tarjima – Matn bevosita tarjima qilingan bo‘lib, tabiiy va aniq ekvivalentlarni o‘z ichiga oladi.
- Oraliq til orqali tarjima – Matn dastlab boshqa tilga tarjima qilingan va keyin turk tiliga yoki o‘zbek tiliga o‘girilgan (masalan, ingliz tilidan tarjima). Bunday tarjimalar ko‘pincha aniq bo‘lmaydi.

- Mashina tarjimasi yordamida yaratilgan matnlar – Sun’iy intellekt yordamida tarjima qilingan, biroq inson tomonidan tahrir qilinmagan matnlar bo‘lishi mumkin.

Korpusga kiritiladigan matnlar asosan to‘g‘ridan-to‘g‘ri tarjima qilingan yoki inson tomonidan tekshirilgan bo‘lishi lozim[3].

Yuqori sifatli tarjima matnning ma’no jihatidan to‘g‘ri kelishidan tashqari, uning tabiiy va ravon bo‘lishini ham talab qiladi. O‘zbek va turk tillari struktura jihatidan o‘xshash bo‘lsa-da, ba’zan sintaktik farqlar yuzaga keladi. Maslan, o‘zbek tilidagi *“U juda aqlli inson”* gapi turk tiliga *“O çok zeki bir insan”* deb Tarjima qilingan bo‘lsa, demak bu gap parallel korpus uchun to‘g‘ri tanlagan moslik hisoblanadi. “*O çok akıllı bir insan”* deb tarjima qilingan bo‘lsa, bu ekvivalentni korpus uchun tanlamagan ma’qul, chunki bu holatda *aqlli* so‘zi turk tilida *akıllı* sifatida tarjima qilinsa ham, *“zeki”* shakli aniqroq va tabiiyroq hisoblanadi.

Fan, texnika va huquq sohalarida ishlatiladigan matnlarda terminologiyaning bir xilligi muhim hisoblanadi[4]. Masalan, o‘zbek tilidagi “*Elektr toki qarshilik orqali o‘tganda energiya ajralib chiqadi.”* gapi turk tiliga “*Elektrik akımı dirençten geçtiğinde enerji açığa çıkar.”* deb tarjima qilingan bo‘lsa, terminlar to‘g‘ri tanlangan gap hisoblanadi. Agar *“Elektrik akışı mukavemetten geçtiğinde güç ayrılır.”* deb tarjima qilingan bo‘lsa, *mukavemet* so‘zi texnik termin sifatida *“direnç”* so‘zi o‘rnida noto‘g‘ri qo‘llanilgan gap hisoblanadi. Shu sababli, texnik terminlar uchun maxsus lug‘atlardan foydalanish kerak.

O‘zbek-turk parallel korpusida tarjima sifatining yuqori bo‘lishi uchun matnlarni tanlashda ekvivalentlik va tarjima sifati mezonlariga rioya qilish muhim ahamiyatga ega. Grammatik, leksik va sintaktik jihatdan to‘g‘ri tarjimalar neyron mashina tarjimasi va boshqa tarjima tizimlarining samaradorligini oshiradi. Formal va dinamik ekvivalentlikni ta’minlash esa tarjimalarning tabiiy va tushunarli bo‘lishiga yordam beradi. Shu sababli, parallel korpus yaratishda lingvistik tahlil va tarjima sifatini baholash mexanizmlaridan foydalanish zarur[5].

Huquqiy mezonlar. Parallel korpuslar, ayniqsa mashina tarjimasi tizimlarini yaratishda muhim resurslar hisoblanadi. Korpusda ishlatiladigan matnlar aniq, to‘g‘ri va huquqiy jihatdan ishonchli bo‘lishi kerak. O‘zbek-turk parallel korpusini yaratishda, faqat huquqiy matnlar emas, balki umumiy matnlarni tanlashda ham huquqiy va normativ mezonlarga rioya qilish zarur. Bu talablar nafaqat tarjima sifatini oshiradi, balki matnning mazmunini ham aniq va to‘g‘ri ifodalashni ta’minlaydi. Korpusda ishlatiladigan har bir matn, ikki til o‘rtasida tushunmovchiliklarni kamaytirish va huquqiy tizimlarning farqlarini hisobga olishga xizmat qilishi lozim[6].

Matnlarni tanlashda eng birinchi mezon, tanlangan matnning qonuniy jihatdan yasalganligi va amaldagi qonunlar bilan mos kelishidir. Har bir mamlakatning qonunlar, normativ hujjatlar va yuridik tizimiga mos matnlarni

tanlash, tarjima jarayonining ishonchliligini ta’minlashda muhim o‘rin tutadi. Bu, ayniqsa, milliy qonunlarni ifodalovchi, iqtisodiy, ijtimoiy va siyosiy matnlar uchun dolzarbdir. Shu bilan birga, masalan, xalqaro huquq, iqtisodiy bitimlar yoki boshqa xalqaro kelishuvlar haqida gapirganda, har bir davlatning o‘ziga xos qonunlariga amal qilish zarur. Misol uchun o‘zbek tilida: *“Shartnoma tomonlar o‘rtasidagi kelishuv asosida tuziladi.”* gapi turk tiliga *“Sözleşme taraflar arasındaki anlaşma temelinde yapılır.”* tarzida tarjima qilingan bo‘lishi lozim. Mazkur gapda ikki tilda ham shartnoma va kelishuv tushunchalari aniq va to‘g‘ri ifodalangan. Bu o‘zgarishlarni amalga oshirishda, har bir davlatning qonuniy talablari, masalan, kelishuv yoki shartnoma qonuniyligi to‘g‘risidagi normativ hujjatlar hisobga olinadi.

O‘zbek-turk parallel korpusini tuzishda, matnlardagi huquqiy atamalar o‘rtasidagi moslikka alohida e’tibor berish zarur. Huquqiy atamalar (masalan, *“shartnoma”, “tartib”, “qonun”, “haq”, “javobgarlik”*) o‘ziga xos aniq va qat’iy ma’noga ega bo‘lishi kerak. Agar tarjimadagi atamalar yuridik jihatdan noaniq bo‘lsa, bu tushunmovchilikka olib kelishi va natijada noto‘g‘ri yuridik qarorlarga sabab bo‘lishi mumkin. Matnni tanlashda, har bir huquqiy atamaning o‘z tilidagi to‘g‘ri ekvivalentini tanlash va ular o‘rtasidagi tafovutlarni hisobga olish muhimdir. Misol uchun, o‘zbek tilida: “*Fuqarolik javobgarligi shartnomada belgilangan asosda amalga oshiriladi.”* gapi turk tiliga: *“Medeni sorumluluk sözleşmede belirlenen esaslar doğrultusunda yerine getirilir.”* deb tarjima qilingan bo‘lishi lozim. Shunda *fuqarolik javobgarligi* va *medeni sorumluluk* atamalari ikki til o‘rtasida aniq va to‘g‘ri tarjima qilingan deb hisoblaniladi. Har bir atamaning yuridik mohiyati, mamlakatlarning huquqiy tizimlaridagi ekvivalentlarga mos ravishda ko‘rsatilgan deb izohlash mumkin[7].

Korpusni tuzishda matnlarning qonuniy va normativ jihatdan muvofiqligi alohida ahamiyatga ega. Masalan, ijtimoiy muhim matnlar, iqtisodiy hujjatlar, siyosiy qarorlar yoki davlatning qonuniy me’yorlariga tegishli bo‘lgan matnlar tanlanishi kerak. Bu matnlar, har ikki til o‘rtasida muvofiqlikni ta’minlashda asosiy rol o‘ynaydi. O‘zbek-turk parallel korpusi uchun matnni tanlashda, davlat qonunlariga mos keladigan rasmiy, normativ yoki huquqiy hujjatlar asosiy mezon bo‘lishi kerak. Masalan, o‘zbek tilidagi: *“Davlat tomonidan tasdiqlangan dastur asosida moliyaviy yordam ko‘rsatiladi.”* turk tiliga: *“Devlet tarafından onaylanan program doğrultusunda mali yardım sağlanacaktır.”* Holatida bo‘lsa, har ikki tilda ham moliyaviy yordamning tasdiqlangan dastur asosida ko‘rsatilishi to‘g‘ri ifodalangan bo‘ladi. Qonuniy normativ hujjatlar va davlatning amaldagi dasturlari hisobga olinib, tarjima aniq va tushunarli bo‘lgan deb qarash mumkin.

Agar parallel korpusda xalqaro huquqiy matnlar mavjud bo‘lsa, bu matnlarning xalqaro konvensiyalar, bitimlar yoki shartnomalar bilan mos bo‘lishi kerak. O‘zbek va turk tillaridagi matnlar xalqaro huquq va kelishuvlarga muvofiq

tarjima qilinishi lozim. Bunday matnlar, yuridik noaniqliklarni bartaraf etish va har ikki mamlakatning xalqaro shartnomalari bilan mos kelishini ta’minlash uchun ehtiyotkorlik bilan tanlanadi. Masalan, o‘zbek tilida: *“Xalqaro huquq me’yorlariga asosan, davlatlar o‘zaro bitimlar tuzishda mustaqildir.”* gapi turk tilida: *“Uluslararası hukuk kurallarına göre, devletler birbirleriyle anlaşmalar yapma konusunda bağımsızdır.”* holatida bo‘lsa, demak bu misolda, *xalqaro huquq me’yorlariga asosan* va *uluslararası hukuk kurallarına göre* atamalari ikki til o‘rtasida to‘g‘ri ekvivalentlarda ifodalangan. Bunday matnlarni tanlash, xalqaro huquqning talablari va konvensiyalariga rioya qilishda muhimdir.

Matnlarni tanlashda, ularning huquqiy vaziyatga qanday ta’sir qilishi, yuridik jihatdan to‘g‘ri va tushunarli bo‘lishi ham muhimdir. Har bir matnni tarjima qilishda, uning maqsadi va kontekstiga e’tibor qaratish kerak[8]. Masalan, biror iqtisodiy bitim yoki huquqiy kelishuvni tarjima qilishda, uning tuzilishi va maqsadi aniq bo‘lishi kerak, bu esa har ikki til o‘rtasida muvofiqlikni ta’minlaydi. Misol uchun, o‘zbek tilida: *“Tashkilot tomonidan taqdim etilgan grantlar xalqaro talablar asosida ajratiladi.”* gapi turk tilida: *“Kuruluş tarafından sağlanan hibeler uluslararası standartlar doğrultusunda verilecektir.”* deyilgan bo‘lsa, bu misolda, matnning konteksti va yuridik vaziyati, xalqaro talablar va shartlarga mos bo‘lishi kerakligi ifodalangan. Bu holatda, matnning mazmuni har ikki til o‘rtasida to‘g‘ri va aniq tarzda aks ettirilgan.

O‘zbek-turk parallel korpusida matnlarni tanlashda huquqiy mezonlarga rioya qilish tarjima sifatining yuqori bo‘lishini ta’minlash uchun zarur. Matnlarning qonuniy jihatdan yasalganligi, atamalarning to‘g‘ri ekvivalentliklari, huquqiy tizimlar o‘rtasidagi farqlarni hisobga olish, xalqaro huquq va normativ talablarning inobatga olinishi bu jarayonning asosi hisoblanadi. Korpusda ishlatiladigan har bir matn, nafaqat tarjimadagi aniqlikni oshirish, balki huquqiy tizimlar o‘rtasida o‘zaro tushunishni mustahkamlashga yordam beradi.

O‘zbek-turk parallel korpusi, tilshunoslik va mashina tarjimasi sohalarida ilmiy tadqiqotlar va amaliyotda keng qo‘llaniladigan asar hisoblanadi. Korpusda ishlatiladigan matnlar tanlashda bir qancha muhim mezonlarga rioya qilish, uning sifatini oshirish va tarjima tizimlarining samaradorligini ta’minlash uchun zarurdir. Bu mezonlar lingvistik, uslubiy, texnik va huquqiy jihatlardan iborat bo‘lib, matnlarning tarjima jarayonida aniq, to‘g‘ri va tabiiy bo‘lishini ta’minlashga qaratilgan. Matnlarni tanlashda lingvistik mezonlar asosiy o‘rin tutadi. Bu mezonlar matnlarning grammatik va leksik to‘g‘riligini ta’minlashga yordam beradi. Tarjima qilinadigan matnlar o‘zbek va turk tillarining semantik va sintaktik xususiyatlariga mos kelishi kerak. Matnlarning uslubiy jihatlari ham katta ahamiyatga ega. Matnlar turli xil uslublarda yozilgan bo‘lishi mumkin: rasmiy, ilmiy, texnik, huquqiy, kundalik va boshqalar. Uslubiy mezonlar matnning to‘g‘ri tarjima qilinishini, ya’ni har bir turdagi matnning o‘ziga xos ifodalanishini ta’minlaydi. Misol uchun, ilmiy

va texnik matnlar aniq va qat’iy tilni talab qiladi, shu bilan birga, huquqiy matnlar qat’iy yuridik terminologiyani talab etadi. Matnlar tarjimasi sifatining yuqori bo‘lishi, ayniqsa, mashina tarjimasi tizimlarining muvaffaqiyatli ishlashini ta’minlashda muhimdir. Tarjima sifatini baholashda ekvivalentlik – ya’ni, asl matnning to‘liq va aniq tarjimasi – asosiy mezonlardan biridir[9]. Texnik va huquqiy matnlar parallel korpusda maxsus e’tiborga loyiqdir. Texnik matnlar ko‘pincha aniq terminologiyani va formatlarga mos kelishni talab qiladi, huquqiy matnlar esa qat’iy yuridik tilni va terminologiyani talab qiladi. Shuning uchun, bu turdagi matnlarni tanlashda ularning texnik va qonuniy jihatlariga e’tibor qaratish zarur[10].

O‘zbek-turk parallel korpusining muvaffaqiyatli ishlashi uchun matnlarni tanlashda yuqoridagi mezonlar qat’iy rioya qilinishi kerak. Har bir mezonning o‘ziga xos ahamiyati bor va ularning birgalikda qo‘llanishi tarjima tizimining samaradorligini oshiradi. Matnlarning lingvistik va uslubiy mosligi, tarjima sifati va ekvivalentligi, shuningdek, texnik va huquqiy jihatlarning hisobga olinishi parallel korpus yaratishda sifatli va ishonchli resurslarni ta’minlaydi. Bu esa, o‘z navbatida, mashina tarjimasi va boshqa til texnologiyalarining yanada yaxshilanishiga olib keladi.

Foydalanilgan adabiyotlar:

1. A. N. Jucker, H. Z. S. Translation and Interpretation: A Comparative Linguistic Approach. Oxford University Press. 2020
2. Aydin, E. Cross-cultural Translation and Technology: Challenges in the Translation of Technical Texts. Istanbul: Bogazici University Press. 2018
3. Kershaw, P. Corpus Linguistics for Translation Studies. Cambridge Scholars Publishing. 2017
4. Nilsen, A. Technical Translation: A Practical Guide. Routledge. 2019
5. Popovic, A. Introduction to Translation Studies. Springer. 2018
6. Saussure, F. Course in General Linguistics. Columbia University Press. 2011
7. Shvidkovsky, S. Legal Translation and Bilingualism. Routledge. 2015
8. Zeng, Y. Corpus-Based Translation Studies: Theory, Methods, and Applications. Springer. 2020–P. 101-109.
9. Yang, Y. The Role of Parallel Corpora in the Development of Machine Translation Systems. Springer. 2016
10. Zare, G. Baghery, M. Translation Quality Assessment: A Case Study in the Ouzbek-Turk Parallel Corpus. International Journal of Translation and Interpretation Studies, 9(3), 2017–P. 25-42.

DIALEKTAL SO‘ZLAR VA NUTQ YOZUVLARINI SAQLASHNING DASTURIY YECHIMI

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Sanjarbek Baxodirov Rahmatali o‘g‘li,
Kompyuter lingvistikasi mutaxassisligi magistranti
sanjarbahodirov9901@gmail.com
ToshDO‘TAU

Ruxsora Muftillayeva Toshmuhammad qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
ruxsoramuftillayeva@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada Django veb-freymvorkidan foydalangan holda dialektal nutq yozuvlarini saqlash modeli tahlil qilinadi. “Recording” modeli audio fayllar, ularning transkriptsiyasi, so‘zlovchining ma’lumotlari va yozib olingan joy kabi atributlarni o‘z ichiga oladi. Model dialektal tahlillar uchun qulay ma’lumotlar bazasini yaratishga imkon beradi. Maqolada modelning tarkibiy tuzilishi, maydonlarining funksional imkoniyatlari va ularning ahamiyati muhokama qilinadi. Django modellarining malumotlarni tartiblash, saqlash va qayta ishlashdagi afzalliklari ko‘rib chiqiladi. Ushbu yondashuv dialektologiya va tilshunoslik tadqiqotlarida amaliy ahamiyat kasb etadi. Modelni yanada takomillashtirish bo‘yicha tavsiyalar ham keltiriladi.

Abstract.This article analyzes a Django web framework-based model for storing dialectal speech recordings. The Recording model includes attributes such as audio files, their transcriptions, speaker information, and recording location. The model enables the creation of a structured database for dialectal analysis. The article discusses the structural composition of the model, the functionality of its fields, and their significance. It also examines the advantages of Django models in organizing, storing, and processing data. This approach has practical importance in dialectology and linguistic research. Additionally, recommendations for further improvement of the model are provided.

Аннотация. В данной статье анализируется модель сохранения диалектных речевых записей с использованием веб-фреймворка Django. Модель Recording включает в себя такие атрибуты, как аудиофайлы, их транскрипция, данные о говорящем и место записи. Модель позволяет создавать удобную базу данных для диалектных исследований. В статье рассматривается структурное строение модели, функциональные возможности ее полей и их значение. Обсуждаются преимущества использования моделей

Django для сортировки, хранения и обработки данных. Такой подход имеет практическое значение для диалектологических и лингвистических исследований. Также приводятся рекомендации по дальнейшему совершенствованию модели.

Kalit so‘zlar: *Django, dialektal nutq, ma’lumotlar bazasi, audio yozuvlar, transkriptsiya, lingvistik tahlil, modellar, tilshunoslik.*

KIRISH

Nutqshunoslik, lingvistika va dialektologiya sohalarida dialektal nutqning tahlil qilinishi va hujjatlashtirilishi katta ahamiyatga ega. O'zbek tilida, mintaqaviy dialektlar va shevalar o'zining o'ziga xos fonetik, leksik va grammatik xususiyatlari bilan boydir. Ushbu xususiyatlarni to‘plash va saqlash orqali tilshunoslar nafaqat tilning rivojlanishini, balki mintaqaviy o‘zgarishlarni ham kuzatib borish imkoniyatiga ega bo‘ladilar. Bugungi kunda bu jarayonlarni raqamlashtirish va tizimlashtirish uchun dasturiy yechimlar ishlab chiqish juda zarur.

Mazkur maqolada Django web-freymvorkidan foydalangan holda dialektal nutq yozuvlarini saqlash va tahlil qilish uchun ishlab chiqilgan “**Recording”** modeli va “**DialectWord”** modeli tahlil qilinadi. Ushbu modellar yordamida audio yozuvlar, so‘zlovchilarning ma'lumotlari, so‘zlarning mintaqaviy dialektlari, ularning transkriptsiyasi, o‘zbek tilining literatür turlariga mos keluvchi dialektal shakllari kabi ma’lumotlar saqlanadi.

Recording modeli, audio fayllar va ularning transkriptsiyalarini saqlash imkoniyatini yaratadi, shu bilan birga, so‘zlovchining shaxsiy ma'lumotlari, yoshi, jinsi va kasbi kabi atributlarni ham kiritadi. **DialectWord** modeli esa so‘zlarning turkumlarini, mintaqaviy o‘zgarishlarni va ularning o‘zbek tilidagi literatür ekvivalentlarini saqlash uchun ishlatiladi. Bu modellarning birgalikda ishlashi dialektal nutqni tizimlashtirish va ilmiy tahlil qilishda katta qulaylik yaratadi.

DialectWord modelidagi **WORD_CATEGORIES** maydoni, so‘z turkumlarini belgilashda yordam beradi va har bir dialektal so‘zning grammatik xususiyatlarini tasniflash imkonini beradi. Ushbu tasniflash tizimi tilshunoslik tadqiqotlarida tilning semantik va sintaktik strukturasi haqida muhim ma'lumotlarni olishda foydalaniladi. Modeldagi boshqa maydonlar, masalan, **region, meaning**, va **example_sentence**, so‘zlarning mintaqaviy tafovutlari va ularning amaliy qo‘llanilishiga oid ma’lumotlarni to‘plashda yordam beradi.

Ushbu yondashuv, nafaqat dialektal so‘zlarni, balki ularning fonetik va semantik xususiyatlarini ham tizimlashtirish orqali o‘zbek tilining boyligini va xilma-xilligini saqlashga xizmat qiladi. Maqolada keltirilgan kod va modellar, tilshunoslar va ilmiy tadqiqotchilarga mintaqaviy so‘zlar va nutq shakllarini tahlil

qilishda yordam berishi mumkin. Django modellarining qulayliklari va ularning ma’lumotlarni saqlashdagi ahamiyati chuqur tahlil qilinadi.

ASOSIY QISM

Modelning Tuzilishi. Recording modeli va uning maqsadi. Recording modeli dialektal audio yozuvlarni saqlash, tahlil qilish va foydalanuvchi ma’lumotlarini jamlash uchun ishlab chiqilgan.

```
class Recording(models.Model):
    dialect_word = models.ForeignKey(
        DialectWord,
        on_delete=models.CASCADE,
        related_name='recordings',
        null=True,
        blank=True
    )
    audio_file = models.FileField(upload_to='recordings/')
    speaker_name = models.CharField(max_length=100, default='Unknown')
    speaker_age = models.IntegerField(default=0)
    speaker_gender = models.CharField(
        max_length=10,
        choices=[
            ('male', 'Erkak'),
            ('female', 'Ayol')
        ],
        default='male'
    )
    speaker_occupation = models.CharField(max_length=100, default='Unknown')
    transcription_text = models.TextField(
        help_text="Audio matni transkriptsiyasi",
        default=''
    )
    recorded_date = models.DateField(default=date.today)
    location = models.CharField(
        max_length=200,
        help_text="Yozib olingan joy",
        default='Nurota'
    )
    notes = models.TextField(
        blank=True,
        help_text="Qo'shimcha ma'lumotlar"
    )

    def __str__(self):
        if self.dialect_word:
            return f"Recording of {self.dialect_word.word} by {self.speaker_name}"
        return f"Recording by {self.speaker_name}"

    class Meta:
        ordering = ['-recorded_date']
```

Ushbu model har bir yozuv haqida quyidagi ma’lumotlarni o‘z ichiga oladi:

- Audio fayl
- So‘zlovchi ma’lumotlari (ismi, yoshi, jinsi, kasbi)
- Transkriptsiya (yozib olingan nutqning matnli shakli)
- Yozib olingan sana va joy
- Qo‘shimcha izohlar

Recording modeli dialektal tadqiqotlar uchun to‘plangan audio ma’lumotlarni tartiblash, saqlash va keyinchalik analiz qilish imkoniyatini beradi.

Modeldagi maydonlarning tavsifi va ularning roli. Quyidagi jadvalda **Recording** modelining har bir maydoni va uning roli ko‘rsatilgan:(Qarang:1-jadval)

1-jadval. Modeldagi maydonlarning tavsifi

Maydon nomi	Ma'lumot turi	Tavsifi
dialect_word	ForeignKey	Yozuvga tegishli dialektal so‘zni aniqlash uchun ishlatiladi
audio_file	FileField	Audio yozuvni saqlash va yuklash imkonini beradi
speaker_name	CharField	So‘zlovchining ismini saqlaydi
speaker_age	IntegerField	So‘zlovchining yoshini belgilaydi
speaker_gender	CharField (choices)	So‘zlovchining jinsini (erkak yoki ayol) belgilaydi
speaker_occupation	CharField	So‘zlovchining kasbi haqida ma’lumot
transcription_text	TextField	Audio yozuvning matnli shakli (transkriptsiya)
recorded_date	DateField (default=date.today)	Yozib olingan sana
location	CharField	Yozib olingan joy haqida ma’lumot
notes	TextField (blank=True)	Qo‘shimcha izohlar uchun maydon

Recording modeli dialect_word maydoni orqali DialectWord modeliga bog‘langan. Bu bog‘lanish **ForeignKey** orqali amalga oshirilgan bo‘lib, har bir audio yozuv muayyan dialektal so‘zga tegishli bo‘lishi mumkin.

ForeignKey orqali Recording modeli DialectWord modeli bilan “ko‘p-birga” (many-to-one) bog‘lanadi.

Ma’lumotlarni tartiblash (ordering) va __str__ metodi

- ordering = ['-recorded_date'] — Ushbu parametr orqali yozuvlar oxirgi yozib olingan sanadan boshlab tartiblanadi.
- __str__ metodi — Model obyektini matn sifatida aks ettirish uchun ishlatiladi:

```
def __str__(self):
    if self.dialect_word:
        return f"Recording of {self.dialect_word.word} by {self.speaker_name}"
```

```
    return f"Recording by {self.speaker_name}"
```

Bu funksiya natijasida ma’lumotlar ko‘rinishi yanada tushunarli bo‘ladi.

Kodning Tahlili. Django model maydonlarining to‘g‘ri tanlanganligi.

Django model maydonlari to‘g‘ri tanlangan va ma’lumotlarning turiga qarab mos keluvchi model maydonlari ishlatilgan. Masalan(Qarang:2-jadval)

2-jadval. Django model maydonlari(fields)

Django maydoni	Ma’lumot turi
CharField	Matn (100-200 belgi)
IntegerField	Butun son (yosh)
FileField	Fayllarni yuklash (audio)
TextField	Katta hajmdagi matn (transkriptsiya)
DateField	Sana (recorded_date)

null=True, blank=True ishlatilishi va ularning farqi(Qarang:3-jadval)

3-jadval. null=True, blank=True ishlatilishi va ularning farqi

Parametr	Vazifasi
null=True	Baza darajasida maydon **NULL qiymat** olishiga ruxsat beradi
blank=True	Forma orqali to‘ldirish **shart emas**ligini bildiradi

Masalan, dialect_word maydonida null=True, blank=True belgilangan bo‘lib, bu foydalanuvchi ba’zi yozuvlarda dialektal so‘zni kiritmasligi mumkinligini bildiradi.

choices argumenti orqali jinsni tanlash. Django’da “choices” argumenti orqali foydalanuvchi faqat ma’lum qiymatlar ichidan tanlash imkoniga ega bo‘ladi. Masalan:

```
speaker_gender = models.CharField(
    max_length=10,
    choices=[
        ('male', 'Erkak'),
        ('female', 'Ayol')
    ],
    default='male'
)
```

Bu yondashuv foydalanuvchining noto‘g‘ri qiymat kiritish ehtimolini kamaytiradi.

AMALIY TATBIQ

Ushbu modeldan qanday foydalanish mumkin?

– Ma’lumotlar bazasida dialektal so‘zlar va ularning audio yozuvlarini saqlash
– O‘zbek tilining mintaqaviy dialektlarini ilmiy tahlil qilish
– Audio yozuvlar orqali lingvistik tadqiqotlar olib borish

Ushbu model lingvistik tadqiqotlar va dialektal tahlillar uchun muhim ahamiyat kasb etadi. Uning imkoniyatlari keng bo‘lib, turli sohalarda samarali qo‘llanilishi mumkin. Ayniqsa, O‘zbek tilining mintaqaviy xususiyatlarini chuqur o‘rganish, audio materiallarni to‘plash va ilmiy ma’lumotlarni tizimli tarzda saqlash imkonini beradi.

Ma’lumotlar bazasida dialektal so‘zlar va ularning audio yozuvlarini saqlash. O‘zbek tilining dialektal so‘zlarini saqlash va ularni mintaqalar bo‘yicha tasniflash tilshunoslik tadqiqotlari uchun muhim ahamiyatga ega. ***Recording*** modeli har bir dialektal so‘zga tegishli ***audio yozuvni*** saqlash imkonini beradi. Bu esa tadqiqotchilarga ***so‘zlarning talaffuzi, fonetik o‘zgarishlari va kontekstda ishlatilishini tahlil qilish*** imkoniyatini taqdim etadi.

Django yordamida foydalanuvchilar dialektal so‘zlar bazasiga **yangi yozuvlar qo‘shishi**, mavjudlarini **tahrirlashi** va ularni **mintaqalar bo‘yicha filtrlashi** mumkin. Har bir so‘z bilan bog‘liq **talaffuz namunasi (audio fayl)** ham saqlanadi, bu esa tadqiqotchilarga **real nutq namunalarini to‘plash va tahlil qilishga** yordam beradi.

Bunday tizimni yaratish orqali:

- **Dialektal lug‘atlar shakllanadi**
- **So‘zlarning talaffuzi saqlanib qoladi**
- **Mintaqaviy shevalarning fonetik jihatlari o‘rganiladi**

Misol uchun, **Recording** modelidan foydalanib, quyidagi ma’lumotlar saqlanishi mumkin:

- Toshkent shevasiga oid **“pachakilamoq”** so‘zi
- Uning adabiy o‘zbek tilidagi muqobili “**cho‘pchak aytmoq”**
- **Audio yozuv** orqali bu so‘zning talaffuzi
- **Transkriptsiya** va **lug‘aviy izoh**

O‘zbek tilining mintaqaviy dialektlarini ilmiy tahlil qilish. O‘zbekistonning turli hududlarida turlicha talaffuz qilinadigan va ishlatiladigan so‘zlarni tahlil qilish orqali ***tilshunoslik va dialektologiya*** yo‘nalishida ilmiy tadqiqotlar olib borish mumkin. Ushbu model ***so‘zlarning semantik va fonetik xususiyatlarini*** o‘rganish imkonini beradi.

Dialektal tadqiqotlar uchun bu model quyidagi jihatlarda qo‘llanilishi mumkin:

- **Mintaqalar bo‘yicha lug‘aviy farqlarni aniqlash**
- **Sheva va adabiy til o‘rtasidagi tafovutlarni o‘rganish**
- **O‘zbek tilining tarixiy rivojlanish jarayonlarini kuzatish**

Masalan, Toshkent shevasida keng tarqalgan **“havo”** so‘zi Xorazm mintaqasida **“avo”** tarzida ishlatiladi. Ushbu model yordamida bu kabi farqlar **audio materiallar** orqali saqlanib, o‘rganilishi mumkin.

Bundan tashqari, tadqiqotchilar **mintaqalar bo‘yicha dialektal so‘zlarni taqqoslab, ularning kelib chiqishiga oid xulosalar** chiqarishlari mumkin.

Audio yozuvlar orqali lingvistik tadqiqotlar olib borish. Nutqning audio shaklda saqlanishi tadqiqotchilarga ***real muloqot jarayonlarini tahlil qilish, fonetik o‘zgarishlarni o‘rganish va talaffuz farqlarini aniqlash*** imkonini beradi. Ovozli materiallardan foydalanish orqali quyidagi tadqiqotlarni amalga oshirish mumkin:

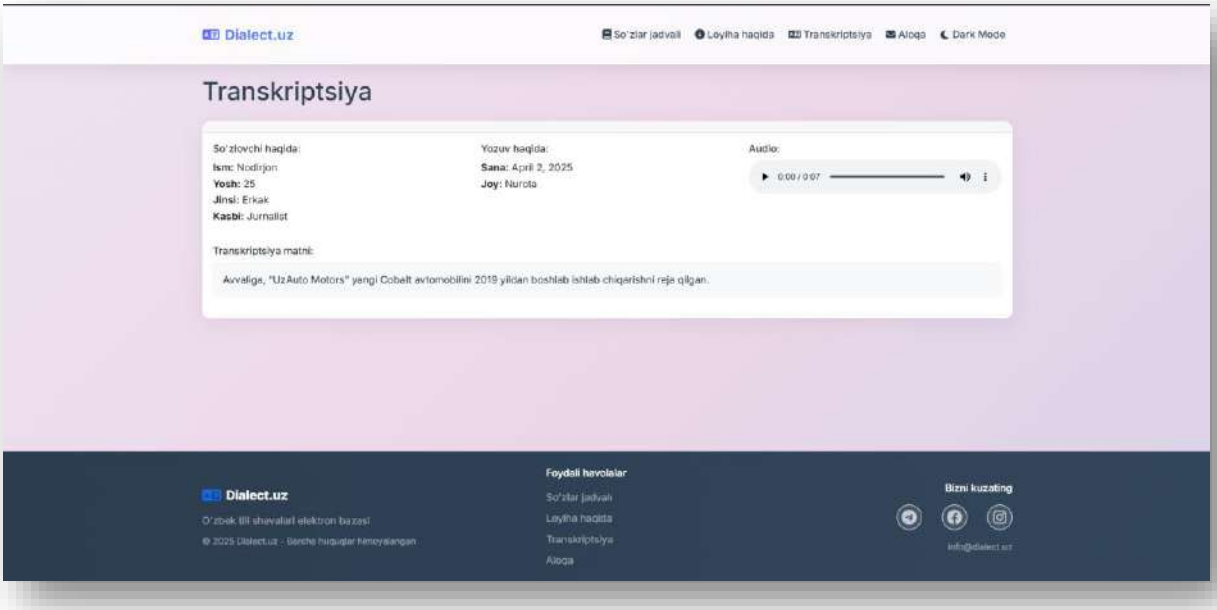
- Sheva va lahjalarning talaffuz xususiyatlarini o‘rganish
- Turli yosh guruhlari va kasblarga mansub shaxslarning nutqini tahlil qilish
- Davrlar o‘tishi bilan o‘zbek tilining o‘zgarishlarini kuzatish

Misol uchun, Recording modeli yordamida **xalq orasida keng tarqalgan iboralar va so‘z birikmalarining talaffuzi** saqlanib, ularning **zamonaviy til bilan bog‘liqligi** tahlil qilinishi mumkin.

Ushbu model nafaqat **tilshunoslar**, balki **sun’iy intellekt** va **nutqni qayta ishlash** (speech processing) sohasida ishlovchilar uchun ham muhim ahamiyat kasb etadi. AI-modellar uchun **dialektal nutqni tushunish** va **turli talaffuzlarni analiz qilish** muhim vazifalardan biri hisoblanadi.

Audio fayllarni yuklash va ularni qayta ishlash imkoniyatlari. Recording modelidagi audio_file maydoni audio fayllarni saqlash va ularni analiz qilish imkonini beradi. Django yordamida media fayllarni yuklash va ularni qayta ishlash quyidagicha amalga oshiriladi(Qarang:1-rasm)

```
recording = Recording.objects.create(
    speaker_name="Ali",
    speaker_age=30,
    speaker_gender="male",
    audio_file="recordings/sample.mp3"
)
```



1-rasm. Dasturdagi ko’rinishi

Ushbu model dialektal nutqni raqamlashtirish, saqlash va ilmiy tahlil qilish imkoniyatini beradi. Django modellaridan foydalanish orqali samarali va tizimli ma’lumotlar bazasini shakllantirish mumkin.

XULOSA

Mazkur maqolada **Recording** va **DialectWord** modellarining tuzilishi, ularning ma'lumotlar bazasida qanday ishlashi va amaliy qo‘llanilishi haqida batafsil tahlil qilindi. Ushbu modellar O‘zbek tilining dialektal xususiyatlarini saqlash, o‘rganish va tahlil qilish imkonini beradi. **Recording** modeli audio yozuvlarni saqlash va ularga tegishli til birliklarini bog‘lashga xizmat qilsa, **DialectWord** modeli dialektal so‘zlarni adabiy til shakllari bilan solishtirish, ularning ma’nosini va mintaqaviy xususiyatlarini aniqlashga yordam beradi.

Modeldagi maydonlarning to‘g‘ri tanlanishi, “**ForeignKey**” bog‘lanishlarining mavjudligi va “**ordering**” parametrining qo‘llanilishi ma’lumotlarni tartibli saqlashga va samarali ishlov berishga imkon yaratadi. Shuningdek, audio fayllarni yuklash va ularni qayta ishlash funksiyasi lingvistik tadqiqotlar uchun yangi imkoniyatlar ochadi.

Ushbu modellar dialektologiya, lingvistika va sun’iy intellekt sohalarida amaliy qo‘llanishga ega bo‘lib, tilshunoslik tadqiqotlari uchun muhim ma’lumotlar bazasini shakllantirishga xizmat qiladi. Shu bilan birga, O‘zbek tilining mintaqaviy xususiyatlarini saqlash va o‘rganish orqali milliy til merosimizni kelajak avlodlarga yetkazishga ko‘maklashadi.

Foydalanilgan adabiyotlar:

1. Холова М. ўзбек миллий шевалари корпусини тузишнинг лингвистик асослари (Бойсун тумани “ж”ловчи шевалари мисолида):Фил. фан. бўйича фалсафа доктори (PhD)...дисс. Автореф. – Термиз, 2022.
2. http://rusling.narod.ru/qqq_corp_nonslav_other.htm
3. http://www.corpusrural.es
4. https://freidok.uni-freiburg.de/proj/
5. Nordic Dialect Corpus (uio.no)
6. https://www.helsinki.fi/varieng/CoRD/corpora/Dialects/fieldwork
7. Региональная этнолингвистика (ethnolex.ru)

SHEVALAR VA MEDIA: O‘ZBEK TILINING RAQAMLI RIVOJIDA XORAZM SHEVASINING O‘RNI

Musayeva Mahina Maqsudbek qizi,
magistrant
musayevamohina41@gmail.com
Urganch davlat universiteti

Annotatsiya. So‘nggi yillarda o‘zbek tilining raqamli shakllanishi va sun’iy intellekt tizimlarida qo‘llanilishi tez sur’atlarda rivojlanmoqda. Ammo bu jarayonda hududiy shevalarning, xususan, Xorazm shevasining o‘rni va ifodalanish darajasi hali ham yetarli emas. Ushbu maqolada O‘zbek tilining shevalari, ayniqsa Xorazm shevasi, zamonaviy media platformalarida (YouTube, TikTok, Telegram va h.k.) qanday aks etayotgani tahlil qilinadi. Shuningdek, Xorazm shevasida yaratilgan kontentlarning hajmi, sifati va ularni til korpuslari yoki AI modellar uchun potensial ma’lumot manbai sifatida foydalanish imkoniyatlari ko‘rib chiqiladi. Shevalarning raqamli kontent orqali saqlanishi va rivojlanishi til siyosati va madaniy meros nuqtai nazaridan ham dolzarb ekani asoslab beriladi.

Abstract. In recent years, the Uzbek language has been rapidly developing in the digital sphere and increasingly integrated into artificial intelligence systems. However, regional dialects-especially the Khorezm dialect-remain underrepresented in this process. This paper analyzes how Uzbek dialects, particularly the Khorezm dialect, are reflected on modern media platforms such as YouTube, TikTok, and Telegram. It explores the volume, quality, and characteristics of content in the Khorezm dialect and its potential as a data source for language corpora and AI models. The importance of preserving and promoting dialects through digital media is emphasized from the perspectives of language policy and cultural heritage.

Аннотация. В последние годы узбекский язык активно развивается в цифровом пространстве и внедряется в системы искусственного интеллекта. Однако региональные диалекты, в частности хоразмский говор, остаются недостаточно представленными в этих процессах. В данной статье анализируется присутствие узбекских диалектов, особенно хоразмского, на современных медиа-платформах (YouTube, TikTok, Telegram и др.). Рассматриваются объем, качество и особенности контента на хоразмском говоре, а также его потенциал как источника данных для лингвистических корпусов и моделей ИИ. Подчеркивается значение сохранения и развития диалектов через цифровой контент с точки зрения языковой политики и культурного наследия.

Kalit so‘zlar: *Xorazm shevasi, media korpus, o‘zbek tili, raqamli tilshunoslik, shevalar, sun’iy intellekt, til korpusi.*

Raqamli texnologiyalarning tezkor rivojlanishi tilshunoslikda yangi yo‘nalishlarning shakllanishiga sabab bo‘lmoqda. Xususan, sun’iy intellekt (SI) tizimlarida tabiiy tilni qayta ishlash (Natural Language Processing – NLP) metodlari asosida til modellarini yaratish uchun katta hajmdagi matnli korpuslar talab etiladi. O‘zbek tilida bunday korpuslar shakllantirila boshlagan bo‘lsa-da, bu jarayonda asosan adabiy til asosiy manba sifatida ishlatilmoqda. Hududiy shevalar, ayniqsa Xorazm shevasi, bu jarayonda yetarlicha e’tibordan chetda qolmoqda.

Shevalar har bir tilning ichki boyligini, tarixiy qatlamlarini va madaniy xilma-xilligini ifodalovchi muhim til resursidir. Agar ularni raqamli formatda saqlab qolish va tizimlashtirish amalga oshirilmasa, bu so‘z boyligi va madaniy qatlamlar vaqt o‘tishi bilan yo‘qolishi mumkin. Shu bois Xorazm shevasining media kontentlar orqali qay darajada aks etayotgani, bu kontentlarni korpus yaratishda qanday foydalanish mumkinligi muhim tadqiqot yo‘nalishidir.

Globalizatsiya va raqamli texnologiyalar rivojlanishi natijasida tillarning ijtimoiy hayotdagi, ayniqsa internet va media platformalaridagi roli keskin o‘zgarib bormoqda. Shu nuqtai nazardan qaralganda, o‘zbek tilining raqamli makondagi o‘rni, uning ichki ko‘rinishlari, ya’ni lahja va shevalarining taqdiri ham dolzarb masalaga aylanmoqda. Chunki raqamli axborot almashinuvi faqat adabiy til orqali emas, balki jonli xalq og‘zaki nutqi - shevalar orqali ham amalga oshmoqda. O‘zbek tilining hududiy shevalari ichida Xorazm shevasi o‘zining boy lug‘aviy zaxirasi, qadimiylik bilan uyg‘unlashgan zamonaviy ifoda uslublari va raqamli kommunikatsiyada faol qo‘llanilishi bilan ajralib turadi. Ushbu maqolada Xorazm shevasining raqamli kommunikatsiyadagi roli, media vositalaridagi ifodalanish shakllari, hamda umumtil taraqqiyotiga qo‘shayotgan hissasi tahlil qilinadi. Shu bilan birga, raqamli tilshunoslik doirasida ushbu shevani o‘rganish va undan zamonaviy til texnologiyalarida foydalanish imkoniyatlari ko‘rib chiqiladi.

Xorazm shevasi fonetik, morfologik va leksik jihatdan o‘ziga xos bo‘lib, o‘zbek tilining boshqa shevalaridan farqlanuvchi ko‘plab xususiyatlarga ega. Ushbu sheva nafaqat muloqot vositasi, balki Xorazm viloyatining madaniy identitetini ifodalovchi element hamdir. Biroq bugungi kungacha bu sheva asosida yaratilgan hech qanday katta korpus yoki lingvistik resurs mavjud emas.

Zamonaviy media platformalarida ayniqsa, YouTube, Instagram, TikTok va Telegram kabi platformalar, xorazm shevasining tabiiy kontekstdagi namunalarini saqlab qolish va tahlil qilish uchun yangi imkoniyatlar yaratmoqda. Ayniqsa, Xorazmda faol blogerlar, kontent yaratuvchilar tomonidan yaratilgan vloglar, sahnalashtirilgan skitlar va suhbat janridagi videolar shevaning zamonaviy shaklini to‘g‘ridan-to‘g‘ri aks ettiradi. Ushbu materiallar og‘zaki tilning real va kontekstli misollari sifatida juda katta ahamiyatga ega.

Mazkur media kontentlar raqamli korpus yaratishda foydali manba bo‘lishi mumkin. Biroq, bu jarayonda bir nechta muammo yuzaga chiqadi:

1. *Standartlashtirish muammosi* - Xorazm shevasida yozilgan yoki talaffuz qilingan iboralarni adabiy tilga yoki fonetik yozuvga keltirish qiyin.

2. *Kontent sifati va barqarorligi* - Media platformalarda joylashtirilgan materiallar vaqtinchalik bo‘lishi, yo‘qolib ketishi yoki o‘chirilishi mumkin.

3. *Annotatsiya muammolari* - Dialektal so‘zlar, ijtimoiy jargon va kontekstual iboralarni aniqlash va ularni standart shaklda belgilash uchun lingvistik bilim va metodik yondashuv talab etiladi.

Shunga qaramay, xorazm shevasida mavjud kontentlarni tizimli yig‘ish, ularni transkripsiya qilish, matnga aylantirish va lingvistik belgilash orqali yirik media korpus yaratish mumkin. Bu korpus keyinchalik SI modellarini shevaga moslashtirish, avtomatik tarjima, ovozdan matnga konversiya, va boshqa NLP vazifalarda foydalanish imkonini beradi.

Xorazm shevasi o‘zbek tilining qadimiy va mustahkam mintaqaviy shevalaridan biri bo‘lib, u fonetik, leksik va grammatik jihatdan ancha farq qiluvchi jihatlarga ega. Bu shevada qadimgi o‘zbek tili unsurlarining saqlanib qolganligi, shuningdek, turkiy tillarning umumiy xususiyatlarini aks ettiruvchi shakl va so‘zlar uchrashi kuzatiladi. Masalan, “bormoqchiman” o‘rniga “barjaqman”, “yaxshimisiz?” o‘rniga “nichiksiz?”, “nima qilay” o‘rniga “nishatin” kabi shakllar fonetik o‘zgarishlarning yorqin namunalaridandir. Xorazm shevasining leksikasida ham ko‘plab o‘ziga xos atamalar mavjud: “sovuq” o‘rniga “suvug‘”, “uyqu” o‘rniga “uxi”, “katta” o‘rniga “ulli” va hokazo. Ushbu leksik birliklar nafaqat tilshunoslik, balki madaniyatshunoslik uchun ham katta ilmiy qiziqish uyg‘otadi. Bugungi kunda Xorazm shevasi nafaqat og‘zaki nutqda, balki ijtimoiy tarmoqlar, audio-video mahsulotlarda ham faol ishlatilmoqda.

Xalqaro tajriba shuni ko‘rsatmoqdaki, ko‘plab tillar uchun dialektal korpuslar yaratilib, ular lingvistik tadqiqotlarda ham, sun’iy intellekt modellarida ham muvaffaqiyatli qo‘llanilmoqda. Masalan, arab tili, italyan tili, nemis tili dialektlari bo‘yicha yaratilgan korpuslar tilshunoslikda muhim o‘rin egallaydi. O‘zbek tilida bu yo‘nalishda hali dastlabki bosqichlar boshlanmoqda, va Xorazm shevasi shu jarayonga kiritilishi zarur.

Xulosa qilib aytganda, shevalar, xususan, Xorazm shevasi o‘zbek tilining muhim tarkibiy qismi sifatida nafaqat madaniy merosni saqlashda, balki zamonaviy kommunikatsiya va raqamli til muhitida ham alohida o‘rin egallamoqda. Raqamli platformalarning keng tarqalishi bu shevaning ommalashuvi va yoshlar orasida qayta qadrlanishiga sabab bo‘lmoqda. Ayniqsa, media kontentlarda Xorazm shevasining faol qo‘llanishi uning ijtimoiy qadriyat, estetik vosita va identitet belgisi

sifatida mustahkamlanishiga xizmat qilmoqda. Xorazm shevasi o‘zbek tilining boyligidir. Uni raqamli asr talablariga mos tarzda saqlab qolish, o‘rganish va rivojlantirish – nafaqat tilshunoslarning, balki jamiyatning har bir a’zosining ham burchidir.

Xorazm shevasining media kontentlarda aks etishi va ularni raqamli til resurslariga aylantirish imkoniyati o‘zbek tilining raqamli rivoji uchun muhim ahamiyat kasb etadi. Bu jarayon nafaqat tilni sun’iy intellektga moslashtirish, balki madaniy boylikni saqlash vazifasini ham bajaradi. Bugun boshlangan harakatlar, ertaga xorazm shevasining yo‘qolish emas, balki rivojlanishiga xizmat qiladi. Shu sababli, xorazm shevasidagi media kontentlar tizimli o‘rganilishi, saqlanishi va korpuslashgan shaklga keltirilishi zarur.

Foydalanilgan adabiyotlar:

1. M. Abdul-Mageed, D. Kuebler, and M. Diab, “SAMAR: A System for Arabic Dialect Annotation of Social Media Data,” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, vol. 2, pp. 250–255, 2018.
2. E. Kuriyozov, D. Vilares, and C. Gómez-Rodríguez, ‘BERTbek: A pretrained language model for Uzbek’, in *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages@ LREC-COLING 2024*, 2024, pp. 33–44.
3. P. Soulos et al., "Structural Biases for Improving Transformers on Translation into Morphologically Rich Languages," *arXiv preprint arXiv:2208.06061*, 2022.
4. R. G. Yuldashovna, S. Sukhrob, K. Barno, P. Mukhayyo, and D. Shokhida, ‘Linguistic Considerations on the Functional Semantics of English Reduplicative Elements’, *Remittances Review*, vol. 8, 2023.
5. A. Vaswani et al., "Attention is All You Need," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5998–6008.
6. B. Mansurov and A. Mansurov, "UzBERT: Pretraining a BERT Model for Uzbek," *arXiv preprint arXiv:2108.09814*, 2021.
7. B. Allaberdiev, G. Matlatipov, E. Kuriyozov, and Z. Rakhmonov, ‘Parallel texts dataset for Uzbek-Kazakh machine translation’, *Data in Brief*, vol. 53, p. 110194, 2024.

NUROTA SHEVASI (QARLUQ LAHJASI) KORPUSINING MA’LUMOTLAR BAZASI

Ruxsora Muftillayeva Toshmuhammad qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
ruxsoramuftillayeva@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada milliy korpusda asta-sekin har xil korpuslarning paydo bo‘lishi, sheva korpuslarini nazariy asoslanishi, yig‘ilgan shevalardan foydalanish va tadqiqotlarni yengillashtirish yo‘lini bosqichma-bosqich bosib o‘tish, zamonaviy tilshunoslikning ushbu yo‘nalishidagi yangi korpuslarning paydo bo‘lishiga bir qadar poydevor bo‘lib xizmat qiladi. Shevashunoslikda so‘zlashuv jarayonlarini transkripsion shaklda yozib olish, shevaga oid materiallarni izohlash orqali dialektning o‘ziga xos xususiyatlarini tadqiq etish, shevalar tadqiqotiga bag‘ishlangan monografiyalar, lug‘atlar, yaratiladi hamda dialektal xaritalar tuziladi.

Annotation: This article discusses the gradual emergence of various corpora within the national corpus, the theoretical foundations of dialect corpora, and the step-by-step process of utilizing collected dialects to facilitate research. It serves as a partial foundation for the development of new corpora in this modern field of linguistics. In dialectology, the transcription of spoken language, the annotation of dialect materials, and the study of distinctive features of dialects are emphasized. Additionally, monographs, dictionaries dedicated to dialect research are created, and dialectal maps are compiled.

Kalit so‘zlar: *shevashunoslik, lug‘at, milliy korpus, dialektal korpus*

KIRISH

Dialektal korpus yaratish, yig‘ilgan ma’lumotlardan samarali foydalanish va lingvistik tahlil uchun zaruriy vosita hisoblanadi. Ushbu jarayon, foydalanuvchilarga kerakli ma’lumotlarni qidirishda vaqtni sezilarli darajada qisqartirish imkonini beradi. Nurota tumani dialektal korpusining ma’lumotlar bazasini tuzishda, ijtimoiy jihatdan tashqi dunyodan qisman uzilgan va internet tarmoqlaridan foydalanish ko‘nikmasiga ega bo‘lmagan keksa avlod vakillari ishtirok etishdi. Ushbu jarayon, shahardan uzoqda joylashgan hududlarda yashovchilar o‘rtasidagi o‘zaro suhbatlarni yozib olish orqali amalga oshirildi. Dialektal korpus uchun yig‘ilgan audio fayllar dastavval matn ko’rinishiga keltiriladi. Audio faylni matn holatiga keltirish uchun biz har xil dasturlarni ishlatib ko‘rdik. Xususan, “speech-to-text” dasturi. Bu dastur imkoniyatlari ancha yuqori. Dasturga audio faylni MP3 shaklida yuklashimiz yoki mikrofon orqali to‘g‘ridan

to‘g‘ri ovoz yozishimiz mumkin. Birinchi qadam, yuklangan faylga ishlov beradi. Ya’ni ovozdagi ortiqcha shovqinlarni olib tashlab, ovoz tiniqligini oshiradi. Keyingi bosqichda audioning matn ko‘rinishi taqdim etiladi. Matnni fayl ko‘rinishida saqlab olish imkoniyati mavjud. Dasturda 100 dan ortiq til mavjud, ammo o‘zbek tili yo‘qligi uchun bu dasturdan foydalana olmadik. Foydalangan keyingi dasturimiz “AIsha” o‘zbek tilida mukammal muloqat qiladigan sun’iy intellekt. Bu dastur ovozni matnga, matnni ovozga kabi xizmatlarini taqdim etadi. Dasturdan foydalanish barcha uchun ochiq. SST faqat 30 soniyagacha bo‘lgan MP3 formatdagi fayllarni qabul qiladi. Shuning uchun audio fayllarni 30 soniyadan qilib kesib chiqdik. Audio fayllarni dasturga yuklaganimizda bizga matn shaklida, transkripsiyasi bilan taqdim etdi. Bu dasturlar har doim ham yuqori aniqlikda ishlamaydi, ayniqsa, mahalliy lahjalarda. Xato bo‘lgan joylari qo‘lda to‘g‘irlandi. Audio fayllar matn ko‘rinishiga kelgandan so‘ng matnlar bilan ishlandi. Matnlarni alohida satrlarga bo‘lib chiqdik. Har bir satrni yonidan matnning audio variantini joylab chiqdik. Audio fayllarni satrlarga moslab kesib, satrlar bilan yonma-yon joylashtirdik. Audiolarni kesish uchun maxsus “Simple MP3 Cutter” programmasidan foydalanib audiolarni satrlarga moslab kesib chiqdik. Dasturdan foydalanish bepul. Dasturdan ovozdagi shovqinlarni kesishda ham foydalandik.

Korpus interfeysi. Interfeys − piegasse so‘zidan olingan bo‘lib, texnik va programmalash vositalari yig‘ilmasi hisoblanib, boshqarish yoki o‘lchash, hisoblashdagi turli vazifaviy qurilmalarning o‘zaro ta’sirini ta’minlab beradi. Berilayotgan axborotlarni kodlash, sinxronlash hamda nisbatan belgilangan qoida hamda kelishuvga muvofiq tayyor modullardan sistemani, qurilmalarning mexanik va elektr birikmalarini signal, ya’ni axborot shaklida to‘plashga imkon beradi[1]. Korpusni loyhalash jarayonida avvalo uning dizayniga alohida e’tibor qaratish talab etiladi. Korpus foydalanuvchilarida ijobiy taassurot qoldirishi uchun shevaga xos elementlar va Nurota tumaniga aloqador bo‘lgan vizual oynalardan foydalanish zarur (1-rasm).

1-rasm. Nurota tumani dialektal korpusi interfeysi

Korpus dizaynida shu belgilarni inobatga olish juda muhim omillardan. Milliy til va uning shevalarini informatsion dunyoga olib chiqish, zamonaviy tadqiqotlar uchun muhim bir asos bo‘lib xizmat qiladi. Bu jarayon nafaqat tilshunoslik, balki madaniyatshunoslik va antropologiya sohalarida ham yangi imkoniyatlarni ochadi. Dastlabki shevalar bazasini yaratish jarayoni bir qator muhim talablar va maqsadlar asosida amalga oshiriladi. Ushbu bazaning asosiy maqsadi o‘zbek milliy shevalarining xilma-xilligini, ularning o‘ziga xos leksik va grammatik xususiyatlarini chuqur o‘rganishdir. Buning uchun quyidagi talablar ko‘zda tutiladi: (Qarang: 1-jadval)

1-jadval. Dialektal korpusda so‘zlar tasnifi

Shevaga oid so‘z	**Shipirgi**
Adabiy til	Supurgi
Turkum	Ot
So‘zning izohi	Supurish-sidirish, tozalash uchun burgan, shuvoq, makkasupurgi va shu kabilardan dastalab bog‘lab yasaladigan ro‘zg‘or buyumi
Sinonimi	Sinonimi mavjud emas

Turli xil dialektal so‘zlardan foydalanish, respondentlarning turli ijtimoiy va madaniy kontekstlarini aks ettirishga yordam beradi. Bu esa tadqiqotchilarga aniq va ishonchli xulosalar chiqarishga imkon yaratadi. Bir so‘zni har xil yo‘l bilan so‘rovnomada berish manbadan keng foydalanish imkoniyatini beradi. Ushbu metodologik yondashuvlar orqali so‘rovnoma natijalari yanada aniqroq bo‘ladi. Turli xil dialektal so‘zlardan foydalanish, respondentlarning turli ijtimoiy va madaniy kontekstlarini aks ettirishga yordam beradi. Bu esa tadqiqotchilarga aniq va ishonchli xulosalar chiqarishga imkon yaratadi. Boshqird dialekti xonqizi (kamka)ning 148 ta dialektal variantlarini berib o‘tadi, shulardan 17 ta varianti faqat sharq dialektlari tarkibida uchraydi[2]. Nurota shevalar elektron lug‘ati internetda onlayn ish yurituvchi bazadir. Bazadagi ma’lumotlar doirasini kengaytirish xususiyatiga ega bo‘lgan har tomonlama qulay dialektal bazalardan biridir. Ushbu elektron baza keng imkoniyatlar taqdim etadi. Baza foydalanuvchilar auditoriyasi uchun quyidagi imkoniyatlarni taqdim etadi.

1. Bosh sahifa. 2. So‘zlar jadvali. 3. Loyihasi haqida. 4. Transkriptsiya. 5. Aloqa.

Bosh sahifa shevalar bazasining markaziy qismi bo‘lib, u foydalanuvchilarga bazadagi ma’lumotlarga oson kirish imkonini beradi. Ushbu sahifa yuqoridagi beshta asosiy bo‘limdan iborat bo‘lib, har bir bo‘lim boshqa sahifalarga yo‘naltiruvchi aloqa vositasi sifatida xizmat qiladi. Bazadagi so‘zlarning turli

kesimlarda tahlil qilinishi foydalanuvchilarga shevaning leksik boyligini to‘liq anglash imkonini beradi. So‘z turkumlari asosan ikkita kategoriya bo‘yicha bo‘linadi: Umumiy tashqi so‘z turkumlari: Mustaqil so‘z turkumi, yordamchi so‘z turkumi alohida so‘z turkumi. Xususiy ichki so‘z turkumlari: ot, son, fe’l, ravish, sifat, olmosh, ko‘makchi, bog‘lovchi, yuklama, undov, modal va taqlid so‘zlar. Bu statistika foydalanuvchilarga shevalarning strukturasini va ularning foydalanish doirasini aniqlashda yordam beradi. Bazadagi shevaga xos leksemalar foydalanuvchilar uchun alohida ahamiyatga ega. Ushbu so‘zlar bazaning keng auditoriyaga mo‘ljallanganligini hisobga olib, foydalanuvchilar uchun qulaylik tug‘dirish maqsadida transkripsiya shaklida taqdim etilgan. Bu esa shevalarni o‘rganish jarayonini osonlashtiradi. Shevalar bazasining bosh sahifasi foydalanuvchilar uchun shevalarni o‘rganish jarayonini osonlashtiradi, ularning leksik boyligi, strukturası va tarixiy konteksti haqida aniq va to‘liq ma’lumotlar taqdim etadi. Bu o‘z navbatida, shevalarning saqlanishi va rivojlanishiga xizmat qiladi.

Shevalar Bazasi: Shevalar bazasida so‘zlar jadvali foydalanuvchilarga turli xil imkoniyatlarni taqdim etadi. Ushbu sahifa barcha qatorlar soni, izlash, turkum bo‘yicha va shevaga xos so‘zlarning etimologik saralash imkoniyatlarini o‘z ichiga oladi[3]. Jadval quyidagi ma’lumotlarni o‘z ichiga oladi: adabiy til, transkripsiya (mavjud jahon standartlariga muvofiq), turkum (so‘zning turkumini ko‘rsatish, saralash imkoniyati bilan), izoh (so‘zning ma’nosi yoki qo‘llanilishi haqida tushuntirish), sinonimi (ushbu so‘zga o‘xshash ma’noga ega so‘zlar)[4]. Shevalar bazasini yangi dialektal so‘zlar bilan to‘ldirish jarayonlari o‘z qoidalariga ega bo‘lib, quyidagi bosqichlarni o‘z ichiga oladi:

Yangi dialektal so‘zlarni aniqlash va ularni bazaga kiritish, har bir so‘zni tekshirib, uning to‘g‘riligini va mosligini aniqlash, so‘zlarni turkumlar bo‘yicha saralash va tegishli joylarga joylashtirish, so‘zlarning transkripsiya qoidalariga mos kelishini ta’minlash. Ushbu jarayonlar shevalar bazasining boyligini oshirishga va foydalanuvchilarga yanada qulay xizmat ko‘rsatishga yordam beradi.

XULOSA

Shevalar bazasi foydalanuvchilarga turli xil dialektal so‘zlar bilan ishlash imkonini beruvchi muhim resursdir. So‘zlar jadvali, adabiy til, transkripsiya, turkum, izoh va sinonimlar kabi ma’lumotlarni o‘z ichiga olgan holda, foydalanuvchilarga so‘zlarni oson topish va ularning ma’nolarini tushunishga yordam beradi. Yangi dialektal so‘zlarni bazaga kiritish jarayoni aniq qoidalar va bosqichlarga asoslangan bo‘lib, bu jarayonlar so‘zlarning to‘g‘riligini va mosligini ta’minlashga, shuningdek, bazaning boyligini oshirishga xizmat qiladi. Natijada, shevalar bazasi foydalanuvchilarga yanada qulay va samarali xizmat ko‘rsatishni maqsad qiladi. Bu resurs tilshunoslik va dialektologiya sohalarida muhim

ahamiyatga ega bo‘lib, mahalliy tillarning saqlanishi va rivojlanishiga hissa qo‘shadi.

Foydalanilgan adabiyotlar:

1. https://uz.wikipedia.org/wiki/Interfeys
2. Sirazitdinov Z.A., Buskunbaeva L.A., Karimova R.N. – P. 212–219, DOI 10.22162/2075- 7794-2016-23-1-212-219
3. Холова М. ўзбек миллий шевалари корпусини тузишнинг лингвистик асослари (Бойсун тумани “ж”ловчи шевалари мисолида):Фил. фан. бўйича фалсафа доктори (PhD)...дисс. Автореф. – Термиз, 2022.
4. Иванцова. Е.В. Томский диалектный корпус: обоснование концепции и перспективы развития. Вопросы лексикографии. 2017. № 11. – С. 62.
5. Tregubova E. N. Multi-level thematic markup as a tool for ethnolinguistic representation of dialect discourse in an e-text corpusю Vestnik. Tomsk State. University. Phil. . Vol.№ 1 (33). 2015. –P. 66-77.

ZAHIRIDDIN MUHAMMAD BOBURNING KORPUSINING INTERFEYSI

Nilufar Muradova Baxodir qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
muradovanilufar06@gmail.com
ToshDO‘TAU

Annotatsiya: Ushbu maqolada Zahiridin Muhammad Boburning asarlariga asoslangan korpusning interfeysini yaratish va nimalarga diqqat qaratish kerakligi tahlil qilingan. Maqolada Z.M.Bobur mualliflik korpusining imkoniyatlari va foydalanuvchilar uchun qulaylik masalalari ko‘rib chiqildi. Tadqiqotda korpusning interfeysini ishlab chiqishda qo‘llanilgan texnologiyalar, foydalanuvchi tajribasini yaxshilash uchun qo‘llanilgan dizayn prinsiplari va interfeysning baholanishi ham muhokama etildi. Zahiridin Muhammad Bobur korpusining interfeysining o‘ziga xos xususiyatlari va uning tadqiqotchilar uchun ahamiyati ta’kidlandi. Tadqiqot natijasida Bobur ijodiy merosini o‘rganishda zamonaviy texnologiyalardan samarali foydalanishning amaliy ahamiyati yoritilgan.

Annotation: This article is devoted to the creation of the interface of the corpus based on the works of Zahiridin Muhammad Babur and the analysis of its functionality. The article examines the main goals of Babur’s author corpus, its main functional capabilities and issues of providing convenience for users. The research also discusses the technologies used in the development of the corpus interface, the design principles used to improve the user experience, and the evaluation of the interface. The specific features of the interface of Zahiridin Muhammad Babur’s corpus and its importance for researchers are emphasized. The results of the research show the practical importance of effective use of modern technologies in the study of Babur’s creative heritage.

Kalit so‘zlar: korpus, korpus lingvistikasi, mualliflik korpusi, interfeys, muallif tili, lug‘at, ma’lumotlar bazasi, qidiruv tizimi, foydalanuvchi.

KIRISH

Korpusga qo‘yiladigan talablardan biri uning interfeysga egaligidir. Korpusning bosh oynasi – yuzi uning interfeysi. Korpus interfeysi – bu korpusning maqsadi, muallifning estetik didi va dunyoqarashiga asoslanib tuziladigan ma’lumotlar jamlanmasidir.

Interfeys qanchalik aniq, sodda va tushunarli bo‘lsa foydalanuvchi uchun shuncha samarali xizmat qiladi. Interfeysning asosiy maqsadi foydalanuvchi va tizim o‘rtasida samarali va qulay aloqa o‘rnatishdir. Bu aloqa orqali foydalanuvchi saytdagi ma’lumotlardan foydalanishi, saytdan ma’lumot olishi mumkin. Interfeys orqali korpus haqida dastlabki xulosa va sayt imkoniyatlari haqida ma’lumot olinadi. Interfeys mukammal ishlab chiqilsa, foydalanuvchi o‘z vazifalarini tezda va

samarali bajarishi, tizim bilan muammosiz ishlashi mumkin. Interfeysda ma’lumotlar tartibli va tizimli shaklda berilsa, bu foydalanuvchiga yanada qulaylik yaratadi.

Nafaqat til korpuslarida balki mualliflik korpuslarining ham interfeysi mukammal ishlab chiqilganligi korpusning imkoniyatlarini oshiradi. Mualliflik korpusi ma’lum bir yozuvchi yoki shoir asarlari jamlangan elektron kutubxonadir. Bu kutubxonani qanday bo‘lishi, tuzilishi, dizayni, muallif asarlari joylashuvi va tartibi foydalanuvchi uchun qulay va sodda bo‘lishi kerak.

ASOSIY QISM

Ma’lum bir yozuvchi yoki shoirning mualliflik korpusi interfeysi va korpusdagi bo‘limlar miqdori va tartibini belgilashda avvalo, muallif ijodiy merosini chuqur o‘rganish talab etiladi. Har bir yozuvchi uchun individual yondashuv zarur. Korpusdagi bo‘limlar sonini ham ijodkorning ijodiy merosining salmog‘i va ijod qilgan janrlar turi bilan bog‘liq. Sababi mualliflik korpusi bir shaxsning umumiy ijodining ko‘zgusidir. Mualliflik korpusini ishlab chiqishda, har bir yozuvchi uchun alohida yondashuv zarur. Avvalo, yozuvchi ijodining o‘ziga xosligini o‘rganish, ijodiy merosi bilan batafsil tanishish darkor. Negaki, mualliflik korpusining tuzilishini, tarkibini ishlab chiqishda yozuvchi ijodi xususiyatidan kelib chiqqan holda yondashiladi.

Zahiriddin Muhammad Bobur mualliflik korpusi interfeysi va bo‘limlar sonini belgilashda ijodiy merosidan kelib chiqqan holda ishlab chiqilgan. Bobur hayotining ko‘p qismini jang-u jadallarda, siyosiy maydonda o‘tkazganiga qaramasdan o‘zidan boy adabiy meros qoldirgan. Boburning umumiy 7 ta asari mavjud. Bular quyidagilar.

1-jadval. Zahiriddin Muhammad Boburning ijodiy merosi

1. “Boburnoma” shoh memuar asar
2. Aruz ilmidan saboq beruvchi “Muxtasar”
3. Soliq ishlariga oid “Mubayyin”
4. “Bobur devoni”
5. Harbiy tizimga oid “Harb ishi”
6. Musiqa borasidagi “Musiqa ilmi” asar
7. “Volidiya” tarjima asari

Yuqoridagi asarlardan “Boburnoma”, “Muxtasar”, “Mubayyin”, “Bobur devoni” hamda “Harb ishi” asarlari yetib kelgan. Ammo “Musiqa ilmi” va “Volidiya” asarlari yetib kelmagan. Ilmiy tadqiqotimizning obyekti sifatida Boburning “Boburnoma” va “Bobur devoni” asarlari tanlab olingan. Korpusga qolgan asarlarini ham kiritishni kelgusi ilmiy ishlarimizda davom ettirish maqsad qilingan.

Zahiriddin Muhammad Bobur mualliflik korpusi interfeysi 6 bo‘limdan iborat (1-rasm).

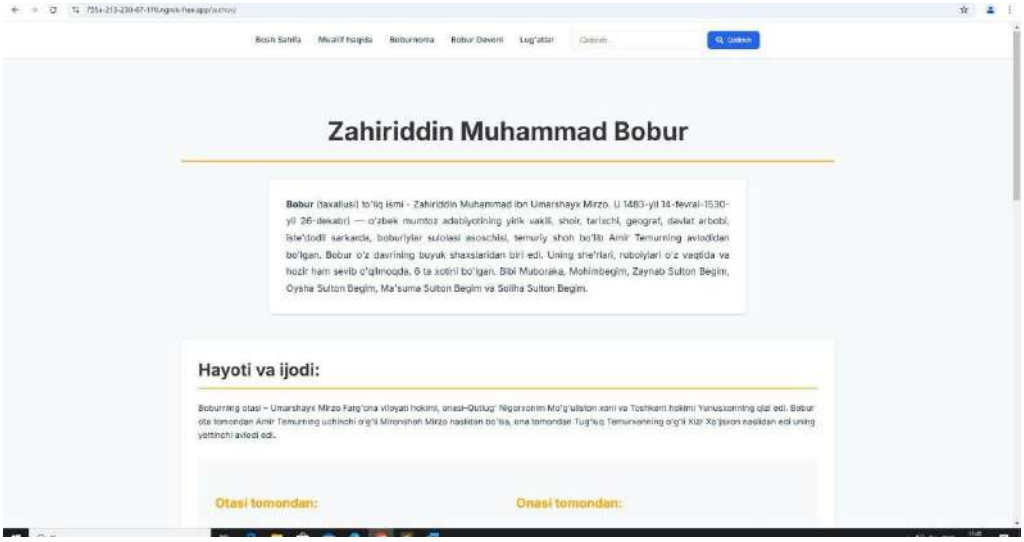
1. Bosh sahifa
2. Muallif haqida
3. Boburnoma
4. Bobur devoni
5. Lug‘at
6. Qidiruv

Zahiriddin Muhammad Bobur muallif korpusi tarkibini tuzishda, avvalo, yozuvchining ijodiy merosini o‘rganish, yozuvchi ijodi o‘ziga xos xususiyatlarini aniqlash talab etiladi. Bosh sahifada Zahiriddin Muhammad Boburning surati aks etgan va prizidentimizning nutqidan quyidagi so‘zlari o‘rin olgan. “Jahon tarixida o‘chmas iz qoldirgan betakror siymolardan biri Zahiriddin Muhammad Bobur buyuk shoir, qomusiy olim, davlat arbobi va mohir sarkarda sifatida butun dunyoda ma’lum va mashhurdir. Uning beqiyos ilmiy-ijodiy merosi nafaqat milliy madaniyatimiz va xalqimiz adabiy-estetik tafakkurining shakllanishida, balki jahon adabiyoti, ilm-fani va davlatchiligi tarixida alohida o‘ringa ega” [https://lex.uz/uz/docs/-6371358].

Bosh sahifa dastlabki oyna bo‘lganligi bois, foydalanuvchida bu korpus haqidagi ilk tasavvurni shakllantiradi. Shu sababli bosh sahifa tartibli va ma’lumotlar tushunarli tarzda joylashtirilgan bo‘lishi zarur.

1-rasm. Zahiriddin Muhammad Bobur korpusi interfeysi

Ikkinchi bo‘lim muallif haqida deb nomlangan. Bu bo‘limda Zahiriddin Muhammad Boburning bolalik va yoshlik davri, ism va taxallusi, nasl-nasabi, hayoti va ijodiga tegishli eng muhim ma’lumotlar joylangan. Korpus foydalanuvchisi nafaqat Bobur asarlari bilan balki, hayotiga oid ma’lumotlar bilan ham tanishish imkoniyatiga ega bo‘ladi. Bu bo‘limda Bobur shaxsiyatiga va ijodiy faoliyatiga oid manbalar joy olgan (2-rasm).



2-rasm. Zahiriddin Muhammad Bobur korpusining 2-bo‘limi

Korpusda uchinchi bo‘lim “Boburnoma” deb nomlangan. “Boburnoma” asari muallifning shoh memuar asari hisoblanadi. Asar uch qismdan iborat bo‘lib, unda O‘rta Osiyo, Afg‘oniston va Hindistondagi hayoti tasvirlangan. Shu bilan birga

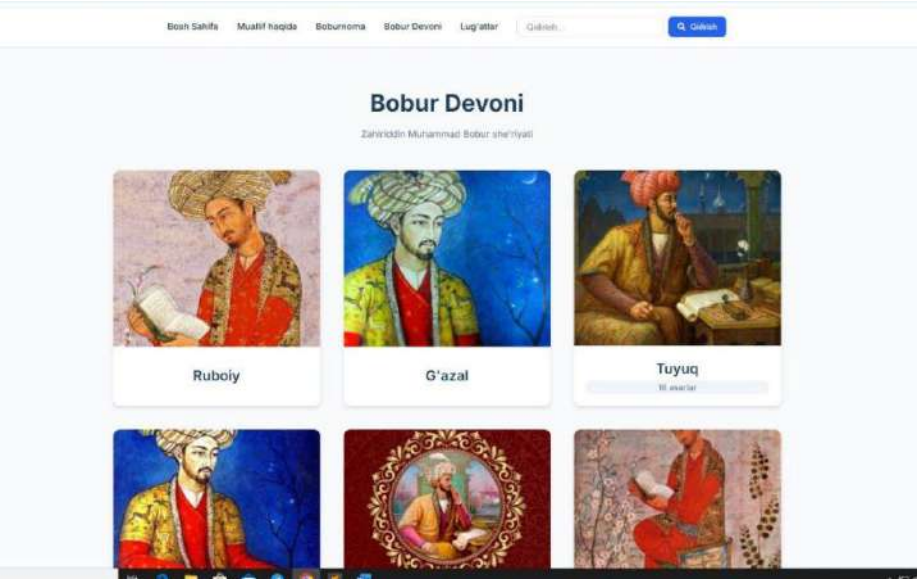
xalqlarning urf-odatlari, geografiyasi, etnografiyasi, turmush tarzi, madaniyati, o‘simlik va hayvonot dunyosiga oid ma’lumotlar o‘rin olgan. “Boburnoma” asari shu kungacha bir necha marta nashr etilgan. Korpusga 2002-yilda “Sharq” nashriyot-matbaasi tominidan nashr etilgan varianti joylandi. Asar krill yozuvida nashr etilgan. Korpus uchun krilldan lotin yozuviga o‘girilgan shaklda kiritildi. Foydalanuvchi asar matni bilan tanishishi va matn tarkibidagi istalgan so‘zni qidirish imkoniyati mavjud.

Korpusda to‘rtinchi bo‘lim “Bobur devoni” deb nomlangan. Bobur ijodining asosiy qismi lirik asarlar tashkil etadi. Devonda aynan lirik janrdagi asarlari jamlangan. Bobur devonidagi janrlar soni va miqdori quyidagi jadvalda berilgan (1-jadvalga qarang). Korpusga jadvalda keltirigan barcha janrga oid ijod namunalari o‘rin olgan.

2-jadval. Bobur devonidagi janrlar soni va miqdori

№	JANRI	SONI
1	G‘azal	220
2	Tuyuq	16
3	Ruboiy	242
4	Qit’a	25
5	To‘rtlik	9
6	Muammo	54
7	Masnu she’r	1
8	Fard	99
9	Ash’ori forsiy	2
10	Ruboiyot	12
11	Muqattoat	7
12	Fardho	17
13	Masnaviy	364
14	Noma	6

Devonning 2022-yilda “Sharq” nashriyoti tomonidan nashr etilgan varianti asos qilib olingan. Foydalanuvchi bu bo‘limga kirganda quyidagi oyna hosil bo‘ladi (2-rasm). Devonda jami o‘n to‘rtga janrga oid lirik she’rlardan iboratligi uchun korpusda ham har bir janr namunalari alohida joylangan. Bu bo‘limda oynada 14 ta janrga oid asarlari aks etagan. Korpusga devon yaxlit tarzda emas janrlariga xos tartibda o‘rin olgan. Devondagi lirik janrlar sarlavhalanmagi uchun, har bir janrga oid asarlari birinchi satri bilan berilgan. Masalan g‘azal qismiga kirilsa devondagi barcha g‘azallarning birinchi satri ketma-ketlikda chiqadi. Foydalanuvchi istalgan g‘azalni birinchi satri belgilansa, butun g‘azal matni bilan tanishishi mumkin.



3-rasm. Bobur devoni bo‘limi

Korpusdagi eng muhim xususiyatlardan biri qidiruv tizimi hisoblanadi. Shu sababli qidiruv tizimi qanchalik mukammal tuzilgan bo‘lsa, korpus ham shunchalik sifatli bo‘ladi. Korpusda qidiruv tizimining bir necha shakllari mavjud. Birinchisi so‘zshaklida, ikkinchisi janr shaklidagi qidiruv tizimi hisoblanadi. Bu yanada muallif asarlarida qidiruvni amalga oshirishda qulaylik yaratadi. Birinchi qidiruv tizimi ya’ni so‘z shaklida qidiruv amalga oshirilganda, bu so‘z “Bobunoma” va “Bobur devoni” asarlaridan olingan natijani taqdim etadi. Ikkinchi qidiruv shakli-janr bo‘yicha qidiruv amalga oshirilganda esa, “Bobur devoni”dagi janrlar tarkibidan qidiruv olb boriladi. Korpus menejerining mukammal ishlab chiqilganligi korpus qidiruv tizimi imkoniyatini yanada oshirishga xizmat qiladi.

XULOSA

Zahiridin Muhammad Boburning asarlariga asoslangan mualliflik korpusning interfeysini yaratish va uning samaradorligini oshirish bo‘yicha o‘tkazilgan tadqiqotlar shuni ko‘rsatdiki, mukammal ishlab chiqilgan interfeys Bobur ijodiy merosini o‘rganish va tadqiq etish jarayonini osonlashtiradi. Korpus interfeysining muhim xususiyatlari sifatida qulay qidiruv tizimi, matnlarni tezkor tahlil qilish imkoniyati, lug‘at va grammatika ma’lumotlariga oson kirish, turli xil so‘rovlarni amalga oshirish va statistik natija olish imkoniyatini beradi. Kelajakdagi tadqiqotlarda korpusga yangi funksionallarni qo‘shish, ma’lumotlar bazasini boyitish va muallifning boshqa asarlarini ham korpusga joylashtirish maqsad qilingan. Ushbu mualliflik korpusi Bobur merosini chuqurroq o‘rganish, tahlil qilish va uning boy tilini o‘rganish uchun yangi imkoniyatlar yaratadi.

‚Foydalanilgan adabiyotlar:

1. Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. Globe edits, 2020.
2. Hoshimova D. “Boburnoma” asarining tanqidiy matnini yaratish muammolari. Filol. Fan. Nomz… diss. – Toshkent, 2006.
3. Rahimov A., Kompyuter lingvistikasi asoslari. –Toshkent: Akademnashr, 2011.
4. Xolmonova Z. “Boburnoma”- til qomusi. – Toshkent: Akademnashr, 2021.
5. Hamroyeva Sh., Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati. 2020.
6. https://lex.uz/uz/docs/-6371358

MUALLIFLIK KORPUSI QIDIRUV TIZIMINI ISHLAB CHIQISH

Sanjarbek Baxodirov Rahmatali o‘g‘li,
Kompyuter lingvistikasi mutaxassisligi magistranti
sanjarbahodirov9901@gmail.com
ToshDO‘TAU

Nilufar Muradova Baxodir qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
muradovanilufar06@gmail.com
ToshDO‘TAU

Annotatsiya. Mualliflik korpusi qidiruv tizimi zamonaviy lingvistik tadqiqotlar uchun samarali vosita hisoblanadi. Ushbu tizim foydalanuvchilarga matnlarni konkordans qidiruv va kengaytirilgan qidiruv yordamida tahlil qilish imkonini beradi. Konkordans qidiruv so‘z shakllari va ularning kontekstlarini aniqlashda yordam bersa, kengaytirilgan qidiruv turli filtrlash parametrlari orqali aniq natijalar taqdim etadi. Mazkur maqolada ushbu qidiruv tizimining asosiy tamoyillari, ishlash mexanizmi va texnik jihatlari yoritiladi. Tizimning samaradorligi va foydalanuvchilarga qulayligi haqidagi tahlillar ham keltiriladi. Mualliflik korpusining ushbu qidiruv imkoniyatlari lingvistika, tarjima va matnshunoslik sohalari uchun muhim ahamiyat kasb etadi.

Abstract. The authorship corpus search system is an effective tool for modern linguistic research. This system allows users to analyze texts through concordance search and advanced search functionalities. The concordance search helps identify word forms and their contexts, while the advanced search provides precise results through various filtering parameters. This article highlights the fundamental principles, operational mechanisms, and technical aspects of this search system. It also includes an analysis of the system's efficiency and user-friendliness. The search capabilities of the authorship corpus are of great importance for linguistics, translation, and text studies.

Аннотация. Система поиска по авторскому корпусу является эффективным инструментом для современных лингвистических исследований. Данная система позволяет пользователям анализировать тексты с помощью поиска по конкордансам и расширенного поиска. Поиск по конкордансам помогает определить формы слов и их контексты, а расширенный поиск предоставляет точные результаты благодаря различным параметрам фильтрации. В данной статье рассматриваются основные принципы, механизм работы и технические аспекты данной поисковой системы. Также приводится анализ её эффективности и удобства использования. Возможности поиска в авторском корпусе имеют важное значение для лингвистики, перевода и текстоведения.

Kalit so‘zlar: *Mualliflik korpusi, konkordans qidiruv, kengaytirilgan qidiruv, lingvistik tahlil, matnshunoslik, qidiruv algoritmlari, korpus lingvistikasi.*

KIRISH

Hozirgi kunda lingvistik tadqiqotlar sohasida korpus lingvistikasi muhim ahamiyat kasb etmoqda. Matnlarni chuqur tahlil qilish, leksik va grammatik xususiyatlarini o‘rganish, tarjima jarayonlarini soddalashtirish va sun’iy intellekt tizimlarini rivojlantirish uchun korpuslardan foydalanish samarali usullardan biri hisoblanadi. Shu nuqtai nazardan, mualliflik korpusi ma’lumotlarni tizimli ravishda to‘plash, saqlash va tahlil qilish imkonini beruvchi qudratli vositadir.

Mualliflik korpusi yordamida matnlar tarkibiy qismlarini o‘rganish, so‘zlarning ishlatilish chastotasini aniqlash, semantik bog‘liqliklarni tadqiq qilish hamda tarjima amaliyotida muhim bo‘lgan parallel matnlarni tahlil qilish imkoniyati yaratiladi. Ushbu maqolada mualliflik korpusi uchun ishlab chiqilgan qidiruv tizimining ishlash tamoyillari, asosiy funksiyalari va texnik jihatlari ko‘rib chiqiladi.

Mazkur tadqiqotda korpus asosida ishlovchi qidiruv tizimlarining samaradorligi va foydalanuvchilarga qulayligi tahlil qilinadi. Shu bilan birga, lingvistik tadqiqotlarda ushbu tizimdan foydalanish imkoniyatlari, kelajakda uni yanada takomillashtirish yo‘nalishlari muhokama qilinadi. Ushbu maqola nafaqat lingvistlar, balki tarjimonlar, tadqiqotchilar va sun’iy intellekt tizimlari bilan shug‘ullanuvchilar uchun ham foydali bo‘lishi mumkin.

ASOSIY QISM

Konkordans qidiruvi. Matnlarni lingvistik jihatdan tahlil qilishda so‘zlarning kontekstini o‘rganish muhim vazifalardan biri hisoblanadi. Shu sababli, bizning qidiruv tizimimizning birinchi qismi Konkordans qidiruvi bo‘lib, u ma’lum bir so‘z yoki iborani uning kontekstida izlash imkonini beradi. Ushbu tizimni Python dasturlash tili va Django veb-frameworkidan foydalangan holda ko‘rib chiqildi.

Konkordans qidiruv tizimi uchta asosiy qism dan iborat:

1. **Ma’lumotlar bazasi modeli** – Konkordans jadvali so‘zlarni va ularning kontekstini saqlaydi.

2. **Qidiruv funksiyasi** – concordance_search funksiyasi foydalanuvchining kiritgan so‘zi bo‘yicha mos keluvchi natijalarni chiqaradi.

3. **Ma’lumotlarni indekslash** – build_concordance kommandasi yordamida matnlardagi so‘zlar avtomatik ravishda **Concordance** bazasiga yuklanadi.

Konkordans modeli (ma’lumotlar bazasi modeli). Konkordans qidiruvi asosiy ma’lumotlar bazasi modeli sifatida **Concordance** modelidan foydalanadi.

Ushbu model har bir so‘zning kontekstini, uning joylashuvi va tegishli asar bilan bog‘liqligini saqlaydi.

```
class Concordance(models.Model):
    work = models.ForeignKey(Work, on_delete=models.CASCADE)
    keyword = models.CharField(max_length=100)
    context_before = models.TextField()
    context_after = models.TextField()
    position = models.IntegerField()

    def __str__(self):
        return f"{self.keyword} in {self.work.title}"
```

Ushbu model quyidagi maydonlar(fields)dan iborat:(Qarang:1-jadval)

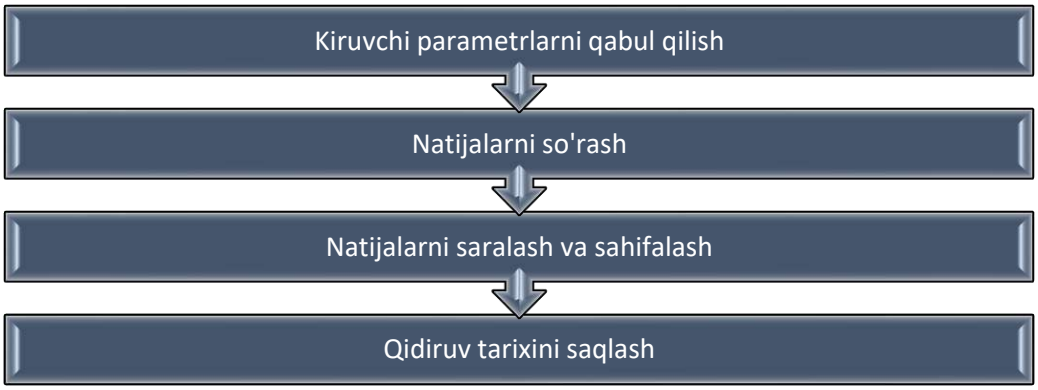
1-jadval. Konkordans qidiruv modeli tuzilishi

Maydon	Turi	Izoh
id	AutoField (PK)	Unikal identifikator
work_id	ForeignKey(Work)	Ushbu so‘z qaysi matnga tegishli ekanligi
keyword	CharField(100)	Qidirilayotgan so‘z
context_before	TextField()	So‘zdan oldingi kontekst
context_after	TextField()	So‘zdan keyingi kontekst
position	IntegerField()	So‘zning matndagi tartib raqami

2. *Concordance qidiruv funksiyasi* – matnda so‘zlarni kontekst (oldingi va keyingi so‘zlar) bilan birga izlash tizimi. Ushbu usul lingvistik tahlil, matnni semantik tekshirish, ma’lumot qidirish va tilshunoslik sohalarida keng qo‘llaniladi.

Django yordamida qurilgan ushbu qidiruv tizimi ma’lum matnlardan so‘zlarni chiqarib, ularning atrofidagi kontekst bilan birga saqlaydi va qidiruv imkoniyatlarini taqdim etadi.

concordance_search funksiyasi quyidagi bosqichlarda ishlaydi:



Kiruvchi Parametrlarni Qabul Qilish. Foydalanuvchi qidiruv so‘zini va qo‘shimcha filtrlarni (janr, davr, qidiruv turi) kiritadi:

```
query = request.GET.get('q', '')
genre = request.GET.get('genre')
period = request.GET.get('period')
search_type = request.GET.get('search_type', 'exact')
```

Qidiruv turlari:

1. exact – aynan mos keluvchi so‘zlarni qidirish (keyword__iexact=query)
2. starts_with – so‘z boshi bo‘yicha qidirish (keyword__istartswith=query)
3. contains – so‘z tarkibida berilgan qator bor-yo‘qligini tekshirish (keyword__icontains=query)

Natijalarni So‘rash. Qidiruv so‘roviga mos keluvchi natijalar Concordance modelidan olinadi:

```
if search_type == 'starts_with':
    results = Concordance.objects.filter(keyword__istartswith=query)
elif search_type == 'exact':
    results = Concordance.objects.filter(keyword__iexact=query)
else:
    results = Concordance.objects.filter(keyword__icontains=query)
```

Agar foydalanuvchi **janr yoki davr bo‘yicha filtr qo‘llasa,** natijalar quyidagicha filtrlanadi:

```
if genre:
    results = results.filter(work__genre__id=genre)
if period:
    results = results.filter(work__period__id=period)
```

Natijalarni Saralash va Sahifalash. Topilgan natijalar *asar nomi* va *so‘zning tartib raqami* bo‘yicha saralanadi:

```
results = results.order_by('work__title', 'position')
```

Natijalar **sahifalarga bo‘linadi** (20 tadan):

```
results_paginator = Paginator(results, 20)
concordances = results_paginator.page(results_page)
```

Qidiruv Tarixini Saqlash. Qidiruv tizimi foydalanuvchi kiritgan so‘rovlarni **tarixga yozib boradi:**

```
SearchHistory.objects.create(
    query=query,
    results_count=total_results,
    filters_used={'genre': genre, 'period': period, 'search_type': search_type}
)
```

Ma’lumotlarni Indekslash. Ma’lumotlarni indekslash – konkordans qidiruv tizimining muhim qismi bo‘lib, u matnlardagi so‘zlarni ajratib, ularning kontekstini aniqlash va ma’lumotlar bazasiga yuklash jarayonidir. Bu jarayon konkordans qidiruvning samarali ishlashi uchun zarur bo‘lib, foydalanuvchilar so‘zlarni qidirganda ularni o‘z konteksti bilan birga topish imkonini beradi.

Indekslash jarayoni quyidagi bosqichlardan iborat:

Birinchi bosqichda barcha matnlar Work modelidan olinadi. Ushbu model ma’lumotlar bazasida saqlanayotgan asar yoki hujjatlarni ifodalaydi. Indekslash funksiyasi har bir matn ustida ishlaydi va uni tahlil qiladi.

Ikkinchi bosqichda matn so‘zlarga bo‘linadi. Buning uchun maxsus matnni tahlil qilish usullari ishlatiladi. Masalan, regular ifodalar (regex) yordamida barcha so‘zlar ajratib olinadi va ular alohida ro‘yxatga joylashtiriladi. Har bir so‘z o‘zining matndagi joylashuv tartibiga ega bo‘ladi.

```
for i, work in enumerate(works, 1):
    self.stdout.write(f'Processing work {i}/{total_works}: {work.title}')

    # 3. Matnni kichik harflarga o‘tkazib, so‘zlarni ajratish
    words = re.findall(r'\b\w+\b', work.content.lower())
```

Uchinchi bosqichda har bir so‘zning konteksti aniqlanadi. Bu bosqichda har bir so‘zning oldidan va keyin keluvchi besh-oltita so‘z ajratib olinadi. Agar so‘z matnning boshida yoki oxirida joylashgan bo‘lsa, u holda mavjud bo‘lgan qismi olinadi. Kontekstni aniqlash natijasida qidiruv tizimi foydalanuvchi izlayotgan so‘z atrofidagi ma’lumotni ham taqdim etadi.

```
for position, word in enumerate(words):
    if len(word) < 2:  # Bitta harfdan iborat so‘zlarni o‘tkazib yuborish
    continue

# 4. So‘zning oldidan va keyin keluvchi so‘zlarni aniqlash
    start_pos = max(0, position - 5)
    end_pos = min(len(words), position + 6)

    context_before = ' '.join(words[start_pos:position])
    context_after = ' '.join(words[position+1:end_pos])
```

To‘rtinchi bosqichda so‘z va uning konteksti Concordance jadvaliga yoziladi. Ushbu jarayonda har bir so‘zga tegishli quyidagi ma’lumotlar bazaga qo‘shiladi:

- So‘zning o‘zi

- Matndagi joylashuv tartibi
- Oldingi so‘zlar (kontekst oldin)
- Keyingi so‘zlar (kontekst keyin)
- Ushbu so‘z tegishli bo‘lgan asar identifikatori

Barcha matnlar tahlil qilinib, kerakli ma‘lumotlar bazaga yuklangandan so‘ng, konkordans qidiruv tizimi ishlashga tayyor bo‘ladi. Endilikda foydalanuvchi so‘zlarni izlaganda, ularning matndagi aniq joylashuvi va konteksti bilan birga ko‘rish imkoniyatiga ega bo‘ladi.

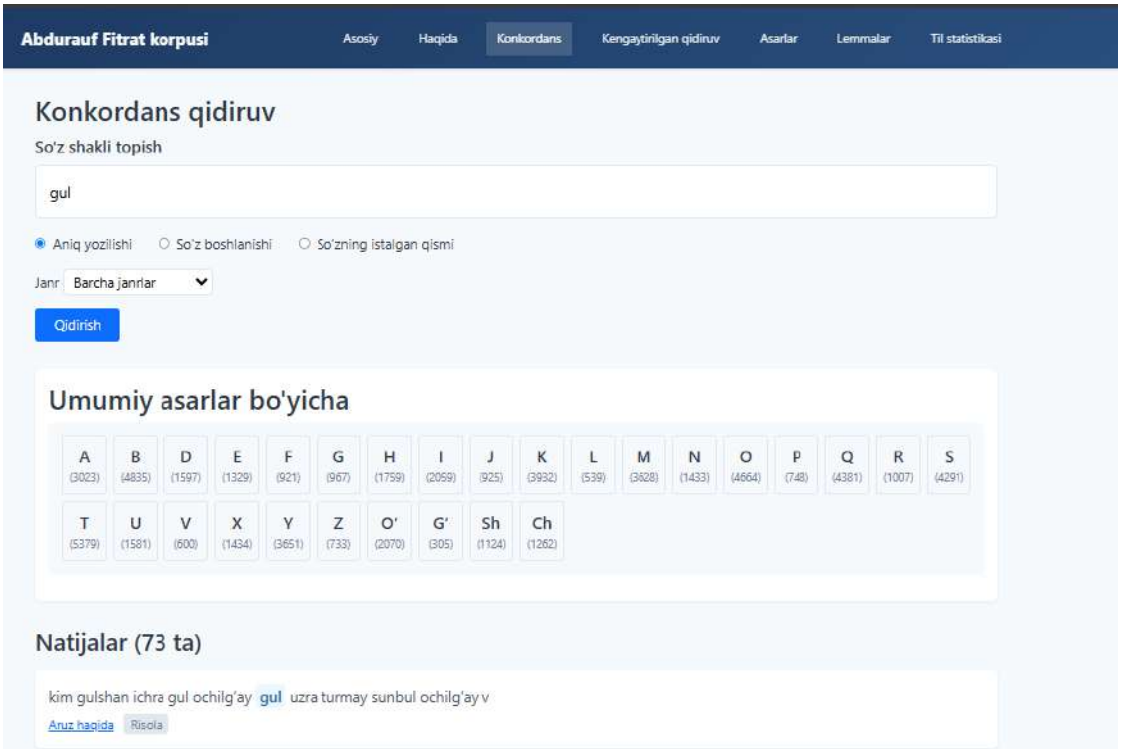
Indekslash jarayoni muntazam ravishda yoki yangi matnlar qo‘shilganda qayta ishga tushirilishi mumkin. Buning uchun maxsus Django management command yoziladi, bu buyruq chaqirilganda barcha mavjud matnlar qayta tahlil qilinadi va konkordans jadvali yangilanadi.

Umuman olganda, indekslash jarayoni konkordans qidiruv tizimining asosiy tayanch qismi bo‘lib, u so‘zlarni samarali qidirish va kontekstda ko‘rsatish imkonini beradi.

NATIJA. Konkordans qidiruv tizimi matn tahlili va lingvistik tadqiqotlar uchun muhim vositalardan biri bo‘lib, matndagi so‘zlarni ularning konteksti bilan birga topish imkonini beradi. Ushbu tadqiqot natijasida concordance qidiruvi amalga oshirishning samarali usullari ishlab chiqildi va tizimning asosiy ishlash tamoyillari aniqlandi.

Tadqiqot davomida matnlarni to‘liq tahlil qilish va qidiruv jarayonini tezlashtirish uchun maxsus indekslash mexanizmi yaratildi. Ushbu indekslash jarayonida matndagi har bir so‘z ajratib olinib, ularning joylashuvi va konteksti qayd etildi. Bu esa qidiruv natijalarini yanada aniqroq va samaraliroq qilish imkonini berdi. Foydalanuvchilar kiritgan so‘zni nafaqat mustaqil holda, balki uning oldidan va keyin keluvchi so‘zlar bilan birgalikda topish imkoniyatiga ega bo‘ldilar.(Qarang:1-rasm)

Qidiruv natijalari matndagi so‘zlarning qanday ishlatilishini tushunish, uslubiy tahlillar o‘tkazish va lingvistik tadqiqotlar olib borish uchun muhim ahamiyat kasb etdi. Tizim real vaqt rejimida tezkor natijalar taqdim etib, katta hajmdagi matnlarni samarali qayta ishlashga yordam berdi. Shuningdek, so‘z shakllari, sinonimlar va matn strukturasi bilan ishlash imkoniyatlari ham qo‘shildi, bu esa qidiruvning aniqroq bo‘lishiga zamin yaratdi.



1-rasm. Konkardans qidiruv natijasi

XULOSA

Ushbu tadqiqot natijasida concordance qidiruv tizimi matn tahlili uchun muhim vosita ekanligi tasdiqlandi. Tizim matndagi har bir so‘zning joylashuvi va kontekstini aniqlashga imkon berib, qidiruv natijalarini aniq va tushunarli shaklda taqdim etadi.

Indekslash jarayoni orqali matn tuzilmasi chuqur tahlil qilinib, so‘zlarning qaysi kontekstda ishlatilganini aniqlash imkoni yaratildi. Ushbu yondashuv lingvistik tadqiqotlar, matnni chuqur o‘rganish va turli tahliliy ishlarda samarali natijalar berishi mumkin.

Natijalar shuni ko‘rsatadiki, konkardans qidiruvi lingvistik izlanishlar, tarjima ishlari, matn muallifligini aniqlash, adabiy tahlillar va axborot qidiruv tizimlarida keng qo‘llanilishi mumkin bo‘lgan kuchli vosita ekanligi isbotlandi. Kelgusida tizimni yanada rivojlantirish uchun tabiiy tilni qayta ishlash (NLP) usullaridan foydalangan holda lemmatizatsiya, ma’noli tahlil va kontekstga asoslangan chuqur qidiruv imkoniyatlarini qo‘shish rejalashtirilmoqda.

Foydalanilgan adabiyotlar:

1. Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. – Globe edit, 2020.

2. Hamroyeva Sh. Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati: Terminologik lug‘at. – Globe edit, 2020.
3. Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005.
4. https://uzlc.tsuull.uz/index.php/uzlangcult/article/download/149/151/473
5. http://compling.navoiy-uni.uz/index.php/conferences/article/view/1038

DIYAKRON DERLEMIN FILOLOJIK ARAŞTIRMALARDAKI ÖNEMI

Yangibayeva Surayyo Gulimboy qizi,
Kompyuter lingvistikasi mutaxassisligi I bosqich magistranti
surayyoyangibayeva4@gmail.com
ToshDO‘TAU

Öz. Bu yazıda diyakron derlem kavramı ve onun filolojik araştırmalardaki önemi ele alınmaktadır. Tarihi dilbilim, dil değişimleri ve metindilbilim alanlarında diyakron derlemler aracılığıyla incelenen temel unsurlar değerlendirilmiştir. Ayrıca, diyakron derlem kullanarak dil sistemindeki değişimleri izleme, sözcüksel ve dilbilgisel yapıları analiz etmenin bilimsel önemi ayrıntılı bir şekilde incelenmiştir. Yazıda bu konu üzerine önde gelen akademisyenlerin bilimsel çalışmaları ve mevcut kaynaklardan yararlanma olanakları ele alınmaktadır.

Annotation. This article explores the concept of a diachronic corpus and its significance in philological research. The key aspects studied through diachronic corpora in historical linguistics, linguistic changes, and textology are examined. Additionally, the scientific importance of tracking language system changes, analyzing lexical and grammatical structures using a diachronic corpus, is discussed in detail. The article reviews the works of leading scholars on this topic and examines the possibilities of utilizing existing resources.

Аннотация. В данной статье рассматривается понятие диахронного корпуса и его значение в филологических исследованиях. Анализируются основные аспекты, изучаемые с использованием диахронных корпусов в историческом языкознании, лингвистических изменениях и текстологии. Кроме того, подробно обсуждается научная значимость наблюдения за изменениями в языковой системе, анализа лексических и грамматических структур с помощью диахронного корпуса. В статье рассматриваются научные труды ведущих исследователей по данной теме, а также возможности использования существующих ресурсов.

Anahtar kelimeler: *diyakron derlem, tarihi dilbilim, metindilbilim, dil değişimi, sözcüksel analiz, filolojik araştırmalar.*

Giriş

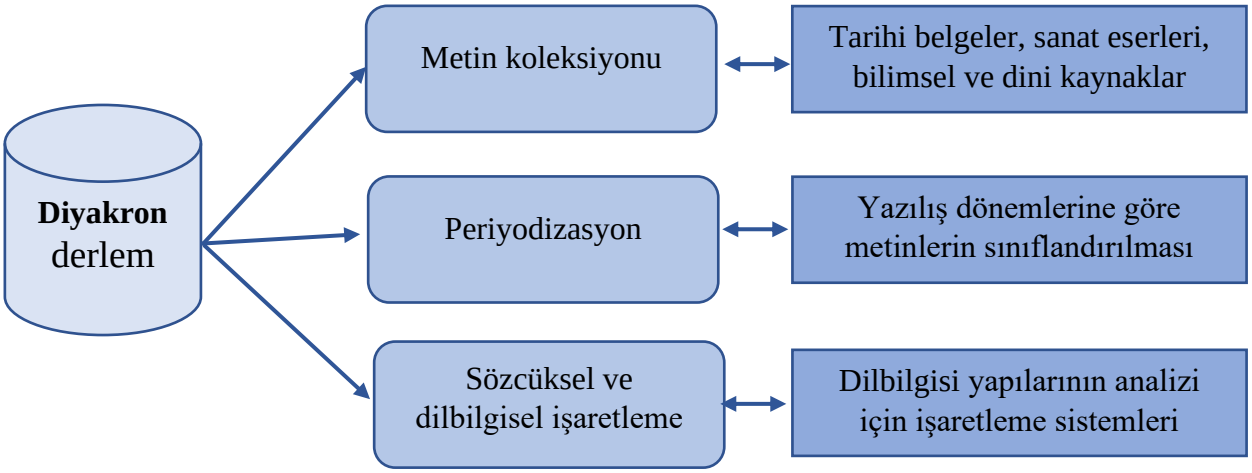
Dilbilim, yıllar içinde gelişiyor, genişliyor ve yenileniyor. Modern dilbilim ise bugün makine çevirisi, NLP, yapay zeka, korpus dilbilimi gibi birçok yeni alanla daha da genişledi. Sonuç olarak, dil derlemleri kavramı ortaya çıkmıştır. üretilen, kelimelerin, ifadelerin ve genel olarak dilin nasıl kullanıldığını analiz etmek için kullanılan çok büyük bir metin koleksiyonlarıdır. Bu derlemler, dilbilimciler, sözlükbilimciler, sosyal bilimciler, beşeri bilimler uzmanları, doğal dil işleme

uzmanları ve daha birçok alanda kullanılmaktadır [1]. Bunun sonucunda, farklı amaçları kapsayan yeni dil derlemleri oluşturulmaya başlanmıştır. Dil, zamanla sürekli değişen dinamik bir sistemdir [2]. Bu değişimleri sistematik bir şekilde incelemek için diyakron derlemler önemli bir araç olarak hizmet etmektedir. Diyakron derlem, farklı tarihî dönemlere ait metin koleksiyonlarından oluşur ve dilin değişim dinamiklerini gözlemleme imkânı sağlar [3]. Filolojik araştırmalar çerçevesinde bu tür derlemler, edebî dilin gelişimi, dilbilgisel yapıların evrimi ve sözcük türetme süreçlerinin derinlemesine incelenmesine olanak tanır. Bu yazıda diyakron derlem kavramı, yapısı ve filolojik araştırmalardaki önemi ele alınmaktadır.

Diyakron derlem kavramı ve yapısı

Diyakron derlem, farklı tarihsel dönemlere ait metinleri içeren bir dilbilimsel kaynak olup, dil değişimlerini analiz etmek, dilin evrimini derinlemesine incelemek, edebî süreçleri değerlendirmek ve filolojik araştırmalar için temel bir kaynak olarak kullanılmaktadır.

Şekil 1. Diyakron derlemin bileşenleri



Diyakron **derlemler** *dilbilim, edebiyat çalışmaları, metin çalışmaları ve kültürel çalışmalar* gibi çeşitli alanlarda yaygın olarak kullanılmaktadır. Örneğin, İngiliz dil biliminde *Helsinki Corpus* [4], *ARCHER Corpus va Corpus of Historical American English (COHA)* Eski ve Modern İngilizcenin gelişimi, dilbilgisi ve sözcük dağarcığındaki değişimler gibi kaynaklar aracılığıyla derinlemesine analiz edilir. Rus dilbiliminde *"Национальный корпус русского языка"* Rus dilinin dilbilgisi değişiklikleri derlem aracılığıyla gözlenip incelenmektedir.

Filolojik araştırmalarda diakronik derlemin rolü ve önemi

Filolojik araştırmalar dilbilimin önemli alanlarından biri olup, analitik yöntemlerin gelişmesiyle doğru ve güvenilir sonuçlara ulaşma olanağı artmaktadır.

Diyakronik metinler bir dilin tarihsel gelişiminin derinlemesine incelenmesinde en önemli araçlardan biri olarak kabul edilir ve aşağıdaki amaçlara hizmet eder.

1. **Sözcüksel değişikliklerin incelenmesi**

Diyakronik metinler, kelimelerin zaman içinde nasıl değiştiğini incelememize olanak tanır [5]. Örneğin Özbek dilinde bazı eski sözcüklerin modern biçimlere geçişi gözlemlenebilir. Zamanla bazı kelimeler tamamen kullanımdan kalkmış olabilir.

2. **Dilbilgisi yapılarının geliştirilmesi**

Bir dilin dilbilgisi sistemi sürekli değişmektedir. Diyakronik metinler aracılığıyla Özbek dilinde fiil zamanlarının tarihsel gelişimi veya cümle kuruluşunun evrimi izlenebilmektedir [6]. Katkı maddelerinin kullanım şekillerinde de değişiklik tespit etmek mümkündür.

3. **Metinsel çalışmalar ve üslup bilimi**

Farklı dönemlere ait metinler incelenerek yazar üslubu ve edebi türlerdeki değişimler incelenmektedir. Örneğin Ali Şir Nevai'nin eserleri ile çağdaş Özbek edebiyatının eserleri, diakronik bir külliyat temelinde karşılaştırılabilir. Yahut Cedit edebiyatı döneminde yaratılan eserleri, bağımsızlık dönemi edebiyatını ve çağdaş Özbek edebiyatını, diakronik bir külliyat üzerinden birbirleriyle karşılaştırabiliriz.

4. ***Dilin tarihini incelemek***

Diyakronik metinler, bir dilin antik, orta ve modern evrelerini karşılaştırmamıza, gelişim aşamalarını belirlememize ve dilsel değişimlerin derinlemesine analizini yapmamıza olanak tanır. Belirli bir yıl içinde kelimelerin kullanım sıklığının belirlenmesiyle hangi kelimelerin kullanımdan düştüğü, hangi kelimelerin yeni kullanıma girdiği incelenebilir.

Diyakronik korpusun pratik önemi göz önüne alındığında, bu tür korpuslar

Bu teknolojinin kullanımı sadece akademik araştırmalarda değil, aynı zamanda pratik alanlarda da önem taşımaktadır. Aşağıdaki alanlarda etkin bir şekilde kullanılabilir:

Tablo 1. Diyakron derlemin çeşitli disiplinlerde uygulanması.

№	Endüstriler	Kullanım amacı
1	***Yapay zeka ve makine çevirisi***	Tarihsel metinlerin modern dile uyarlanması sürecinin otomasyonunda faydalıdır.
2	***Kelime bilgisi***	Tarihî sözlüklerin derlenmesinde önemli bir kaynaktır.
3	***Kültürel mirasın korunması***	Antik yazılı eserlerin sayısallaştırılması ve analizinde kullanılır.
4	***Öğretim ve öğrenme süreci***	Araştırmacıların dilin tarihini derinlemesine incelemelerine malzeme teşkil eder.

5	***Edebiyat ve Metin Çalışmaları***	Edebi eserlerdeki dönemsel değişikliklerin izlenmesine ve üslup araştırmalarının yapılmasına yardımcı olur.
6	***Dil planlama ve standardizasyon***	Dildeki değişiklikleri analiz ederek biçimsel yazım ve dil bilgisi kurallarının iyileştirilmesine olanak sağlar.

Diyakron derlemler yardımıyla gerçekleştirilen araştırmalar

Günümüzde derlem dilbilimi alanındaki araştırmalar hızla gelişmektedir. Diyakronik metinler, dilin tarihsel gelişimini, sözcüksel-anlamsal değişimleri ve dilbilgisi yapılarının evrimini incelemede özellikle önemlidir. Bu metinler aracılığıyla dil birimlerinin farklı dönemlerdeki ortaya çıkışı, gelişimi ve yok oluşu bilimsel bir temele oturtulmaktadır. Çeşitli diller üzerinde diakronik korpus tabanlı araştırmalar yürütülmekte, dilsel özellikleri ve tarihsel etkenler incelenmektedir. Örneğin İngilizce, Almanca ve Fransızca dillerindeki değişim eğilimleri bu yöntem kullanılarak derinlemesine analiz edilmiştir. Bu çalışmalar sonucunda dil değişiminin iç ve dış etkenleri belirlenerek dilin dönemsel gelişim özellikleri açıklığa kavuşturulmaktadır. Aynı zamanda, hesaplamalı dilbilim ve yapay zekâ araçları kullanılarak diakronik metinler derinlemesine analiz edilmekte, böylece sonuçların doğruluğu artırılmaktadır. Filoloji araştırmalarında diakronik metinlerden yararlanılarak yapılmış çok sayıda çalışma bulunmaktadır. Bu çalışmalar dilbilimin çeşitli alanlarına uygulanmıştır. Örneğin Dawn Archer, Jonathan Culpeper ve Paul Rayson (2015) İngilizcede kiplik kategorisinin gelişimini incelemiştir. Araştırmaları modal fiillerin (örn *“must”* Ve *“shall”*) tarihsel süreçte çeşitli anlam değişimlerine uğramıştır. Ayrıca, diachronik derlemlerden yararlanılarak morfolojik ve sentaktik değişimler de incelenmektedir. Douglas Biber ve Susan Conrad (2009) akademik ve bilimsel üslupta cümle yapılarının tarihsel gelişimini incelemiştir. Sonuçlar, uzun cümleler yerine kısa ve net yapıların daha sık kullanıldığını göstermektedir. Türk dilbiliminde diakronik metinlere dayalı çalışmalar arasında Aksan ve Aksu-Koç (2016) tarafından yapılan çalışma özel bir öneme sahiptir. Türkçedeki morfolojik değişimleri incelemişler ve fiil formlarındaki değişimleri ve bunların tarihsel dönemlere göre kullanımını analiz etmişlerdir [7]. Ayrıca Koç H. ve Uçar M. (2019) diachronik derlemleri kullanarak Eski ve Modern Türkçe'deki sözcüksel değişimleri incelemiştir [8]. Bu çalışmalar Türk dilinin art zamanlı gelişiminin aydınlatılması açısından büyük önem taşımaktadır.

Sonuç

Filolojik araştırmalarda dil gelişiminin dinamiklerinin incelenmesi açısından diakronik metinler önemlidir. Bu makale aracılığıyla, sözcüksel değişim, dilbilgisi yapılarının evrimi ve metin çalışmaları alanlarında diakronik metinlerin yaygın kullanımına bir kez daha tanıklık etmiş olduk. Bu nedenle her dil için diyakronik derlemlerin oluşturulması büyük önem taşır. Gelecekte diyakron derlemlerin daha

da gelişmesi ve kullanım olanaklarının genişlemesi, filolojik araştırmaların verimliliğini artırmaya hizmet edecektir.

Kaynakça:

1. Elov B., Xusainova Z., (2024) Til korpuslarini lingvistik teglash bosqichlari.

2. Labov W. (1994). Principles of Linguistic Change: Internal Factors. Blackwell.

3. McEnery T. and Hardie A. (2012). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.

4. Rissanen M. (1999). Helsinki Corpus of English Texts: Diachronic and Synchronic Approaches. Helsinki University Press.

5. Brinton L.J. and Arnovick L.K. (2006). The English Language: A Linguistic History. Oxford University Press.

6. Vinogradov V.V. (1982). Russkiy yazyk: Grammaticheskoe uchenie o slove. Moskva: Nauka.

7. Aksan Y. ve Aksu-Koç A. (2016). “Turkish Morphological Change in Diachronic Corpora”. Turkish Journal of Linguistics, 33(2), 102–120. Google Scholar

8. Koç H., Uçar M. (2019). “Lexical Changes in Historical and Modern Turkish: A Diachronic Corpus Analysis”. Journal of Turkish Studies, 45(3), 223–245. ResearchGate

9. Archer, D., Culpeper J. and Rayson, P. (2015). “Modality in the history of English”. Language and Computers, 72(1), 1–26. DOI: 10.1075/lab.72.1.01arc

10. Apresyan Y.D. (2006). Natsionalniy korpus russkogo yazyka: Novye podkhody k issledovaniyu. Moskva: Yazyki slavyanskoy kultury.

SIFAT SO‘Z TURKUMINI GRAMMATIK POS TEGLASH

Xudayberganov Nizomaddin Uktambay o‘g‘li,
O‘qituvchi
nizomaddin@navoiy-uni.iz
ToshDO‘TAU

Mansurova Shaxinabonu Najmiddin qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
tiladabiyot@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada sifat so‘z turkumining grammatik jihatdan qismlariga ajratilishi va tabiiy tilni qayta ishlash (NLP) doirasida so‘z turkumlarini belgilash (POS tagging) jarayoni tahlil qilinadi. Sifatlar til tuzilishida muhim o‘rin tutib, ularning morfologik va sintaktik xususiyatlarini to‘g‘ri aniqlash matnlarni avtomatlashtirilgan qayta ishlashda katta ahamiyatga ega. Maqolada sifatlarning darajalari, shakllari va qo‘llanilishiga asoslangan POS teglash usullari hamda zamonaviy sun‘iy intellekt algoritmlari yordamida amalga oshiriladigan yondashuvlar haqida ma‘lumot beriladi.

Annotation. This article analyzes the grammatical classification of the adjective part of speech and the process of part of speech tagging (POS tagging) within the scope of Natural Language Processing (NLP). Adjectives play an important role in language structure, and correctly identifying their morphological and syntactic properties is crucial for the automated processing of texts. The article provides information on POS tagging methods based on the degrees, forms, and usage of adjectives, as well as approaches implemented using modern artificial intelligence algorithms.

Аннотация. В данной статье анализируется грамматическая классификация прилагательного как части речи и процесс тегирования частей речи (POS tagging) в рамках обработки естественного языка (NLP). Прилагательные играют важную роль в языковой структуре, и правильное определение их морфологических и синтаксических свойств имеет большое значение для автоматизированной обработки текстов. В статье представлена информация о методах тегирования частей речи на основе степеней, форм и использования прилагательных, а также подходах, реализуемых с использованием современных алгоритмов искусственного интеллекта.

Kalit so‘zlar: *sifat, so‘z turkumlari, grammatik tahlil, POS tagging, tabiiy tilni qayta ishlash, morfologik tahlil, sintaktik tahlil, NLP, sun‘iy intellekt, mashinani o‘rganish.*

Kirish. Sifat so‘z turkumi, lingvistikada sifat sifatida tanilgan so‘zlar to‘plami bo‘lib, ob‘ektlar yoki hodisalarni ta‘riflash va ularga xususiyatlar berish uchun ishlatiladi. Tilshunoslikda sifatlar odatda o‘zi bilan bog‘liq bo‘lgan nomlarni ta‘riflash uchun ishlatiladi, masalan, "yaxshi kitob", "baland tog‘" va "quyoshli kun". Bu so‘zlar ma‘lum bir ob‘ektning xususiyatlarini aniqlab, uni boshqalardan ajratib turadi. Shuning uchun sifatlar tabiiy tilni qayta ishlash va matnni avtomatik tahlil qilishda muhim ahamiyatga ega[1].

Grammatik pos teglash (Part-of-Speech Tagging, POS tagging) – tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining muhim vazifalaridan biri bo‘lib, unda har bir so‘zning grammatik rolini belgilash orqali matnni avtomatik ravishda tahlil qilish amalga oshiriladi. Grammatik pos teglash tilni analiz qilishda tilning morfologik va sintaktik xususiyatlarini aniqlashga yordam beradi va bu jarayon ko‘plab tilshunoslik va kompyuter fanlarida keng qo‘llaniladi. Ushbu jarayon yordamida har bir so‘zning grammatik kategoriyasi, masalan, ot, fe‘l, sifat, predlog va boshqa so‘z turkumlari aniqlanadi.

Grammatik POS teglash jarayonida sifatlar matndagi o‘rniga va uning grammatik funktsiyasiga qarab teglar bilan belgilanishi kerak. Bu jarayon tillarning morfologik va sintaktik xususiyatlariga asoslanadi, va tahlil qilishda tilning turli xususiyatlari, masalan, sintaktik roli, kontekst va ma‘no o‘zgarishi inobatga olinadi. Sifat so‘zlarining grammatik teglari, tilning tuzilishiga qarab, turli xil bo‘lishi mumkin, masalan, "yaxshi" sifatining "JJ" yoki "ADJ" teglariga ega bo‘lishi mumkin.

Bugungi kunda sifatlarni POS teglash tizimlarida to‘g‘ri aniqlash uchun bir qator ilg‘or metodlar mavjud. Masalan, mashina o‘rganish, chuqur o‘rganish (deep learning) va tabiiy tilni qayta ishlashning boshqa zamonaviy usullari sifat so‘zlarni aniqlashda yordam berishi mumkin. Ayniqsa, chuqur o‘rganish algoritmlari va neyron tarmoqlarining qo‘llanilishi POS teglash jarayonini yanada samarali va aniqlik bilan bajarishga yordam beradi. Shu bilan birga, sifatlar o‘zgarishi mumkin bo‘lgan kontekstda ham to‘g‘ri tahlil qilish zarurati paydo bo‘ladi.

Shu sababli, bu maqola sifat so‘z turkumi va uning grammatik POS teglashini amalga oshirishdagi qiyinchiliklar va hozirgi kunda mavjud bo‘lgan usullarni o‘rganish uchun mo‘ljallangan. U shuningdek, tilni qayta ishlashda sifatlarni to‘g‘ri belgilashning ahamiyatini va bu jarayonni yaxshilashda qo‘llaniladigan zamonaviy texnologiyalarni tahlil qilishni maqsad qiladi[2].

Muhokama va natijalar

Sifatni o‘rganish tizimi materiali leksik va grammatik jihatdan izchillikni saqlagan holda rivojlantirib borish, shuningdek, murakkablashtirishni nazarda tutadi. 1-sinfda o‘quvchilar sifatning leksik ma‘nosini va uning qanday? va qanaqa?

so‘roqlariga javob berishini o‘rganadilar. 2-sinfda sifat so‘z turkumi sifatida to‘liq o‘zlashtiriladi, 3-sinfda esa ilgari o‘rganilganlarni takrorlab, grammatik materialga asoslangan –roq qo‘shimchasi bilan yasalgan sifatlarning yozilishi o‘rgatiladi (terminlar ishlatilmaydi). Ona tili va o‘qish darslarida o‘quvchilarning nutqi yangi sifatlar bilan boyitiladi va ilgari o‘rganilgan sifatlarning ma‘nosiga aniqlik kiritiladi.

Sifat so‘z turkumining grammatik teglanishi quyidagicha:

Yaxshidan bog‘ qoladi, yomondan dog‘.

Keltirilgan gapda, yaxshi va yomon so‘zlari sifat so‘z turkumiga mansub bo‘lib, yaxshi(xususiyat sifati, asliy sifat, oddiy daraja, sodda sifat), yomon(xususiyat sifati, asliy sifat, oddiy daraja, sodda sifat) ko‘rinishida grammatik teglanadi.

Sifatni o‘rganish metodikasi uning lingvistik xususiyatlariga asoslanadi, chunki sifat predmetning belgilarini (rangi, shakli, hajmi, ko‘rinishi, ta‘mi, hidi va h.k.) bildiradi. Sifatning leksik ma‘nosini tushunish uchun 1-sinfda bolalar sifatning ot bilan bog‘lanishini o‘rganadilar. O‘quvchilar predmetning belgisini aytib, so‘rov yordamida sifat va ot birikmasini aniqlashni o‘rganadilar. Keyingi sinflarda bu bog‘liqlik yanada chuqurlashtiriladi. Shunday qilib, sifatni o‘rganish leksik va grammatik (morfologik va sintaktik) jihatdan amalga oshiriladi.

Boshlang‘ich sinflarda sifat quyidagi izchillikda o‘rganiladi: [3]

1. Sifat bilan tanishtirish (1-sinf),
2. Sifat haqida tushuncha berish (2-sinf),
3. Grammatik materialga bog‘liq sifatlar yozilishini o‘zlashtirish (3-sinf).

Birinchi bosqichda sifatning leksik ma‘nosi va so‘roqlariga diqqat qaratiladi. Predmetning belgilarini, uning rangini, mazasi va shaklini ta‘riflash uchun sifatlar tanlanadi. Shunday qilib, o‘quvchilar sifatlarni tanib olish uchun predmetlar yoki rasmlar asosida mavzularni muhokama qiladilar. Bu darslarda ko‘rgazmalar, mashqlar va suhbatlar keng qo‘llaniladi.

Ikkinchi bosqichda o‘quvchilarga sifatning grammatik xususiyatlari va uning nutqdagi rolini o‘rganish muhimdir. O‘quvchilar sifatlarni mazmunga mos ravishda ishlatishni, fikrni aniq ifodalash uchun sifatlardan foydalanishni o‘rganadilar. Ayniqsa, sifatlarning belgilovchi va bog‘lanuvchi xususiyatlari ta‘kidlanadi. O‘quvchilar predmetning rangini, shaklini, mazasi va boshqa sifatlarini tanlab, ularni tasvirlashda foydalanadilar.

Uchinchi bosqichda sifatlar ustida ishlash davom etadi, ayniqsa, –roq qo‘shimchasi bilan yasalgan sifatlar va ba‘zi maxsus sifatlarni to‘g‘ri yozish o‘rgatiladi. Bu bosqichda o‘quvchilarning mustaqil ishlash qobiliyati oshadi, mashqlar asta-sekin murakkablashtiriladi. Mashq materialida yaxshiroq, aqilliroq

kabi sifatlar hamda tip-tiniq, sap-sariqiq kabi sifatlar ishlatiladi. Bu jarayon sifatni o‘rganishda uning ot bilan o‘zaro bog‘lanishini aniq tushunishni ta‘minlaydi.

Sifat so‘z turkumining grammatik POS teglashidagi ahamiyati tilshunoslikda alohida o‘rganilayotgan sohalardan biridir. Sifatlar tilning eng muhim qismlaridan biri bo‘lib, ular nafaqat ob‘ektning xususiyatlarini ta‘riflaydi, balki uning boshqa ob‘ektlar bilan taqqoslanishiga imkoniyat yaratadi. Sifatlarni to‘g‘ri tanib olish va belgilash, so‘zlarni to‘g‘ri tushunish va tilni qayta ishlashda aniq natijalarga erishish uchun zarur hisoblanadi.

1-jadval. O‘zbek tilidagi POS teglar ro‘yxati (qisqa variant)[4]

№	POS teg	Ingliz	Belgilanishi
1	Ot	Noun	N
2	Sifat	Adjective	JJ
3	Fe‘l	Verb	VB
4	Son	Number	NUM
5	Ravish	Adverb	RR
6	Olmosh	Pronoun	P
7	Bog‘lovchi	Conjunction	C
8	Ko‘makchi	preposition	II
9	Yuklama	particle	Prt
10	Modal so‘z	modal words	MD
11	Undov so‘z	exclamatory words	UH
12	Taqlid so‘zlar	imitation words	IM

Grammatik POS Teglashning muhimligi

POS teglash (Part of Speech tagging) – bu tabiiy tilni qayta ishlashning eng muhim qadamlaridan biridir. Bu jarayon so‘zlarni ma‘lum bir grammatik turkumlarga ajratishni anglatadi. Sifat so‘zlarining to‘g‘ri aniqlanishi, tilni qayta ishlash tizimlari uchun asosiy maqsad bo‘lib, so‘zning sintaktik rolini aniqlash va uning kontekstiga qarab grammatik funktsiyasini belgilashni talab qiladi. Sifatlar odatda, nomlar bilan bog‘lanib, o‘zgaruvchi xususiyatlarni ko‘rsatadi, va bu xususiyatlarni to‘g‘ri tasvirlash uchun zamonaviy tizimlar yuqori aniqlikda ishlashini ta‘minlash zarur[5].

Sifatlarni belgilashdagi qiyinchiliklar

Sifat so‘zlarini to‘g‘ri belgilashda bir qancha qiyinchiliklar mavjud. Masalan, ba‘zi sifatlar ko‘p ma‘nolilikka ega bo‘lishi mumkin. Bu holatlarda kontekstni hisobga olgan holda sifatning qanday funktsiya bajarayotgani aniqlanishi kerak. Misol uchun, "katta" so‘zi biror narsa o‘lchamining kattaligini bildirishi mumkin, ammo ba‘zan bu sifat “katta odam” kabi iboralarda ob‘ektning ijtimoiy yoki siyosiy

holatini ko‘rsatishi mumkin. Shunday qilib, sifatlarni kontekstga qarab to‘g‘ri belgilash muhimdir.

Yana bir qiyinchilik shundaki, ba‘zan sifatlar boshqa so‘zlar bilan bog‘lanib, boshqa grammatik funktsiyalarni bajarishi mumkin. Misol uchun, "yaxshi" so‘zi sifat sifatida ishlatilsa-da, ba‘zan u modifikator sifatida ishlatilishi mumkin. Bu holatlarda POS teglash tizimi sifatning qanday rolda ishlashini aniqlashda qiyinchiliklarga duch kelishi mumkin.

Zamonaviy texnologiyalar va yondashuvlar

Bugungi kunda sifatlarni grammatik POS teglash jarayonida qo‘llaniladigan texnologiyalar yangi imkoniyatlar yaratmoqda. Mashina o‘rganish (machine learning) va chuqur o‘rganish (deep learning) algoritmlari sifatlarni aniqlashda yuqori samaradorlikni ta‘minlashga yordam beradi. Neyron tarmoqlarining rivojlanishi POS teglash tizimlarining aniqligini sezilarli darajada oshirgan. Ayniqsa, chuqur o‘rganish usullari, matnning ma‘nosini yaxshiroq tushunish va sifatlarning to‘g‘ri funktsiyasini aniqlashga imkon beradi. Biroq, bu texnologiyalarni samarali qo‘llash uchun katta hajmdagi ma‘lumotlar va kuchli hisoblash quvvatiga ehtiyoj bor.

Tadqiqotlar shuni ko‘rsatdiki, sifatlarni grammatik POS teglash jarayonida qo‘llaniladigan zamonaviy yondashuvlar yuqori darajada samarali bo‘lishi mumkin. Ayniqsa, chuqur o‘rganish va mashina o‘rganish usullari yordamida sifatlarning to‘g‘ri aniqlanishi va tahlil qilinishi tilni qayta ishlash tizimlarining aniqligini oshiradi. Biroq, bu usullarni qo‘llashda kontekstual omillarni, ya‘ni so‘zlarning ma‘nosini va sintaktik roli o‘zgarishini inobatga olish zarur. Bu o‘z navbatida, POS teglash tizimlarining aniqligini va samaradorligini oshirishga imkon beradi.

Shuningdek, sifatlar o‘zgaruvchan va ko‘p ma‘noli bo‘lgani sababli, POS teglash tizimlari turli kontekstlarni hisobga olib, ularni aniq ajratib belgilashni talab qiladi. Bu o‘z o‘rnida, texnologiyalarni takomillashtirish va yanada yuqori aniqlikni ta‘minlashga olib keladi[6].

Takliflar:

1. **Grammatik belgilash tizimlarini takomillashtirish**: POS teglash tizimlarining samaradorligini oshirish uchun chuqur o‘rganish usullarini qo‘llashni davom ettirish va ularni yanada takomillashtirish zarur.
2. **Kontekstual yondashuvni kuchaytirish**: Sifatlarning ma‘nosini aniqlashda kontekstni hisobga olish tizimning aniqligini oshiradi.
3. **Ma‘lumotlar to‘plamlarini kengaytirish**: Yangi va yanada diversifikatsiyalangan ma‘lumotlar to‘plamlarini yig‘ish sifatlarni to‘g‘ri aniqlashda yordam beradi.

4. **Mashina o‘rganish modellarini yanada takomillashtirish**: Sifatlarni grammatik belgilashda mashina o‘rganishning yangi metodlarini qo‘llash va ularni yuqori darajada avtomatlashtirish tizim samaradorligini oshiradi.

Natijada, sifatlarni grammatik POS teglashda zamonaviy texnologiyalarni to‘g‘ri qo‘llash tilni qayta ishlash va lingvistik tahlilni yanada samarali va aniq qilish imkonini beradi.[7]

Sifatlarni tahlil qilish, umumiy grammatik ma‘noni belgilashni boshlaydi. Sifatlar o‘z xususiyatlariga ko‘ra ikki turga bo‘linadi: 1) asliy sifat, 2) nisbiy sifat. Sifat turkumi sintaktik jihatdan keng emas, lekin uning grammatik xususiyatlari muhimdir. Sifat turkumida egalik kategoriyasi "keyingi sifatni oldingi so‘zga bog‘lash, mansubiyat va xoslik ma‘nolarini ifodalash" orqali aniqlanadi. Kesimlik kategoriyasi sifat turkumida juda cheklangan darajada namoyon bo‘ladi va ko‘pincha bog‘lama vositasi orqali ifodalanadi. Sifat so‘z turkumining teglanishi kengaytirilgan tizimda quyidagicha tasniflanadi: − o‘lik: [sifat], [nisbiy sifat], [oddiy daraja], [yasama sifat], [sodda sifat], [jismoniy harakat LMG]; − tuzuk: [sifat], [asliy sifat], [oddiy daraja], [tub sifat], [yasama sifat], [xususiyat LMG]; − yiroq: [sifat], [asliy sifat], [oddiy daraja], [sodda sifat], [tub sifat], [xususiyat LMG]; − yo‘g‘onroq: [sifat], [asliy sifat], [qiyosiy daraja], [sodda sifat], [tub sifat], [shakl LMG]. O‘tkazilgan tajribalar natijalariga ko‘ra, o‘zak ma‘lumotlarni sintaktik vazifalar bilan birlashtirish morfologik jihatdan boy til uchun POS teglash natijalarini yaxshilaydi, bu esa NLP vazifalarini samarali hal qilishga yordam beradi. O‘zbek tili korpusi matnlarini POS teglash xususiyatlari va kichik kategoriyalar ro‘yxati keltirilgan. Fe‘l va sifat turkumlari grammatik kategoriyalari, tasnifi va misollar izohlangan. Maqola o‘rganilgan usullarni o‘zbek tili morfologik analizatori va boshqa murakkab NLP vazifalarida qo‘llash mumkinligini ko‘rsatadi[8].

Xulosa

Sifat so‘z turkumini grammatik POS teglash jarayoni tilshunoslikda muhim o‘rin tutadi, chunki sifatlar predikatdagi boshqa so‘zlar bilan bog‘lanib, ularning semantik va sintaktik rolini aniqlashga yordam beradi. Sifatlarni to‘g‘ri teglash orqali o‘quvchilar va tadqiqotchilar lingvistik tuzilmalarning yanada aniqroq va samarali tahlilini amalga oshiradilar. Sifatlarning leksik va grammatik xususiyatlarini o‘rganish, ularni nutqdagi o‘rni va qo‘llanilishining izchil tarzda o‘rganilishi, tilning murakkab strukturasi haqida kengroq tasavvurga ega bo‘lishga imkon yaratadi.

Shuningdek, sifatlarni grammatik teglash jarayonida so‘zning morfologik xususiyatlari, jumladan, ularning qo‘shimchalar yordamida qiyosiy yoki orttirma darajalarga o‘tishini, shuningdek, sifatlar va otlarning o‘zaro bog‘lanishini aniqlash

muhimdir. Bu metodik yondashuv til o‘qituvchilariga va tilshunoslarga sifatlarni o‘rganishda to‘g‘ri va samarali yo‘nalishlarni belgilashga yordam beradi.

Shunday qilib, sifat so‘z turkumini grammatik POS teglash, nafaqat tilshunoslik, balki til o‘rganish va ta‘lim jarayonlarida ham samarali vosita sifatida xizmat qiladi, o‘quvchilarga sifatlarning nutqda qanday ishlashini o‘rgatishda muhim o‘rin tutadi.

Foydalanilgan adabiyotlar:

1. Ahmadaliyev, B. (2020). O‘zbek tili grammatikasi. Toshkent: O‘zbekiston Milliy Universiteti.
2. Karimov, S. (2018). Tildagi morfologik va sintaktik tuzilmalar. Toshkent: Sharq.
3. Tohirov, M. (2019). O‘zbek tilida sifat so‘z turkumining sintaktik o‘rni. Toshkent: Ma‘naviyat.
4. Mirzayev, A. (2021). Grammatik POS teglash va uning tilshunoslikdagi ahamiyati. Tashkent: Science and Education.
5. Vaxabov, A. (2017). Tilmashinalar va morfologiya. Toshkent: Universitet.
6. Shamsiyev, Z. (2022). Sifat va uning grammatik xususiyatlari. Toshkent: O‘zbekiston Yozuvchilar Uyushmasi.
7. Smith, J. (2016). Linguistic Analysis of Adjectives in Syntax and Semantics. Oxford University Press.
8. Gorman, J. (2015). Understanding POS Tagging in Natural Language Processing. Cambridge: Cambridge University Press.

ABDULLA QAHHORNING “BEMOR” HIKOYASI PARALLEL KORPUSI

Bekmurzayeva Maftuna Baxtiyor qizi,
Kompyuter lingvistikasi fakulteti 1-kurs magistranti
baratovamaftuna09@gmail.com
ToshDO‘TAU

Anotatsiya. Mazkur maqolada korpus lingvistikasi, xususan, parallel korpuslar asosida badiiy matnlarni lingvistik tahlil qilish imkoniyatlari ko‘rib chiqilgan. Abdulla Qahhorning “Bemor” hikoyasi va uning ingliz tiliga tarjimasi asosida tuzilgan ikki tilli parallel korpus tahlili orqali tarjima jarayonidagi leksik, stilistik va grammatik muvofiqliklar tahlil qilinadi. Tadqiqotda SketchEngine korpus menejeri yordamida yaratilgan korpusdagi kalit so‘zlar, konkordanslar va N-gramm tahlil natijalari asosida tarjimaning lingvistik xususiyatlari, milliy leksik birliklarning o‘ziga xos tarjima strategiyalari yoritiladi. Qiyosiy tahlillar orqali o‘zbek va ingliz korpuslarining texnik va lingvistik imkoniyatlari o‘rganilgan, mavjud farqlar ilmiy jihatdan asoslab berilgan.

Abstract: This article explores the possibilities of linguistic analysis of literary texts based on corpus linguistics, particularly parallel corpora. The study focuses on a bilingual parallel corpus compiled from Abdulla Qahhor’s short story “Bemor” and its English translation. Through the analysis of lexical, stylistic, and grammatical consistencies in the translation process, the paper demonstrates how linguistic features are preserved or transformed. Using the SketchEngine tool, the research analyzes key words, concordances, and N-gram structures to reveal translation strategies, particularly in dealing with culturally specific lexical items. A comparative analysis identifies technical and linguistic differences between the Uzbek and English corpora, supported by empirical evidence.

Аннотация: В данной статье рассматриваются возможности лингвистического анализа художественных текстов на основе корпусной лингвистики, в частности параллельных корпусов. В качестве материала анализа используется двуязычный параллельный корпус, составленный на основе рассказа Абдуллы Каххара «Бемор» и его перевода на английский язык. В процессе исследования анализируются лексические, стилистические и грамматические соответствия между оригиналом и переводом. С помощью программы SketchEngine проводится анализ ключевых слов, конкордансов и N-грамм, что позволяет выявить особенности перевода и стратегии передачи национально-культурной лексики. В результате сравнительного анализа выявлены технические и лингвистические различия между узбекским и английским корпусами, подтверждённые научными данными.

Kalit so‘zlar: *korpus lingvistikasi, parallel korpus, tarjimashunoslik, badiiy matn, konkordans, N-gramm, SketchEngine, kalit so‘zlar tahlili, lingvistik moslik, o‘zbek-ingliz tarjimasi.*

Kirish

Korpus lingvistikasi – bu tabiiy tilni ilmiy asosda o‘rganish va tahlil qilish uchun katta hajmdagi matnlar majmuasidan, ya’ni korpuslardan foydalanishga asoslangan fan sohasidir. Bu yo‘nalish XX asrning ikkinchi yarmidan boshlab rivojlana boshlagan bo‘lib, hozirgi kunga kelib jahon bo‘yicha lingvistika, sun’iy intellekt, tarjima tizimlari, ta’lim va axborot texnologiyalarida keng qo‘llanilmoqda.

Korpus lingvistikasining paydo bo‘lishi, eng avvalo, ingliz tilida yaratilgan dastlabki elektron matnlar bazalari bilan bog‘liq. Jumladan, 1964-yilda AQShda yaratilgan Brown Corpus bu sohadagi ilk yirik loyiha hisoblanadi. Keyinchalik, Buyuk Britaniyada yaratilgan British National Corpus (BNC) va AQShda tuzilgan Corpus of Contemporary American English (COCA) kabi korpuslar butun dunyo tilshunoslari uchun tayanch resurslarga aylandi.

Bugungi kunda korpus lingvistikasi dunyoning ko‘plab yetakchi universitetlari va ilmiy markazlarida chuqur o‘rganilmoqda. Buyuk Britaniyada Lancaster va Oxford universitetlari, AQShda Brigham Young va Stanford universitetlari, Germaniyada Leipzig universiteti, Rossiyada esa Moskva Davlat universiteti kabi markazlar ushbu sohada yetakchilik qilmoqda.

Korpus lingvistikasi faqat nazariy soha emas – u til materiallarini avtomatik tahlil qilish, sun’iy intellekt modellari yaratish, tarjima tizimlarini rivojlantirish, nutqni tanish texnologiyalarini ishlab chiqish va o‘quv materiallarini tuzishda amaliy vosita sifatida keng qo‘llaniladi. Masalan, EuroParl, OpenSubtitles, Universal Dependencies, Sketch Engine kabi xalqaro loyihalar orqali turli tillarning matnlari tizimli tarzda yig‘ilib, avtomatlashtirilgan tarzda tahlil qilinmoqda.

Korpus lingvistikasi orqali tilga doir qarorlar empirik asosda, ya’ni real til qo‘llanilishi asosida qabul qilinadi. Bu esa uni an’anaviy intuitiv yoki nazariy tahlil metodlaridan ustun qiladi. Bugungi kunda bu sohada Mark Davies, Tony McEnery, Geoffrey Leech, Douglas Biber kabi olimlar jahon ilm-fani taraqqiyotiga katta hissa qo‘shmoqda.

Asosiy qism

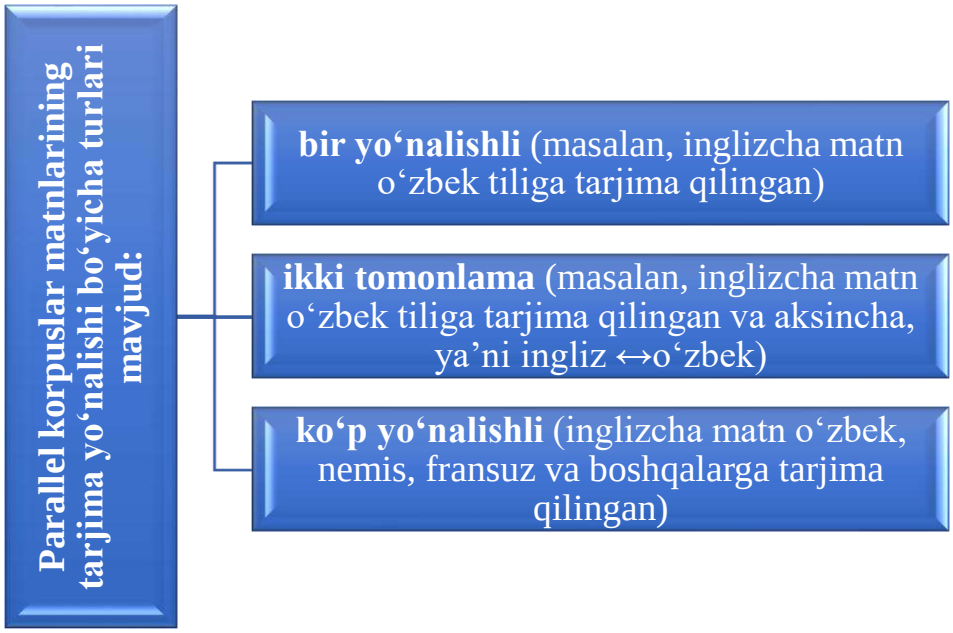
Korpus (lot. *Corpus* – tan, to‘plam) – tilshunoslikda tabiiy til materiallarining katta hajmdagi elektron to‘plamidir. Bu to‘plamlar ilmiy tadqiqotlar, sun’iy intellekt tizimlari, mashina tarjimasi, nutqni aniqlash va boshqa ko‘plab maqsadlar uchun ishlatiladi.

Korpus tabiiy tilning yozma yoki og‘zaki shakllarini o‘z ichiga olgan, maxsus tamoyillar asosida yig‘ilgan, raqamlashtirilgan, ko‘pincha morfologik, sintaktik, semantik jihatdan annotatsiyalangan, kompyuterda qayta ishlash uchun qulay formatda tuzilgan katta hajmdagi til materiallari to‘plami.

Korpuslar zamonaviy kompyuter lingvistikasi, sun’iy intellekt, ma’lumotlar tahlili, tilshunoslik va o‘qitish texnologiyalarida keng foydalaniladi.

Parallell korpus – bu ikki yoki undan ortiq tilda yozilgan bir xil mazmunga ega bo‘lgan matnlar to‘plami bo‘lib, har bir til versiyasi bir-biriga mos qilingan (ya’ni, segmentlar yoki gaplar darajasida muvofiqlashtirilgan) holda saqlanadi. Oddiy qilib aytganda, parallell korpus – bir tilning matnlari va ularning tarjimalaridan iborat bazadir. Bu turdagi korpuslar mashina tarjimasi, kontrastiv tilshunoslik, tarjima tadqiqotlari, va ko‘p tilli til modellari yaratishda keng qo‘llaniladi.

Parallel korpuslar yohud tarjimalar korpusi (ingl. parallel / translation corpus) – bu asl tildagi matnlar va uning boshqa tildagi tarjimasi mavjud korpus. Qoidaga ko‘ra, bunday korpuslar faqat ikkita tilda tayyorlanadi, shu bois undagi o‘zaro mutanosib matnlar bitekst ham deyiladi. Bimatn (bitekst) – asliy til matni va uning boshqa til(lar)dagi tarjimasi mavjud matnlar. Parallel korpuslar – partekst – bimatn – bitekst[*Abjalova, 2023:38*].



Parallell korpusni yaratish bir qator muhim bosqichlarni o‘z ichiga oladi. Bu jarayon texnik va lingvistik yondashuvlarni uyg‘unlashtirgan holda olib boriladi.

Parallell korpusni yaratish bir qator muhim bosqichlarni o‘z ichiga oladi. Bu jarayon texnik va lingvistik yondashuvlarni uyg‘unlashtirgan holda olib boriladi.

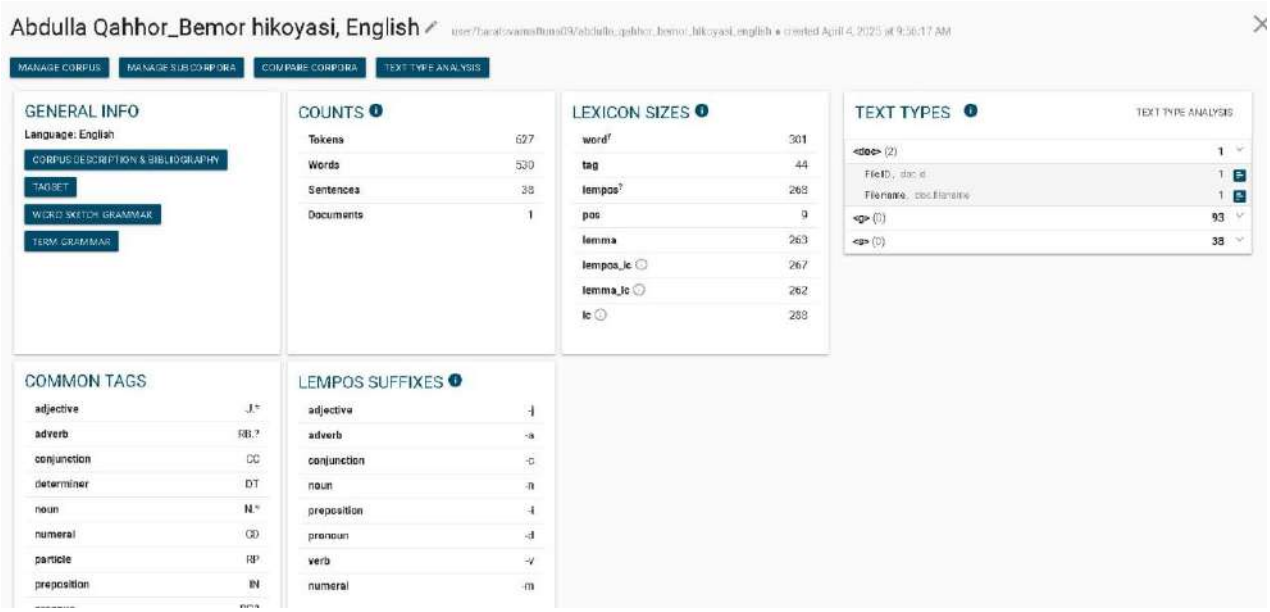
Bunda dastlab matnlarni tanlash, tozalash, segmentlash, moslashtirish, anotatsiyalash, formatlash va saqlash ishlari bajariladi. Parallel korpus yaratishda ayrim qiyinchiliklarga duch kelishimiz mumkin. Bunda avvalo, yuqori sifatli tarjimalar topish va ularni tahlil qilishda yuzaga keladigan muammolarni aytishimiz mumkin. Ikkinchidan, ko‘p tilli korpuslarni yaratish tildagi matnlarning o‘zaro muvofiqlashgan(alignment) shaklini talab qiladi. Ammo, Parallel korpusning bir qator afzalliklari ham mavjud. Misol uchun, zamonaviy va keng qamrovli yondashuvlarni amalga oshiradi. Bu esa bir necha tillar yuzasidan keng qamrovli tahlillarni olib borish mumkin. U tilni o‘rganish va tarjimani tahlil qilishda qo‘llaniladi. Badiiy asar bo‘yicha parallel matnlar korpusi, avvalo, ikki yoki bir necha tillar grammatikasidagi o‘xshash va farqli jihatlarni tahlil qilish imkonini beradi. Parallel matnlar korpusi har ikki tildagi badiiy uslub xususiyatlarini qiyoslashda ahamiyatlidir[*Akramdjanova, 2024:35*].

Quyida yaratilgan Abdulla Qahhorning “Bemor” hikoyasi va uning ingliz tiliga tarjima qilingan nusxasi asosidagi parallel korpus ham tilshunoslik, tarjimashunoslik va adabiyotshunoslik uchun muhim ahamiyatga ega. Korpusni yaratishda “SketchEngine” dasturidan foydalanilgan. Korpus haqidagi statistik ma’lumotlarni quyidagi jadvallarda ko‘rish mumkin.

1-jadval. O‘zbek-ingliz hamda Ingliz-o‘zbek parallel korpusi ma’lumotlari

		O‘zbek-ingliz PK	*Ingliz-O‘zbek PK*
1	*Words*	*462*	*530*
2	*Tokens*	*640*	*627*
3	*Sentences*	*37*	*38*
4	*Documents*	*1*	*1*

1-rasm. Parallel korpusda o‘zbek tilidagi asar ma’lumotlari



2-rasm. Parallel korpusda ingliz tilidagi asar ma’lumotlari

Ushbu jadvalda bir asarning ikki tildagi nusxalaridagi farqlarni yaqqol ko‘rish mumkin. Jumladan, parallel korpusda asar ingliz tiliga tarjima qilinganda o‘zbek tilidagi nusxasida mavjud so‘zlarning 60 tadan ziyodroqqa ko‘payganligini ko‘rish mumkin. Yoki tarjima nusxada tokenlarning qariyb 13 taga qisqargani ko‘zga tashlanadi. Shuningdek tarjima matnda gaplar soni ham 1taga ko‘p.

Ta’kidlash joizki, bugungi kunda mashina tarjimasida so‘zma-so‘z tarjima usulidan voz kechilgan bo‘lib, ushbu usul hatto an’anaviy tarjima jarayonida ham o‘zini oqlamagan. Hatto asl matndagi katta bir xatboshi bir necha gap bilan tarjima matnda o‘z mazmunini topgan. Shu bois parallel korpuslarda ham tarjimalar jarayonida jumlama-jumla moslashish bo‘lmasligi mumkin[*Abjalova, 2023:38*].

Parallel korpus tuzishdan avval asarning ikki tildagi ayni nusxalarining elektron formatlari topib olindi. Keyin dasturga muvaffaqiyatli yuklanib, ikki tilli parallel korpus ishga tushdi (1-rasm).

3-rasm. Parallel korpus interfeysi

Parallel korpusdagi asar tahliliga kirishishdan oldin kalit so‘zlarni topib olish lozim bo‘ladi. Quyidagi jadvalda ikki tildagi korpusdan topilgan dastlabki 10 ta kalit so‘zlarni ko‘rish mumkin (2-jadval).

2-jadval. Parallel korpusdagi kalit so‘zlarning qiyosiy tahlili.

O ‘zbek tili PK			***Ingliz tili PK***		
Kalit so‘zlar	***Takrori***	***Foiz***	***Kalit so‘zlar***	***Takrori***	***Foiz***
Sotiboldi	6	0.94	Sotiboldi	9	1.4
doktorxona	3	0.47	Abduganiboy	2	0.32
sotiboldining	3	0.47	chilyosin	2	0.32
chilyosin	2	0.31	devonai	1	0.16
saxarlari	2	0.31	baqavli	1	0.16
beygin	2	0.31	shaydullo	1	0.16
o‘ladigan	1	0.16	banomi	1	0.16
qabzali	1	0.16	bakhshi	1	0.16
pashshalardan	1	0.16	handkerchief	2	0.32
yiroqdan	1	0.16	cabbage	1	0.16

Ushbu jadvalni taqqoslaganda o‘zbek-ingliz va ingliz-o‘zbek parallel korpusidagi kalit so‘zlar tarjimada o‘zgarganini ko‘rish mumkin. Masalan, birgina “Sotiboldi” ismini oladigan bo‘lsak o ‘zbek tilida 6 marta ingliz tilida 9 marta qo‘llanganligini ko ‘rishimiz mumkin. Buning sababi esa o‘zbek tilida “Sotiboldi” va “Sotiboldining” so‘zlari alohida berilgan biri 6 marta biri 3 marta qo‘llangan.

Ingliz tilida esa “Sotiboldining” so‘zi tarjimada “ s' ” bilan qo‘llaniladi shuning uchun ham korpusda bir xil so‘z sifatida olingan. Bundan tashqari, asl tilda *“*Sotiboldi*”, “doktorxona”, “*Sotiboldining*”, “chilyosin”* kabi so‘zlar eng ko‘p qo‘llanilgan bo‘lsa, ingliz tiliga tarjimada *“Sotiboldi”, “abduganiboy”, “chilyosin”, “handkerchief”* kabi otlar ko‘p takrordan foydalanilgan. Ushbu qiyosiy tahlildan yana bir xulosaga kelish mumkin bo‘ladi. Asarning tarjima nusxasida urf-odat, marosim va kishilarning lavozimlari o‘zbek tilidan to‘g‘ridan-to‘g‘ri olib kirilgan. Tarjimon o‘zbek xalqi madaniyati va tarixiga oid ushbu so‘zlar ustida tarjimaga qo‘l urmagan balki o ‘z holicha asarga kiritgan.

Parallel korpuslar tilshunoslikda, xususan, tarjimashunoslikda katta ilmiy-amaliy ahamiyat kasb etadi. Ayniqsa, badiiy matnlar tahliliga doir tadqiqotlarda *parallel konkordanslar* muhim manba vazifasini bajaradi. Konkordans – bu muayyan kalit so‘zning matndagi kontekstda qanday ishlatilganini ko‘rsatib beruvchi tizimli ro‘yxat bo‘lib, parallel korpusda bu ro‘yxatlar ikkala til variantida ham o‘zaro muvofiqlashtirilgan holda taqdim etiladi.

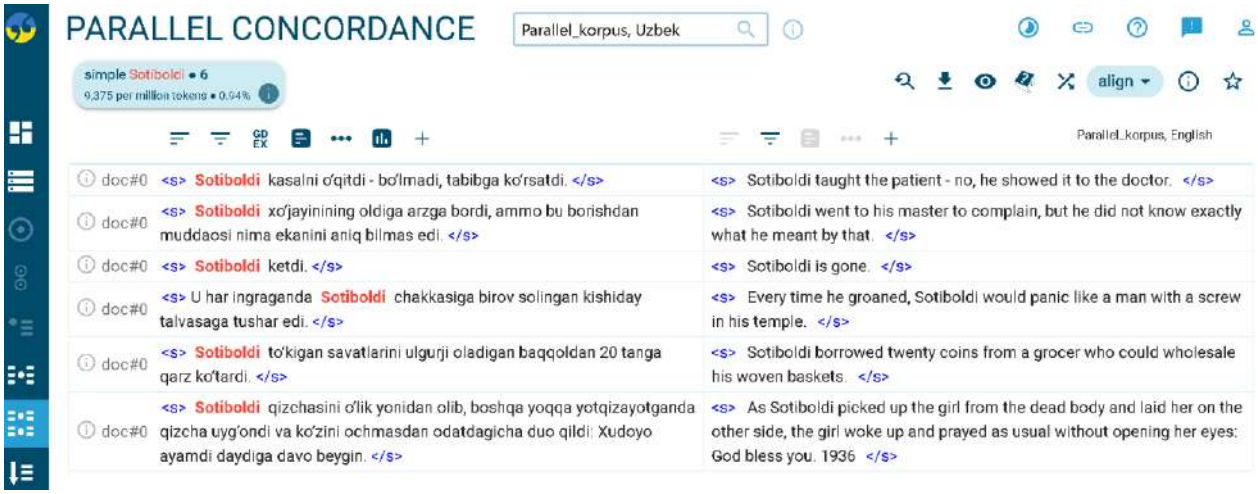
Bunday konkordanslar orqali ikkala til versiyasidagi segmentlarning (gaplar, iboralar) mosligi aniqlanadi, shuningdek, tarjimada yuzaga chiqqan leksik va stilistik o‘zgarishlar tahlil qilinadi. Ayniqsa, milliy koloritga ega bo‘lgan atamalar, toponimlar, shaxs ismlari va ularning tarjimada qanday ifodalangani, qanday strategiyalar asosida ko‘chirilgani o‘rganiladi.

Masalan, *“Sotiboldi”* so‘zini parallel korpusda qidirish orqali quyidagi tahlillar amalga oshirildi:

1. *Segmentlararo muvofiqlik:* Avvalo, ikkala til matnlari segmentlar darajasida to‘g‘ri moslashtirilgani ko‘zga tashlandi. Bu esa korpusga asoslangan tahlilning aniqligini ta’minlaydi.
2. *Kontrastiv leksik tahlil: “Sotiboldi”* so‘zining tarjimada qanday ishlatilgani, uning to‘g‘ridan-to‘g‘ri transliteratsiya qilingani yoki kontekstga mos tarzda stilistik moslashuvga uchragani o‘rganildi.
3. *Semantik tahlil:* Bu ism atrofida ishlatilgan fe’llar va boshqa bog‘lovchi birliklar ikkala tilda qanday semantik tusga ega bo‘lishi aniqlashtirildi.

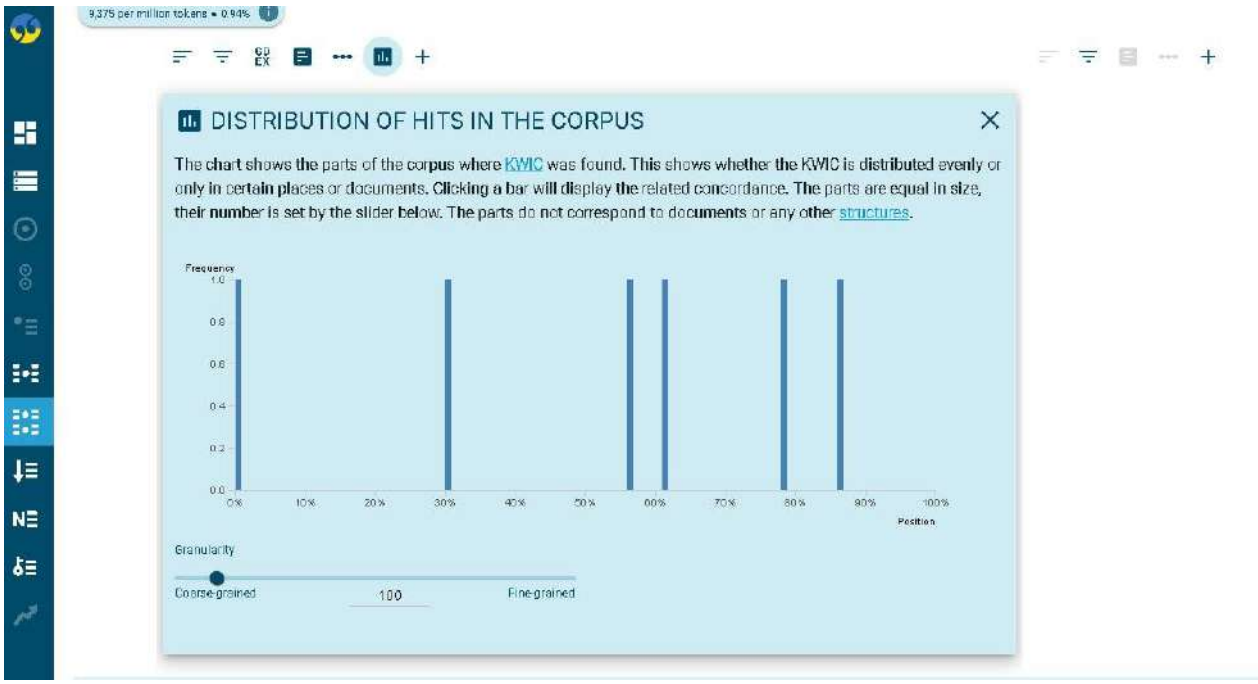
Ushbu tahlillar shuni ko‘rsatadiki, parallel korpuslar faqat tarjima moslamalarini o‘rganish bilan cheklanmay, balki matn mazmuni va strukturasini ko‘p darajali lingvistik yondashuvlar asosida tahlil qilish imkonini beradi. Ayniqsa, badiiy uslubda yozilgan asarlar ustida ishlayotganda, matnning emotsional, madaniy va konnotativ qatlamlarini aniqlashda parallel korpuslar benazir vosita hisoblanadi. Shu jihatdan olib qaralganda, parallel konkordanslar orqali badiiy tarjima usullarini tadqiq qilish, tarjimonning lingvistik tanlovi, madaniy o‘zlashtirish mexanizmlari,

va o‘quvchiga yetkazilayotgan emotsional ta’sir darajasini ham aniqlashga imkon beradi.

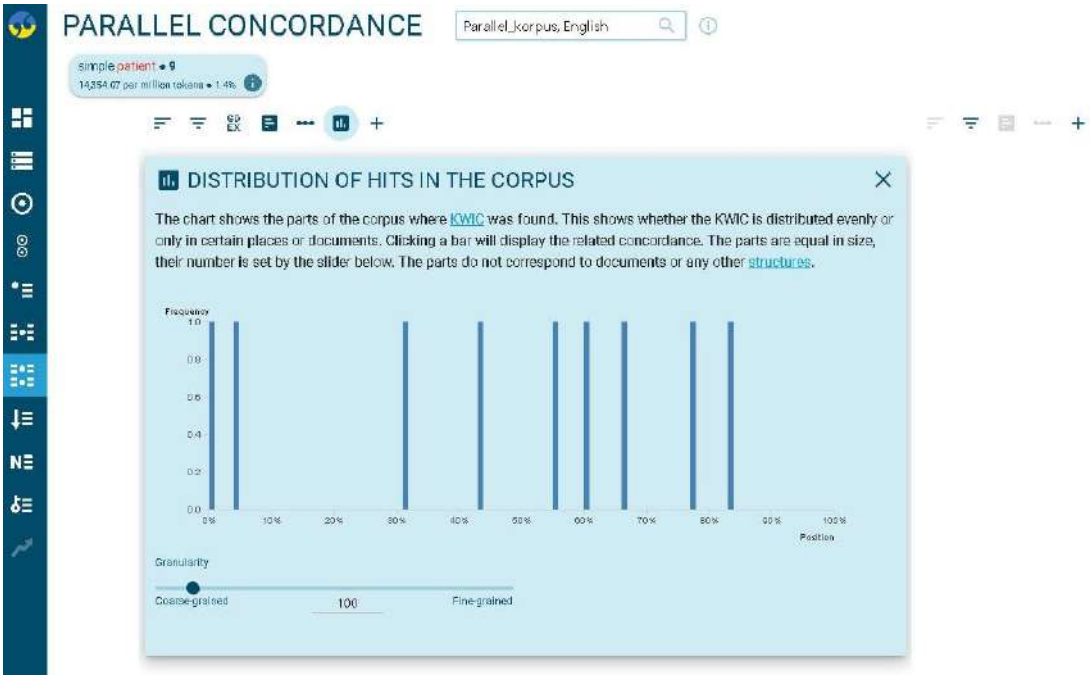


4-rasm. Matnlar muvofiqlashgan parallel korpus.

Keyingi qiyosiy tahlil *“Sotiboldi”* shaxs nomining ikki korpusda ham asarda qo‘llanilganlik darajasini diagrammada ifodalanishi bo‘ldi(5-rasm).



5-rasm. O‘zbek tilli parallel korpusda “Sotiboldi” shaxs nomining qo‘llanilishi.

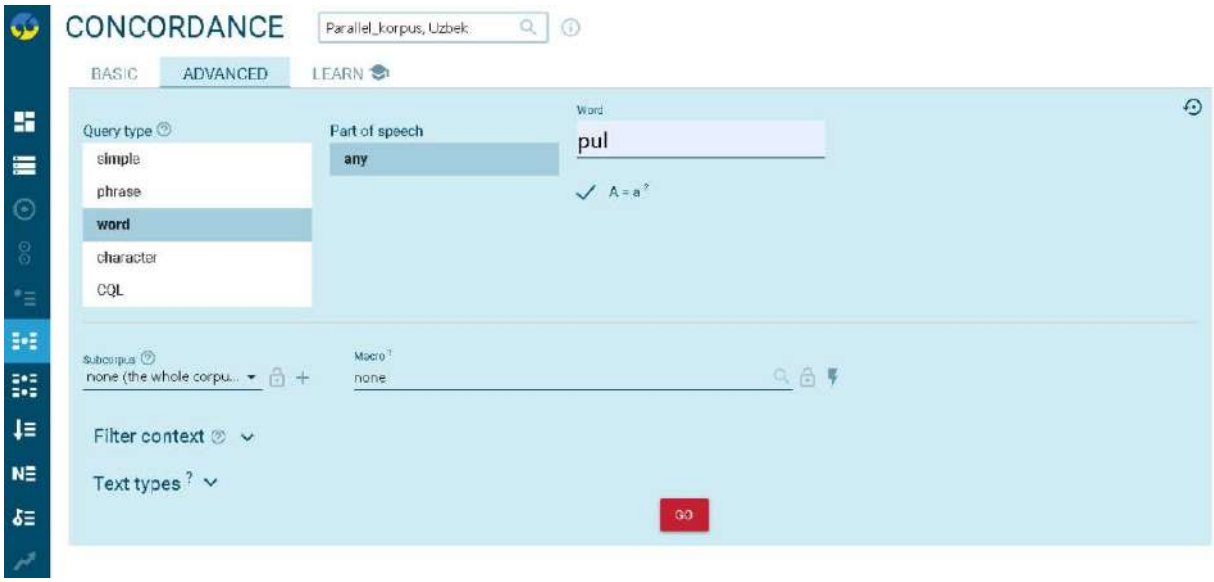


6-rasm. Ingliz tilli parallel korpusda “Sotiboldi” shaxs nomining qo‘llanilishi.

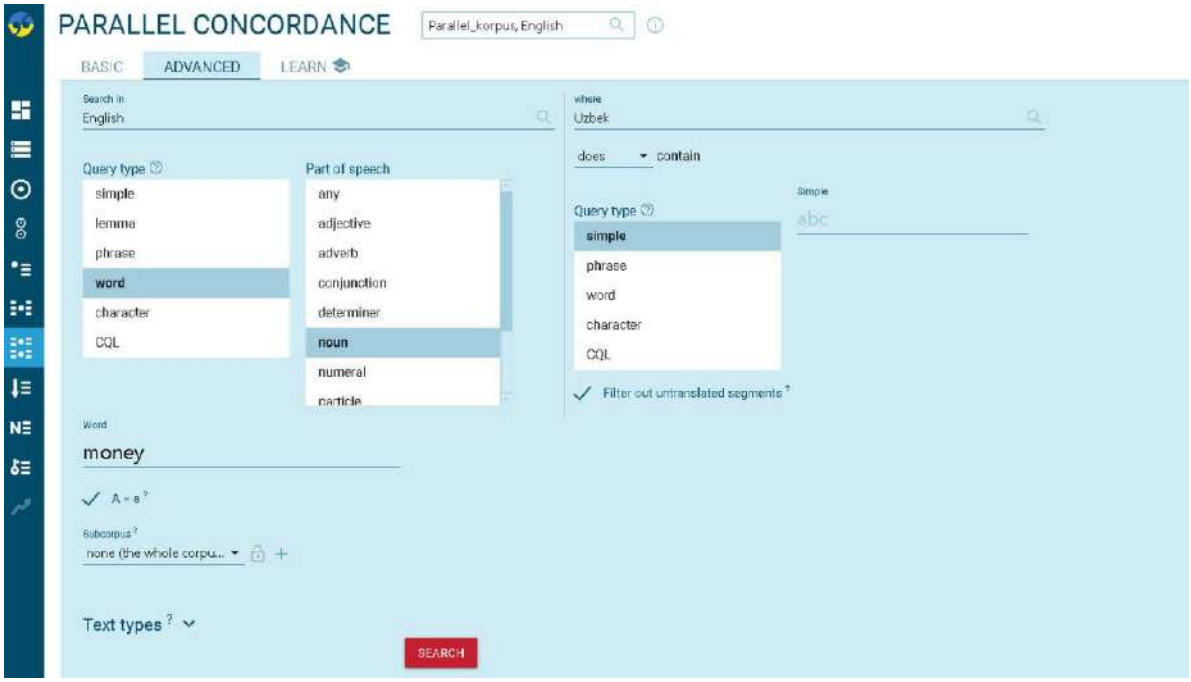
Parallel korpuslar ikki yoki undan ortiq tildagi matnlarni muvofiqlashtirilgan holatda saqlash orqali tilshunoslik tadqiqotlarida muhim o‘rin tutadi. Ayniqsa, ularning ichki qidiruv tizimlari til birliklarini chuqurroq tahlil qilish imkonini beradi. Keyingi bosqichdagi qiyosiy tahlil ingliz va o‘zbek tillaridagi parallel korpuslarning texnik va lingvistik salohiyatini qiyosiy o‘rganish orqali mavjud farqlar aniqlanadi. Dastlabki kuzatuvlar shuni ko‘rsatdiki, har ikki til parallel korpuslarida yakka so‘zlar, frazalar, va turli birliklar bo‘yicha qidiruv imkoniyatlari mavjud. Ya’ni foydalanuvchi biror so‘zni yozish orqali korpusdagi barcha mos segmentlarni kontekst doirasida ko‘rishi mumkin. Bu jihat umumiy konkordans tizimining mavjudligini anglatadi. Biroq, rus tilidagi parallel korpuslarning keyingi bosqichlarda taklif etayotgan imkoniyatlari ancha kengroq va zamonaviyroq texnologiyalarga asoslangani kuzatildi. Masalan, ingliz korpuslarida quyidagi lingvistik atributlar bo‘yicha izlanishlar olib borish mumkin: *so‘zning morfologik tavsifi (so‘z turkumi, jins, son, kelishik); sintaktik vazifasi (otli iboraning ergash gapdagi roli, fe’lning ko‘chirma yoki buyruq shakli va h.k.); matndagi pozitsiyasi (gap boshida yoki oxirida kelishi, kontekstga ta’siri).* Bu holat tahlilning chuqurligini oshiradi va ingliz tilida yozilgan yoki tarjima qilingan matnlarni formal tahlil modellari asosida o‘rganish imkonini yaratadi.

Aksincha, o‘zbek tilidagi parallel korpuslarda bu kabi morfologik yoki sintaktik belgilash (ya’ni, avtomatik annotatsiya) tizimlari hali joriy etilmagan. O‘zbek korpusida faqat matn kontekstida so‘zlarni qidirish va ularning tarjima variantlarini ko‘rish mumkin, biroq ularning grammatik rolini aniqlash, so‘z

turkumiga ko‘ra filtrlash yoki morfologik shakllarini solishtirish imkoniyati mavjud emas. Bu esa, o‘zbek tili uchun yaratilayotgan korpuslarning texnologik va lingvistik darajasini rivojlantirish zarurligini ko‘rsatadi. Xususan, avtomatik tahlil vositalarini joriy etish, so‘z turlarini belgilash tizimlarini ishlab chiqish va sintaktik analizatorlarni integratsiya qilish o‘zbek tilshunosligi uchun dolzarb yo‘nalish hisoblanadi.



7-rasm. Parallel konkordans qidiruv imkoniyatlari. (O‘zbek tili parallel korpusida)

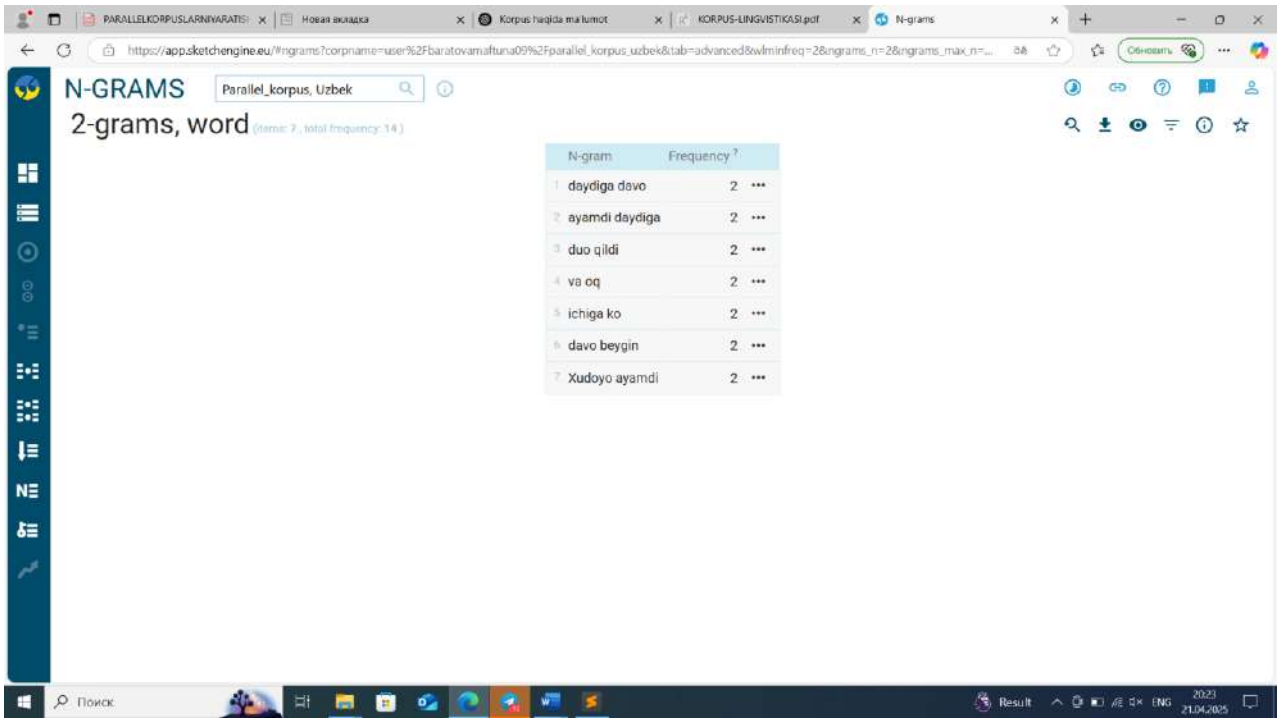


8-rasm. Parallel konkordans qidiruv imkoniyatlari. (Ingliz tili parallel korpusida)

Qiyosiy tahlillar shuni anglatadiki, parallel korpuslardan samarali foydalanish, faqatgina matn segmentlarini muvofiqlashtirish bilan cheklanmay, balki ularni morfologik va sintaktik jihatdan chuqur qayta ishlash imkoniyatiga ham bevosita bog‘liqdir.

Navbatdagi tahlil bosqichlari parallel korpuslarning yana bir muhim funksional imkoniyatini namoyon bo‘ldi. N-gramm tahlili orqali matnda ishlatilgan turli leksik birliklarning qanday ketma-ketlikda va qanday kontekstda qo‘llanilgani aniqlandi. Bu usul orqali matn ichidagi so‘zlar o‘rtasidagi bog‘liqliklar, ularning birgalikda ishlatilish tendensiyalari va ustuvorlik darajalari aniqlab boriladi. Shu orqali badiiy matnning lingvistik xususiyatlari, ya’ni muallif uslubi, matnning leksik boyligi, frazeologik birliklarning ishlatilishi va stilistik tuslari yanada ravshanlashadi. Shuningdek, bu tahlil natijalari orqali tarjimaning sifatini baholash, ikki til o‘rtasidagi ifoda vositalarining qanday moslashtirilgani haqida xulosa chiqarish mumkin bo‘ladi.

Masalan, “Bemor” asarining asl tilida 2-grammlardan foydalanib ko‘rdik. Bunda eng ko‘p qo‘llanilgan so‘z juftliklari, asosan, *“daydiga davo”, “ayami daydiga”, “duo qildi”, “davo beygin”* kabilar bo‘lsa, ingliz tilidagi tarjima nusxasida faqatgina *“in the”* birikmasi 5 marotaba qo ‘llanilgan.



9-rasm. O‘zbek parallel korpusida N-gramlar asosida qidiruv.

10-rasm. Ingliz parallel korpusida N-gramlar asosida qidiruv.

3-jadval. 2-gramm asosida qiyosiy tahlil.

	O‘zbek tili PK	Ingliz tili PK
2-gramm natijasi	**7 ta**	**1ta**
Takrori	**2 marta**	**5 marta**

Xulosa

Olib borilgan tadqiqot asosida parallel korpuslar, ayniqsa badiiy matnlar asosida yaratilgan hollarda, tilshunoslik va tarjimashunoslik sohalarida keng tahlil imkoniyatini taqdim etishi aniqlandi. Abdulla Qahhorning “Bemor” hikoyasi va uning ingliz tilidagi tarjimasi asosida tuzilgan parallel korpus yordamida amalga oshirilgan tahlillar shuni ko‘rsatdiki, bunday korpuslar nafaqat matnlar orasidagi grammatik va leksik muvofiqliklarni aniqlash, balki tarjima jarayonida yuzaga keladigan semantik o‘zgarishlar, stilistik yondashuvlar va madaniy unsurlarni ham chuqur o‘rganish imkonini beradi.

Korpus asosida olib borilgan konkordans tahlili orqali matn segmentlarining tarjima variantlari bilan mosligi tekshirildi. Kalit so‘zlar asosidagi qiyosiy izlanishlar tarjimada muhim rol o‘ynagan leksik birliklar soni va ularning kontekstda qo‘llanish xususiyatlarini aniqlashga xizmat qildi. Shuningdek, N-gramm tahlili orqali asardagi so‘zlar orasidagi sintaktik va semantik bog‘lanishlar, frazeologik birliklar va ularning tarjimadagi ifodasi haqida muhim lingvistik xulosalar chiqarildi.

Tadqiqot davomida SketchEngine korpus menejeri asosida tuzilgan korpus texnik imkoniyatlari bo‘yicha ham tahlil qilindi. Aniqlanishicha, ingliz tilidagi parallel korpuslarda morfologik va sintaktik annotatsiyalar, kalit atributlar bo‘yicha filtrlangan qidiruvlar keng joriy etilgan bo‘lsa, o‘zbek tilidagi parallel korpuslar hali bu darajadagi chuqur strukturaviy belgilash imkoniyatlariga ega emas. Bu esa o‘zbek korpus texnologiyalarini yanada takomillashtirish zarurligini ko‘rsatadi. Tadqiqot shuni isbotladiki, parallel korpuslar tarjima sifati, tarjima strategiyasi, madaniy unsurlarni yetkazish darajasi va muallif uslubini tahlil qilishda samarali

vosita hisoblanadi. Ayniqsa, badiiy tarjima misolida bu korpuslar nafaqat tilshunoslik, balki adabiyotshunoslik va madaniyatshunoslik uchun ham muhim empirik manba bo‘lib xizmat qiladi. Kelgusida bunday korpuslar sonini ko‘paytirish, ularni zamonaviy texnologiyalar bilan boyitish va o‘zbek tiliga moslangan avtomatik tahlil vositalarini yaratish ushbu sohadagi ilmiy izlanishlarning sifatini oshirishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. M.Abjalova. Parallel korpuslarni yaratish asoslari / Kompyuter lingvistikasining zamonaviy texnologiyalari-CTCL.2023. –Tashkent, 2023. 35-42-b.

2. Sh.Hamroyeva. Til korpuslarining morfologik tahlilida so‘zshakllarga belgi (разметка) qo‘yish masalalariga doir (ayrim so‘z turkumlari misolida)/ Til, ta’lim, tarjima. – Toshkent-2020. 36-40-b

3. M.Akramdjanova. Parallel matnlar korpusini yaratishning amaliy va nazariy ahamiyati / Ilm-fan va innovatsiya ilmiy-amaliy konferensiyasi(in-academy.uz/index.php/si) –Tashkent, 2023.

4. Abjalova M. Korpus lingvistikasi. [Matn]: uslubiy qo‘llanma / M.A. Abjalova. –Toshkent: Nodirabegim, 2022. – 110 b.

5. Mengliyev B, Hamroyeva Sh. (2020). Korpus lingvistikasi: Korpus tuzish va undan foydalanish. Tashkent

TOHIR MALIK ASARLARIDA SALBIY SO‘ZLARNING STATISTIK TAHLILI

Ibragimova Mohitob Eldor qizi,
Kompyuter lingvistikasi yo‘nalishi IV bosqich talabasi
mohitobibragimova@gmail.com
ToshDO‘TAU

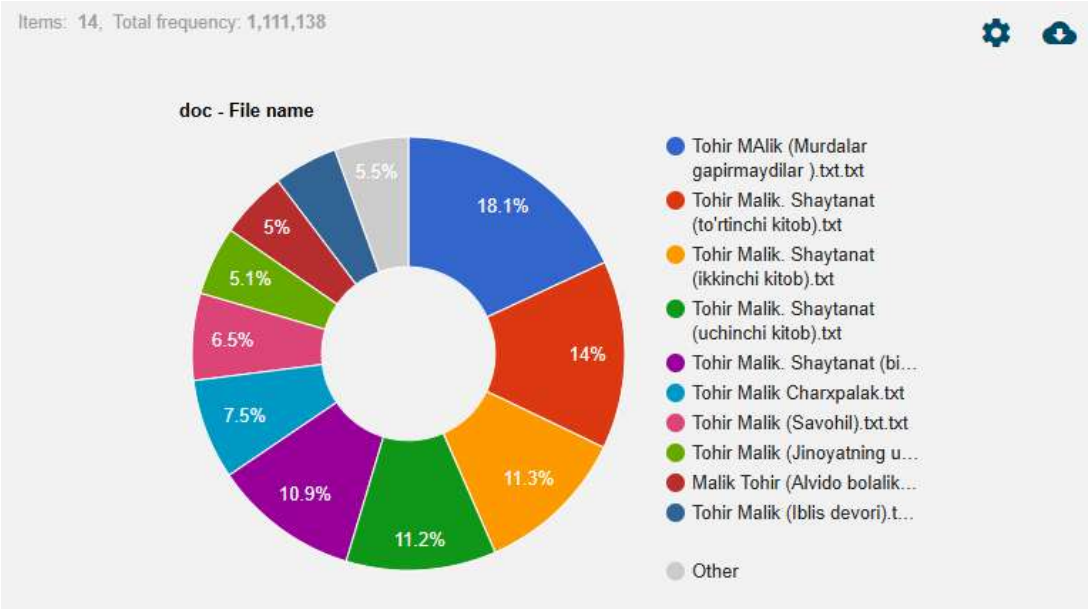
Annotatsiyasi. Maqolada taniqli o‘zbek yozuvchisi Tohir Malik asarlaridagi salbiy so‘zlarning statistik tahlili amalga oshirilgan. Sketch Engine platformasida Tohir Malik asarlarining mualliflik korpusi shakllantirilib, salbiy so‘zlarning chastotasi, ularning tahlili va kontekstual birikish imkoniyatlari o‘rganilgan. Tadqiqot davomida muallif asarlaridagi leksik birliklarning qo‘llanilish chastotasi va lingvistik xususiyatlari aniqlangan. Ushbu ish mualliflik korpuslarini yaratish va ularni lingvistik tahlil qilish bo‘yicha muhim amaliy yondashuvlarni taklif qiladi.

Kalit so‘zlar: *Mualliflik korpuslari, Tohir Malik asarlari, Shekspir mualliflik korpusi, salbiy so‘zlar, statistik tahlil va tarkiblar.*

Mualliflik korpusi – bu matnlar to‘plami bo‘lib, u ma’lum bir muallif yoki mualliflar guruhiga tegishli bo‘lgan asarlarni o‘z ichiga oladi. Ushbu korpus odatda tilshunoslik, adabiyotshunoslik, va kompyuter lingvistikasi sohalarida tilning xususiyatlarini o‘rganish uchun ishlatiladi. Mualliflik korpuslari yordamida, masalan, tilning grammatika, stilistika, leksik va semantik xususiyatlari, muallifning tilga bo‘lgan yondashuvi, va asarlarning o‘ziga xosligini tahlil qilish mumkin.

Jahon yozuvchisi Shekspir mualliflik korpusi – bu ingliz adabiyotining klassik yozuvchisi Uilyam Shekspirning asarlarini o‘z ichiga olgan va uning asarlarini tahlil qilish uchun yaratilgan ma’lumotlar to‘plamidir. Shekspirning asarlari dunyo bo‘ylab keng tarqalgan va uning tilidagi o‘ziga xos xususiyatlar, uslubi, mavzulari, dramatik strukturalari, va boshqa jihatlarni o‘rganish uchun ko‘plab tadqiqotlar olib borilgan. Shekspir mualliflik korpusining tuzilishi: Matnlar to‘plami, semantik va stilistik tahlil, qidiruv va tahlil vositalari[1].

Hozirgi kunda jahon kompyter lingvistikasida ko‘plab korpuslar, xususan mualliflik korpuslari tizimli shaklda yaratilgan. O‘zbek tilshunosligida bugungi kunga qadar A.Navoiy va A.Qahhor mualliflik korpusining lingvistik bazasi yaratilgan bo‘lsa-da, dasturiy ta’minoti ishlab chiqilmagan. Shu sababli ham mualliflik korpusini ham lingvistik, ham dasturiy ta’minotini ishlab chiqish dolzarb masalalardan biri hisoblanadi. Ma’lum yozuvchi asarlari uchun mualliflik korpusini tuzishda muallif ijodiy merosini mukammal o‘rganish zarur. Negaki, korpus tarkibi, bo‘limlar soni muallif ijodining o‘ziga xos xususiyatidan kelib chiqqan holda tuziladi[2].



1-rasm (Tohir Malik asarlari foizlarda diogramma shaklda)

Tohir Malik (1949-2020) – o‘zbek adibi, yozuvchi va jurnalist. “Iblis devori” qissasi 2018-yil “O‘qituvchi”, “Jinoyatning uzun yo‘li” qissasi 2021-yil “Tohir Malik nomidagi nashriyot uyi”, “Talvasa” romani 2012-yil “Sharq” nashriyoti”, “Ozod inson haqidagi qo‘shiq” qissasi 2012-yil “Sharq” nashriyoti, “Charxpalak” qissasi 2014-yil “Sharq” nashriyoti, “Shaytanat 1,2,3,4” romanlari 2019-yil “Sharq” nashriyotlari tomonidan nashr etilgan.

№	Asarlari	Yozilgan yili	Janri	Uslubi	Adabiy turi	Manbasi	Nashriyoti	Nashr yili
1	Shaytanat 1	1992-yil	roman	badiiy	epik	Tohir Malik. Tanlangan asarlar. 18-jild	“Sharq” nashriyoti	2019-yil
2	Shaytanat 2	1995-yil	roman	badiiy	epik	Tohir Malik. Tanlangan asarlar. 19-jild	“Sharq” nashriyoti	2019-yil
3	Shaytanat 3	1997-yil	roman	badiiy	epik	Tohir Malik. Tanlangan asarlar. 20-jild	“Sharq” nashriyoti	2019-yil
4	Shaytanat 4	2000-yil	roman	badiiy	epik	Tohir Malik. Tanlangan asarlar. 21-jild	“Sharq” nashriyoti	2019-yil
5	Talvasa	1992-yil	roman	badiiy	epik	Tohir Malik. Tanlangan asarlar. 11-jild	“Sharq” nashriyoti	2012-yil
6	Alvido bolalik	1989-yil	qissa	badiiy	epik	Tohir Malik. Alvido bolalik. Qissa	Yangi Asr Avlodi	2015-yil
7	Charxpalak	1999-yil	qissa	badiiy	epik	Tohir Malik. Charxpalak. Qissa	“Sharq” nashriyoti	2014-yil
8	Iblis devori	2006-yil	qissa	badiiy	epik	Tohir Malik. Iblis devori. Qissa	O‘qituvchi	2018-yil
9	Jinoyatning uzun yo’li	2004-yil	qissa	badiiy	epik	Tohir Malik. Jinoyatning uzun yo’li. Qissa	Tohir Malik nomidagi nashriyot uyi	2021-yil
10	Murdalar gapirmaydilar	1999-yil	qissa	badiiy	epik	Tohir Malik. Murdalar gapirmaydilar.Qissa	“Sharq” nashriyoti	2019-yil
11	Ozod inson haqidagi qo’shiq	1987-yil	qissa	badiiy	epik	Tohir Malik. Tanlangan asarlar 11-jild	“Sharq” nashriyoti	2012-yil
12	Savohil	1987-yil	qissa	badiiy	epik	Tohir Malik. Savohil. Qissa	O‘zbekiston milliy ensiklopediyasi	2018-yil
13	Qarg'ish		hikoya	badiiy	epik	Tohir Malik.Qarg'ish. Hikoya	Ziyouz.com	
14	Xotira		hikoya	badiiy	epik	Tohir Malik. Xotira. Hikoya	Ziyouz.com	
15	Zilzila		hikoya	badiiy	epik	Tohir Malik. Zilzila. Hikoya	Ziyouz.com	

“Iblis devori”(2018)

Ushbu qissa, ijtimoiy-siyosiy mavzularga qaratilgan bo‘lib, insonning ichki maqsadlari, ishonchliligi va axloqiy dilemmasi haqida hikoya qiladi. “Iblis devori” atamasi diniy va ruhiy jihatlarni ko‘rsatadi, uning asosi – insonning o‘z-o‘zini tanishi va shaxsiy qiyinchiliklarini yengishdir.

“Jinoyatning uzun yo‘li” (2021)

Bu qissada jinoyat va adolat, jinoyatchilikning kelib chiqishi va uning ijtimoiy jihatlari yoritilgan. Tohir Malik asarda jinoyatlarning nafaqat tashqi ko‘rinishini, balki uning inson ruhiyatiga, ichki dunyosiga ta'sirini ham ko‘rsatadi.

“Ozod inson haqidagi qo‘shiq” (2012)

Asarda erkinlik va insonning o‘ziga xosligini qidirishi, shuningdek, jamiyatdagi ijtimoiy cheklovlarga qarshi kurashi mavzusi ko‘tarilgan. Tohir Malik o‘z asarida shaxsning o‘z haq-huquqlarini izlash va jamiyatning ma’naviy qiyinchiliklariga qarshi kurashni yoritadi.

“Charxpalak”(2014)

Ushbu qissada, insonning hayotdagi o‘zgarishlarga bo‘lgan munosabatini, xususan, ijtimoiy tizimdagi va shaxsiy hayotdagi o‘zgarishlarga moslashishning qiyinchiliklarini ko‘rsatadi. Bu asarda, hayotning doimiy aylanishi va insonning bu aylanishdagi o‘rni haqida chuqur tahlil mavjud.

“Shaytanat”1,2,3,4 (2019)

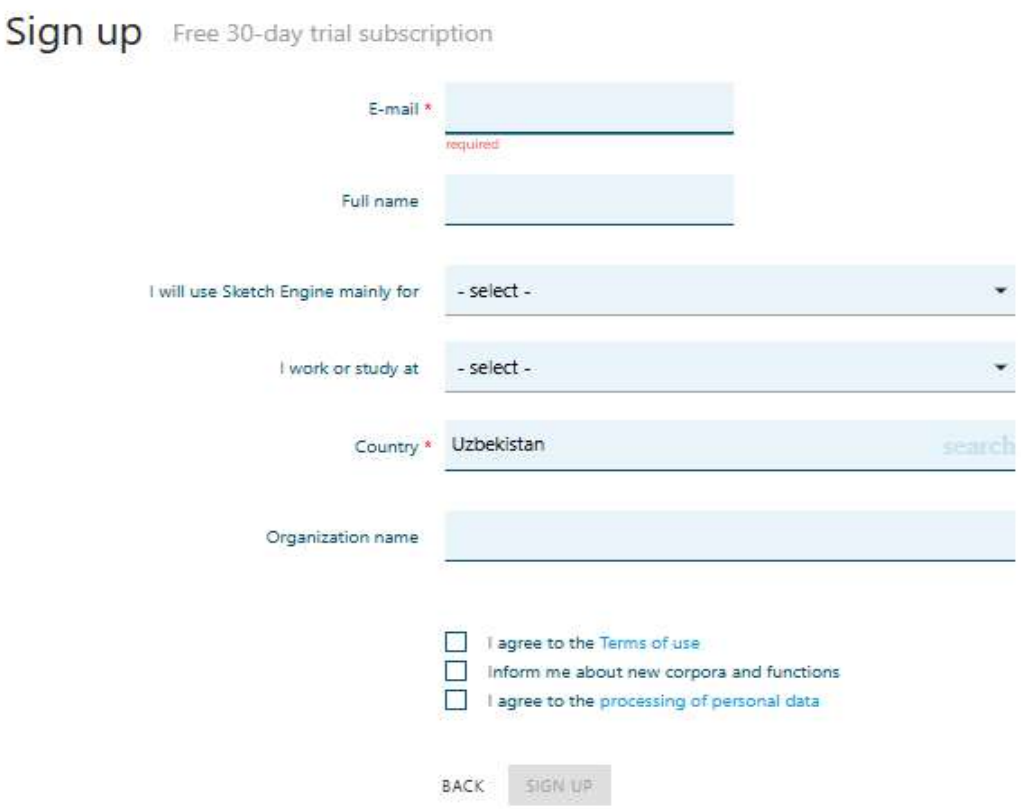
Ushbu romanlar to‘plamida, insonlarning ma’naviy va axloqiy jihatlari, shuningdek, yovuzlik va adolatning tasviri mavjud. Tohir Malikning bu asarlarida ijtimoiy va ma’naviy muammolarni yoritish orqali, insonning ijtimoiy tizimdagi o‘rni va o‘zining ichki kurashlarini aks ettiradi. Romanda insonning qiyinchiliklari va murakkab hayotiy sharoitlarining yomonlashuvi aks etgan.

Tohir Malik asarlarida insonning ichki dunyosi, psixologik holatlari va axloqiy dilemmasi asosiy o‘rin tutadi. Uning asarlari chuqur tahliliy, psixologik va falsafiy xarakterga ega bo‘lib, insonning jamiyatdagi o‘rni, adolat, erkinlik va hayotdagi maqsadlarni izlashning murakkabligini o‘rganadi. U o‘z asarlarida nafaqat individual tajribalarni, balki jamiyatning ijtimoiy tuzilishini ham tasvirlaydi, bu esa o‘quvchiga o‘zgaruvchan va murakkab dunyoda yashashni anglashda yordam beradi. Tohir Malikning asarlari kitobxonni chuqur o‘ylashga undaydi, ijtimoiy va axloqiy masalalar bo‘yicha muhim savollarni qo‘yadi.

Avvalo, Tohir Malik asarlari yig‘ib olindi. Korpus uchun Tohir Malikning 14 asaridan foydalanildi. Ular qissalar, romanlar, hikoyalardan iborat. Birinchi, asarlarni ziyo.uz saytidan foydalangan holda pdf shaklda yuklab olindi va uni Sketch

Engine platformasiga joylashtirish uchun standart shaklga keltirildi. Ikkinchi, avval barcha pdf formatdagi fayllar txt formatga o‘tkaziladi. Uning uchun onlayn convertio.co/ru/pdf-txt/ [3] saytidan foydalaniladi va ba’zi fayllar krilcha shaklda bo‘lganligi tufayli, ularni lotin tiliga o‘tkaziladi, uning uchun ham tayyor onlayn saytlardan foydalansa bo‘ladi[4]. Bu sayt pdf formatdan txt formatga o‘tkazib beradi. Uchinchi, bu txt fayllarni Subline Textdan foydalangan holda tozalab olinadi. Tozalashda, har bir gaplar alohida qilinadi, ba’zi so‘zlar xatoliklari, keraksiz simvollar, tinish belgilari olib tashlanadi.

Barcha asarlar yig‘ilib, tozalanib bo‘lgach, fayllar Sketch Engine platformasiga joylash boshlanadi. Avvalo, ushbu platformadan ro‘yxatdan o‘tish kerak (2-rasm). Ro‘yxatdan o‘tgandan so‘ng platformadan faqat 1 oy mobaynida tekin versiyada ishlatishingiz mumkin.

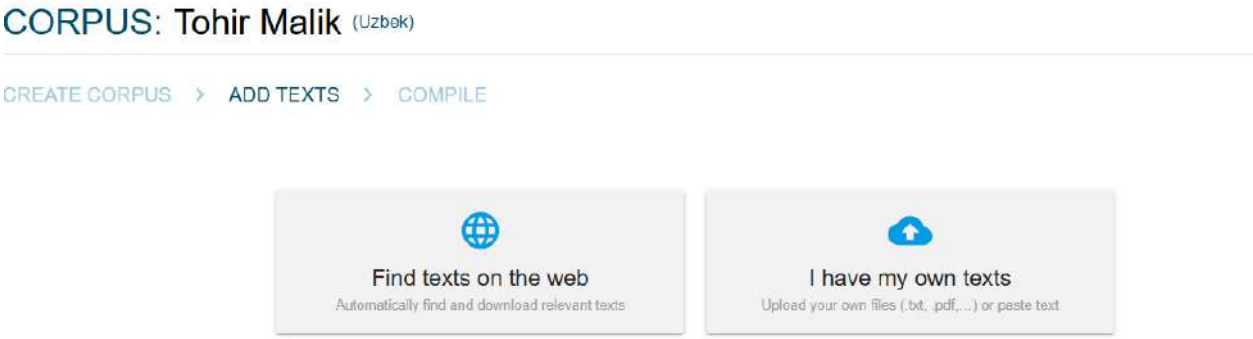


2-rasm (Ro‘yxatdan o‘tish)

3-rasm (Korpusni nomlash)

Undan so‘ng o‘z mualliflik korpusingizni yaratishni boshlashingiz mumkin. Buning uchun **New corpus** bo‘limidan, korpus nomi va tilini kiritib, korpusning dastlabki holati yaratib olinadi (3-rasm).

Bu bosqichdan so‘ng korpusga ma’lumotlarimizni kiritishni boshlaysiz. Bizning qo‘limizda mavjud bo‘lgan matnlardan foydalanib, **I have my own texts** bo‘limiga, txt fayllar yuklanadi (4-rasm).



4-rasm (txt fayllarni korpusga yuklash)

Sketch Enginening 1 oylik tekin versiyada faqat 1 mln so‘zdan iborat korpus yaratish mumkin. Fayllar yuklangach **Corpus content** qismida matnlarning necha so‘zdan iborat ekanligini ko‘rsatadi. Joylashtirilgan Tohir Malik asarlari **826055 so‘zdan** iborat bo‘ldi.

Tohir Malik asarlarida salbiy so‘zlar statistikasini aniqlash uchun birinchi navbatda Sketch Engine platformasidan

bo‘limiga o‘tiladi.

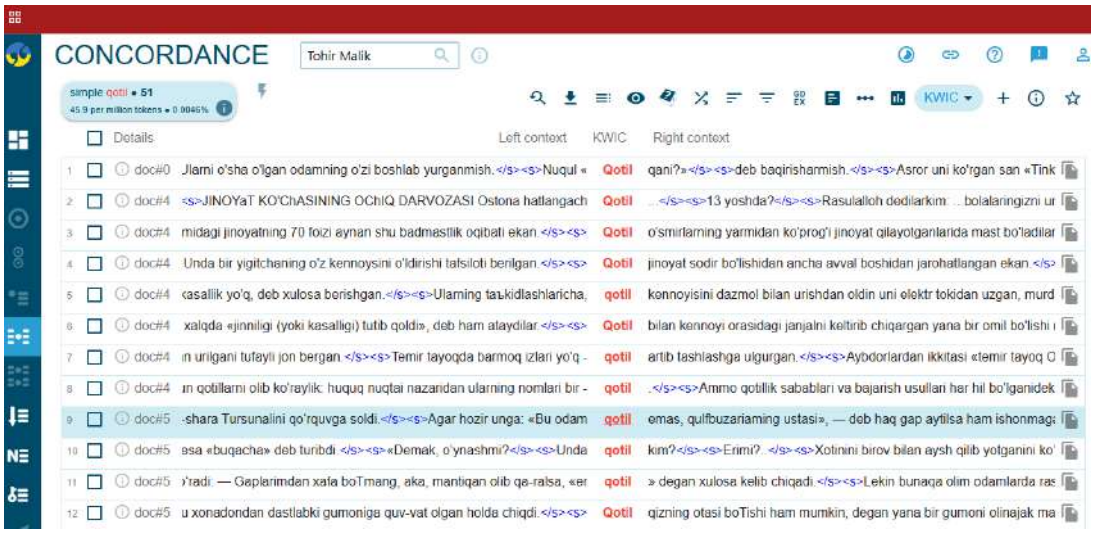
U bo‘limda 3 xil ko‘rinishda qidiruv berish mumkin. **Starting with** (bilan boshlanib), **ending with** (bilan tugaydigan), **containing** (o‘z ichiga olgan) qismlari bor (5-rasm), ushbu holatda **starting with** qismi bilan qidiruv berildi.

5-rasm (Qidiruv berish)

Bu yerga salbiy so‘zlar “yomon”, “o‘lim”, “qotil”, “xunuk” so‘zlarini qidiruvga berildi (6-rasm) va ushbu jadval hosil bo‘ladi (1-jadvalga qarang).

1-jadval. Qidiruv natijasi

So‘z	Soni	Qidiruv shakli
Yomon	356 marta	starting with
Azob	144 marta	starting with
Xunuk	74 marta	starting with
Qotil	51 marta	starting with
O‘lim	44 marta	starting with



6-rasm (Qidiruv natijasining korpusdagi ko‘rinishi)

Yanada yuqori aniqlikka erishish uchun ko‘proq salbiy bo‘lgan so‘zlarni qidiruvga berish kerak bo‘ladi. Ushbu so‘zlarning so‘z shakllarini ham ko‘rib chiqishim mumkin, buning uchun “qotil” so‘zining so‘z shakllarini olindi (2-jadvalga qarang).

2-jadval. “Qotil” so‘zining so‘z shakllari

№	So‘z	Soni	№	So‘z	Soni	№	So‘z	Soni
1	qotil	51	13	qotilsiz	3	25	qotilligini	1
2	qotillik	45	14	qotillarning	3	26	qotilim	1
3	qotilni	20	15	qotilga	3	27	qotillikni	1
4	qotillar	19	16	qotillikning	2	28	qotillikdan-chi	1
5	qotillikdan	11	17	qotillik-u	2	29	qotillarga	1
6	qotillarni	10	18	qotilligi	2	30	qotillardan	1
7	qotillikni	9	19	qotillarmi	2	31	qotilliklar	1
8	qotili	8	20	qotillari	2	32	qotilliklarning	1
9	qotillikka	7	21	qotilimni	2	33	qotillaridan	1
10	qotilini	7	22	qotilnimi	1	34	qotilingiz	1
11	qotillikda	6	23	qotillarigacha	1			
12	qotilning	6	24	qotillik-chi	1			

Xulosa. Mazkur tadqiqotda Tohir Malik asarlaridagi salbiy so‘zlarning statistik tahlili amalga oshirildi va ularning matn tuzilishidagi o‘rni, ma'no yuklanishi hamda til uslubiga ta'siri chuqur tahlil qilindi. Sketch Engine platformasida yaratilgan mualliflik korpusi yordamida, salbiy so‘zlarning chastotasi, birikish imkoniyatlari va turli kontekstlardagi ma'nodorligi o‘rganildi.

Tadqiqot natijalari shuni ko‘rsatdiki, yozuvchi salbiy leksik birliklarni o‘ziga xos uslubda qo‘llash orqali o‘quvchida kuchli hissiy rezonans uyg‘otadi va asarlarining mavzuviy chuqurligini oshiradi. Tohir Malikning ijodi inson psixologiyasi va jamiyat muammolarini aks ettirishda o‘ziga xos bo‘lib, salbiy so‘zlardan foydalanish asarlarining umumiy falsafiy yo‘nalishini belgilab beradi.

Ushbu ish nafaqat lingvistik korpuslarni yaratish sohasidagi yangi yondashuvlarni rivojlantiradi, balki o‘zbek adabiyoti va tilshunosligida muhim ilmiy manba sifatida xizmat qiladi. Kelgusida bu kabi tadqiqotlar o‘zbek yozuvchilari ijodiy uslubini aniqlash va ularning asarlarini tahlil qilish imkoniyatlarini kengaytirishga yordam beradi.

Foydalanilgan adabiyotlar

1. opensourceshakespeare.org
2. https://compling.navoiy-uni.uz/index.php/conferences/article/view/1038/1033

3. convertio.co/ru/pdf-txt/
4. https://parsing.uz/
5. https://app.sketchengine.eu/#wordlist?corpname=user%2Fmohimibra0204%2Ftohir_malik&tab=basic&keyword=qotil&filter=startingWith&wlattr=lc&wlminfreq=0&include_nonwords=1&itemsPerPage=50&showresults=1&cols=%5B%22frq%22%5D&showtimelines=0&diaattr=&showtimelineabs=0&timelinesthreshold=5
6. https://www.researchgate.net/profile/Manzura-Abjalova/publication/364348858_KORPUS_LINGVISTIKASI
7. https://tsuull.uz/sites/default/files/dissertatsiya_gulomova_compressed.pdf

ABDULLA QODIRIY ASARLARIDA QO‘LLANGAN TARIXIY SO‘ZLARNING KORPUS TAHLILI

Asal Aytbayeva Xudoybergen qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
asalaytboyeva027@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur maqola o‘zbek adabiyotining asoschilaridan biri Abdulla Qodiriyning yozgan asarlari va unda qo‘llangan tarixiy so‘zlar mualliflik korpusi tizimida o‘rganiladi. Unda yozuvchining asarlari matni *Sketch Engine* platformasida korpusga yuklanib tarixiy qo‘langan so‘zlar chastotasi, necha marta qo‘langanligi va hozirgi kunda tarixiy so‘zlarning muqobil varianti tahlil qilinadi. Mualliflik korpusida mashhur asarlari – “O‘tgan kunlar”, “Mehrobdan chayon”, “Uloqda”, “Baxtsiz kuyov”, “Fikraylagil”, “Ahvolimiz”, “Jinlar bazmi”, “To‘y” , “Millatimga”, “Fikr aylagil” asarlari tahlil qilinadi. Ushbu maqola Abdulla Qodiriy shaxsini va uning adabiyotimizdagi beqiyos ro‘lini o‘rganish uchun muhim manba hisoblanadi.

Kalit so‘zlar: *korpus, mualliflik korpusi, tarixiy so‘zlar, muqobil variant, tadqiqot metadologiyasi.*

KIRISH

Abdulla Qodiriy – o‘zbek adabiyotining zabardast namoyandalaridan biri bo‘lib, uning ijodiy merosi nafaqat badiiy qiymatga ega, balki xalqning o‘tmishini anglash va o‘zlikni qadrlash uchun ham muhim ahamiyat kasb etadi. Qodiriy o‘z asarlarida o‘zbek xalqining tarixini, madaniyati va urf-odatlarini chuqur tasvirlash bilan birga, tarixiy so‘zlar va iboralardan mahorat bilan foydalangan. Bu uslub o‘quvchini o‘sha davr muhitiga olib kirib, milliy ruhni his qilishga yordam beradi. Qodiriy ijodida ishlatilgan tarixiy leksika, o‘tmishda yashagan avlodlarimizning hayot tarzi, qadriyatlari va madaniyatidan dalolat beradi. Ushbu tarixiy lug‘at zamonaviy o‘zbek tilining shakllanish jarayonini o‘rganish uchun ham bebaho manba hisoblanadi. Ammo zamonaviy o‘zbek tilida bu so‘zlarning bir qismi unutilgan, ayrimlari esa boshqa shakl yoki ma’no kasb etgan. Shuning uchun bunday tarixiy lug‘atni o‘rganish nafaqat lingvistik, balki madaniy ahamiyatga ham ega. Ushbu maqola Abdulla Qodiriy asarlarida qo‘llangan tarixiy so‘zlar va iboralar bilan zamonaviy variantlar o‘rtasidagi o‘zaro bog‘liqlikni tadqiq qilishga qaratilgan. Tadqiqot davomida tarixiy lug‘atning tilshunoslikka oid jihatlari va madaniy meros sifatidagi ahamiyati yoritiladi. Bu orqali Qodiriy asarlarida ko‘rinish bergan tarixiy til qatlamining bugungi kundagi o‘rnini tushunish va qadrlash imkoniyati yaratiladi.

Hozirgi kunda Abdulla Qodiriy asarlari bo‘yicha maxsus mualliflik korpusi mavjud emasligi haqida ma’lumotlar bor. Lekin bu sohada ayrim tadqiqotchilar tomonidan qilingan ishlar va elektron shaklda chop etilgan materiallar mavjud. Masalan: **H.Akbarovning “Abdulla Qodiriy va sanoyi nafisa” kitobi**: Ushbu asarda Abdulla Qodiriyning ijodi va uning san’atga bo‘lgan munosabati haqida batafsil ma’lumot berilgan[2].

N.Karimovning “Abdulla Qodiriy qomusi” kitobi: maqolasida adibning yangi o‘zbek adabiy tilining shakllanishidagi xizmatlarini yoritgan[3].

M.Abdiyevning “Abdulla Qodiriyning o‘zbek adabiy tili va matbuoti rivojidagi xizmatlari” maqolasida yozuvchining matbuot sohasidagi faoliyatini tahlil qilgan[4].

N.Normatovning “Abdulla Qodiriy hajviyoti” maqolasida adibning hajviy asarlari va publitsistik mahoratini o‘rgangan[5].

Maqolaning mualliflik korpusi haqida. Mualliflik korpusi Qodiriy ijodidagi tarixiy so‘zlarning to‘liq ro‘yxatini tuzish, ularning hozirgi kundagi muqobillarini aniqlash va semantik tahlil qilishga qaratilgan. Ushbu korpus quyidagi vazifalarni bajarishga xizmat qiladi:

Leksik tahlil: Asarlarda ishlatilgan so‘zlarni lug‘aviy ma‘nolari bilan aniqlash.

Hozirgi kundagi muqobillar: Tarixiy so‘zlarning zamonaviy tildagi ishlatilishini tahlil qilish.

Tadqiqot metodologiyasi. Tadqiqot davomida Abdulla Qodiriyning mualliflik korpusini tuzishga erishildi. Abdulla Qodiriyning ijodi va asarlarida qo‘llagan noyob so‘zlarni kengroq o‘rganish, tahlil qilish va ilmiy tadqiqotlar uchun foydalanishni maqsad qilib, ushbu mualliflik korpusi tuzildi. Bu korpus adibning asarlaridagi tarixiy va lingvistik o‘zgarishlarni o‘rganish uchun muhim manba bo‘lib, ularni tilshunoslik nuqtai nazaridan tahlil qilish imkonini beradi.

Korpusni yaratish jarayoni bir nechta bosqichlardan iborat bo‘ldi:

Mualliflik korpusini yaratish jarayonida Abdulla Qodiriy asarlari bo‘yicha bir nechta bosqichlarni amalga oshirish mumkin. *Mualliflik korpusi* — bu muayyan yozuvchi yoki adabiyotshunoslik obyekti asosida yaratilgan matnlar to‘plami bo‘lib, u tilshunoslik, adabiyotshunoslik, leksikografiya va boshqa sohalarda o‘rganish, tahlil qilish uchun ishlatiladi. Quyida Abdulla Qodiriy asarlari bo‘yicha mualliflik korpusini yaratish jarayonidagi asosiy bosqichlar keltirilgan:

Asarlarni tanlash: Korpus yaratish uchun Abdulla Qodiriy asarlarini tanlash kerak. Bu bosqichda, asarlarning to‘liq va ishonchli nusxalari tanlandi.

***Matnlarni raqamlashtirish**:* Agar matnlar elektron shaklda bo‘lmasa, ular skaner qilinib, OCR texnologiyasi yordamida raqamlashtiriladi.

***Korpusni tozalash**:* Raqamlashtirilgan matnlardan xatolar va keraksiz ma‘lumotlar olib tashlandi.

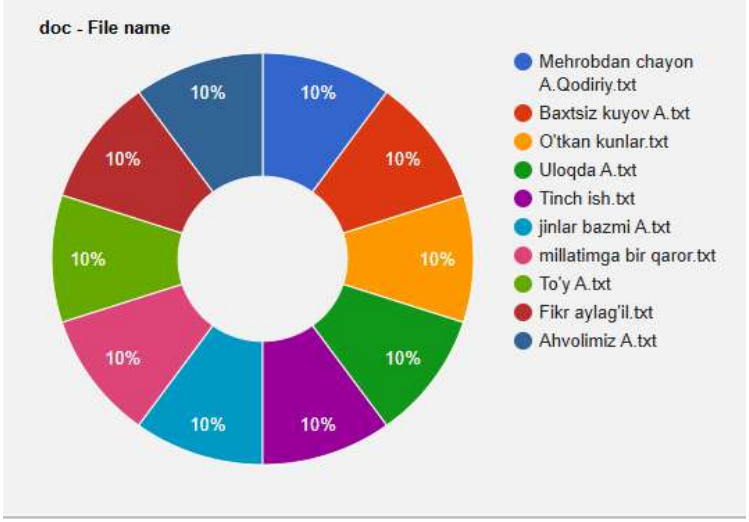
***Elektron platformaga joylash**:* Tayyor matnlar Sketch Engine https://app.sketchengine.eu/ platformasiga yuklanib, tahlil uchun tayyorlandi.

Ushbu amallar yakunida to‘plangan asarlar haqidagi ma’lumotlarni quyidagi jadvalda ko‘rishingiz mumkin. (Qarang: *1-jadval*).

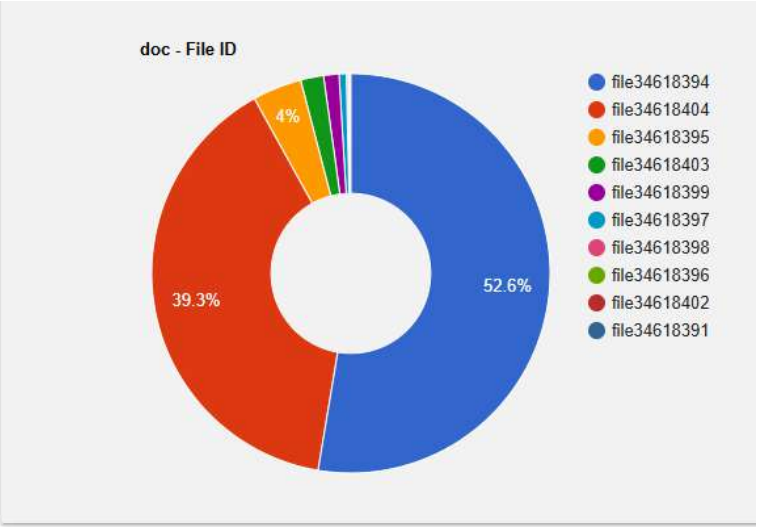
1-jadval. (Korpusdagi barcha asarlarning jamlanmasi)

No	Asar nomi	Muallif	Yashagan yillari	Yozilgan yili	Adabiy tur	Janri	Uslubi	Nashriyoti	Nashr yili
1	O'tkan kunlar	Abdulla Qodiriy	1894-1938	1920	epik	roman	badiiy	G‘afur G‘ulom nomidagi nashriyot-matbaa ijodiy uyi	1994
2	Mehrobdan chayon	Abdulla Qodiriy	1894-1938	1928	epik	roman	badiiy	G‘afur G‘ulom nomidagi nashriyot-matbaa ijodiy uyi	1994
3	Uloqda	Abdulla Qodiriy	1894-1938	1923	epik	hikoya	badiiy	-	-
4	Baxtsiz kuyov	Abdulla Qodiriy	1894-1938	1920	epik	dramma	badiiy	-	-
5	Jinlar bazmi	Abdulla Qodiriy	1894-1938	1916	epik	hikoya	badiiy	“Sharq chechagi” jurnalida bosilgan.	1921
6	Tinch ish	Abdulla Qodiriy	1894-1938	1920	epik	hikoya	badiiy	«Sharq» NMAK Bosh tahririyati	2005
7	To'y	Abdulla Qodiriy	1894-1938	1912	Lirik	she'r	badiiy	«Sharq» NMAK Bosh tahririyati	2005
8	Ahvolimiz	Abdulla Qodiriy	1894-1938	1912	Lirik	she'r	badiiy	«Sharq» NMAK Bosh tahririyati	2005
9	Millatimga	Abdulla Qodiriy	1894-1938	1912	Lirik	she'r	badiiy	-	-
10	Fikr aylag'il	Abdulla Qodiriy	1894-1938	1912	Lirik	she'r	badiiy	-	-

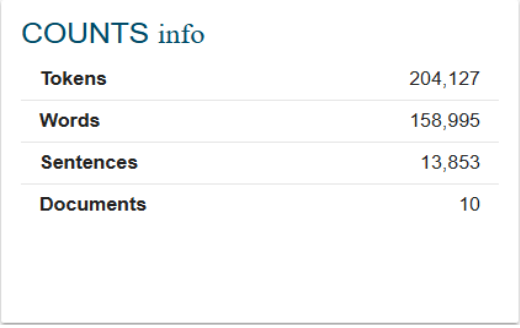
Matnlarni yuklash: Yig‘ilgan va tayyorlangan matnlarni *Sketch engine* tizimiga yuklandi. Bu jarayondagi yuklangan fayllar natijasida quyidagi diagramma hosil bo‘ldi. (Qarang: 1-rasm).



1-rasm. (Korpusga kiritilgan asarlar)



2-rasm (Qodiriy asarlari foizlarda)



COUNTS info

Tokens	204,127
Words	158,995
Sentences	13,853
Documents	10

3-rasm. (Korpusdagi barcha asarlar haqida)

Bu diagrammada yuklangan fayllarning taqsimoti ko‘rsatilgan (Qarang: 2-rasm). Keyingi diagrammada esa har bitta strukturlangan matnli ma’lumotlarimiz uchun tokenlar statistikasini ko‘rsatadi (Qarang: 3-rasm).

Bu korpusdagi umumiy tokenlar, gaplar, so‘zlar taqsimoti va nechta asar borligini ko‘rishingiz mumkin.

Tahlil va natijalar. Abdulla Qodiriy asarlarida so‘zlar va iboralarni kontekstual tahlil qilish uchun “Concordance” vositasidan foydalanib yozuvchi asarlarida qaysi so‘zlarni ko‘p ishlatilganini bilib olishimiz mumkin. (Qarang: *2-jadval*).

2-jadval. (Korpusdagi eng ko‘p ishlatilgan so‘zlar)

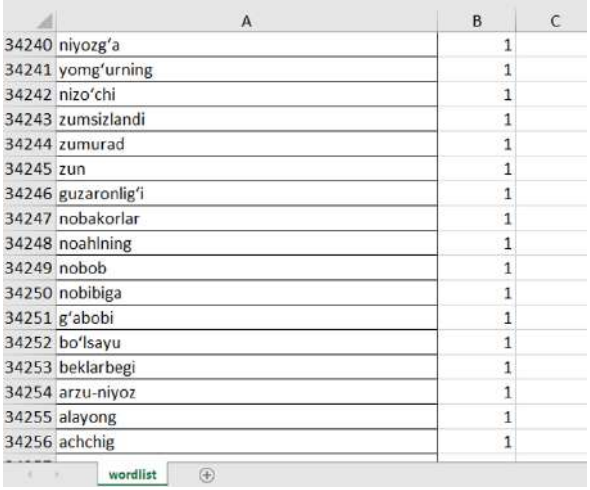
№	So‘z	Takrorlanish soni
1	Shu	797
2	Men	740
3	Anvar	624
4	Otabek	611
5	Yana	599

6	Uning	575
7	Bo‘lsa	497
8	Nima	472
9	Oyim	451
10	Qilib	408

Mualliflik korpusiga yuklangan barcha asarlar Excel platformasiga yuklanadi (Qarang: 4-rasm).

4-rasm (Ko‘p ishlatilgan so‘zlarni yuklash)

Excel orqali korpusdagi 34.456 ta so‘zlarni necha marta qatnashganligi yuklab olindi, bu yanada tarixiy so‘zlarni topishimizga yordam beradi(Qarang: 5-rasm).



	A	B	C
34240	niyozg‘a	1	
34241	yomg‘urning	1	
34242	nizo‘chi	1	
34243	zumsizlandi	1	
34244	zumurad	1	
34245	zun	1	
34246	guzaronlig‘i	1	
34247	nobakorlar	1	
34248	noahlning	1	
34249	nobob	1	
34250	nobibiga	1	
34251	g‘abobi	1	
34252	bo‘lsayu	1	
34253	beklarbegi	1	
34254	arzu-niyoz	1	
34255	alayong	1	
34256	achchig	1	

wordlist

5-rasm (Excelda yuklangan so‘zlar)

34,456 so‘zning ichidan 10 ta ajratib olindi va Abdulla Qodiriy asarlarida ishlatilgan tilning boyligi hamda noyob so‘zlarning takrorlanmasligi yana bir bor tasdiqlandi(Qarang: *3-jadval*).

3-jadval(Qodiriy asarlarida ishlatgan tarixiy so‘zlar)

№	Tarixiy so‘z	Mazmuni	Hozirgi kundagi muqobil varianti	Asarlar
1	Xon	Markaziy Osiyodagi hukmron yoki yuqori darajadagi rahbarni anglatadi.	Hukmdor, boshqaruvchi	O‘tkan kunlar
2	Madrasa	Islomiy ta’lim beradigan maktab yoki diniy o‘quv muassasasi.	Diniy maktab, o‘quv yurtlari	O‘tkan kunlar
3	Bek	Xonliklar davrida davlat arboblariga, shahar yoki viloyat hokimlariga berilgan faxriy unvon.	Rahbar, boshliq, lider	O‘tkan kunlar, Mehrobdan chayon
4	Mulla	Diniy ta’lim olgan shaxs.	Imom, domla, shayx, ustoz	O‘tkan kunlar, Mehrobdan chayon
5	Qozil	Qasrda yoki saroyda ishlaydigan xizmatkorlar.	Xizmatkorlar, uy tozalovchilar	Jinlar bazmi
6	Ulamo	Islom ilm-fanini o‘rgangan va diniy masalalarda yetakchi o‘qituvchilar.	Din olimlari, diniy mutafakkirlar	O‘tkan kunlar
7	Taqsir	Hurmat bilan murojaat (olim yoki din xodimiga nisbatan).	Janob	O‘tkan kunlar, Mehrobdan chayon
8	Qul	Erkin bo‘lmagan, sohibi tomonidan xizmatga yoki ishlashga majbur qilingan shaxs.	Mehnatkash, ishchi	O‘tkan kunlar
9	Qasr	Mukammal qurilgan saroy yoki hukmronning yashash joyi.	Saroy, hashamatli bino, villa	Tinch ish
10	Qozikalon	Oliy darajadagi qozilik lavozimi.	Sud raisi yoki sudya	O‘tkan kunlar

Eng ko‘p qo‘llanilgan gaplar tahlili. Abdulla Qodiriy mualliflik korpusi yordamida yozuvchining tarixiy so‘zlar bilan ishlash uslubi qisman o‘rganildi. Adib asarlarida shu so‘zlarning ko‘p marotaba ishlatilgani aniqlandi(Qarang: *4-jadval*) va asarda necha marotaba qo‘llanilgani aniqlandi(Qarang: *5-jadval*).

4-jadval(Qodiriy asarlaridagi ko‘p uchragan tarixiy so‘zlar)

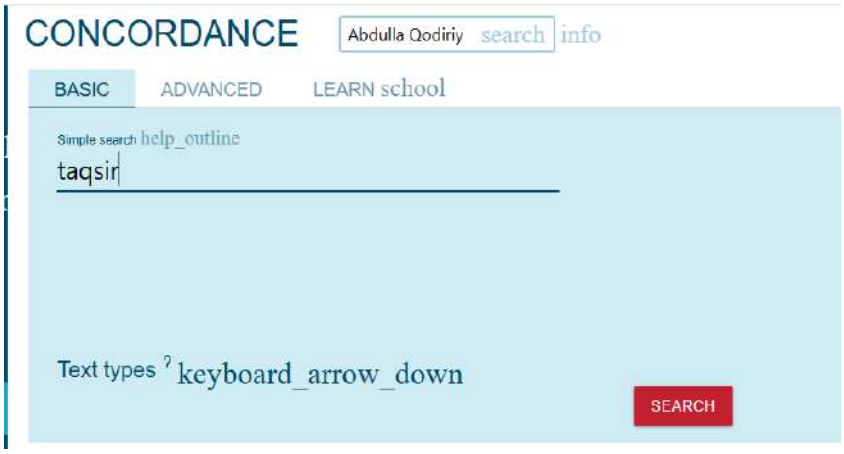
№	Tarixiy so‘zlar:	Natija
1.	“Taqsir”	Hurmat bilan murojaat (olim yoki din xodimiga nisbatan).
2.	“Mulla”	Diniy ta’lim olgan shaxs.
3.	“Xon”	Markaziy Osiyodagi hukmron yoki boshqaruvchi.

4.	“Bek”	Turkiy xalqlarning ba’zilarida xonliklar davrida davlat arboblariga, shahar yoki viloyat hokimlariga, ularning bolalariga berilgan faxriy unvon hamda shunday unvonga sazovor bo‘lgan shaxs.

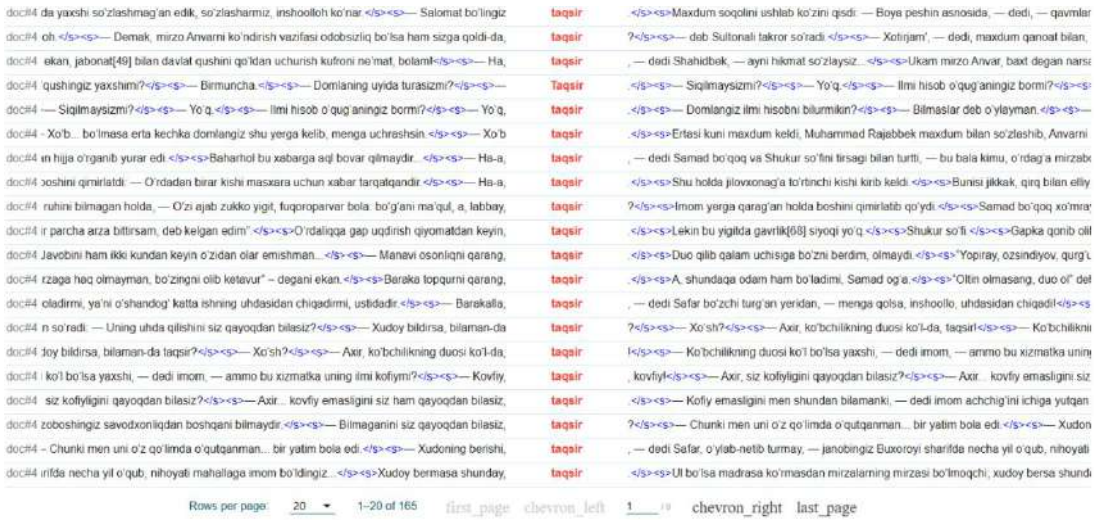
5-jadval(Qodiriy asarlarida qo‘llagan so‘zlar miqdori)

№	Korpusda ishlatilgan eng ko‘p tarixiy so‘zlar:	Natija
1.	“Taqsir”	165 ta gapda qo‘llanilgan
2.	“Mulla”	159 ta gapda qo‘llanilgan
3.	**“Xon”**	134 ta gapda qo‘llanilgan
4.	**“Bek”**	119 ta gapda qo‘llanilgan

CONCORDANCE bo‘limida eng ko‘p qo‘llanilgan tarixiy so‘zlardan biri “taqsir” so‘zi qidirib ko‘rildi (Qarang: 5-rasm) va quyidagi natijalar olindi (Qarang: 6-rasm).



5-rasm (Eng ko‘p ishlatilgan so‘z)



doc#4 da yaxshi so'zlashmag'an edik, so'zlasharmiz, inshoolloh ko'nar.</s><s>— Salomat bo'lingiz	taqsir	</s><s>Maxdum soqolini ushlab ko'zini qisdi: — Boya peshin asnosida, — dedi, — qavmlar
doc#4 oh.</s><s>— Demak, mirzo Anvarni ko'ndirish vazifasi odobsizliq bo'lsa ham sizga qoldi-da,	taqsir	?</s><s>— deb Sultonali takror so'radi.</s><s>— Xotirjam', — dedi, maxdum qanoat bilan,
doc#4 ekan, jabonat[49] bilan davlat qushini qo'ldan uchurish kufroni ne'mat, bolam!</s><s>— Ha,	taqsir	, — dedi Shahidbek, — ayni hikmat so'zlaysiz...</s><s>Ukam mirzo Anvar, baxt degan narsa
doc#4 qushingiz yaxshimi?</s><s>— Birmuncha.</s><s>— Domlaning uyida turasizmi?</s><s>—	Taqsir	</s><s>— Siqilmaysizmi?</s><s>— Yo'q.</s><s>— Ilmi hisob o'qug'aningiz bormi?</s><s
doc#4 — Siqilmaysizmi?</s><s>— Yo'q.</s><s>— Ilmi hisob o'qug'aningiz bormi?</s><s>— Yo'q,	taqsir	.</s><s>— Domlangiz ilmi hisobni bilurmikin?</s><s>— Bilmaslar deb o'ylayman.</s><s>—
doc#4 - Xo'b... bo'lmasa erta kechka domlangiz shu yerga kelib, menga uchrashsin.</s><s>— Xo'b	taqsir	</s><s>Ertasi kuni maxdum keldi, Muhammad Rajabbek maxdum bilan so'zlashib, Anvarni
doc#4 in hija o'rganib yurar edi.</s><s>Baharhol bu xabarga aql bovar qilmaydir...</s><s>— Ha-a,	taqsir	, — dedi Samad bo'qoq va Shukur so'fini tirsagi bilan turtti, — bu bala kimu, o'rdag'a mirzab
doc#4 oshini qimirlatdi: — O'rdadan birar kishi maxxara uchun xabar tarqatqandir.</s><s>— Ha-a,	taqsir	</s><s>Shu holda jilovxonag'a to'rtinchi kishi kirib keldi.</s><s>Bunisi jikkak, qirq bilan elliy
doc#4 ruhini bilmagan holda, — O'zi ajab zukko yigit, fuqoroparvar bola: bo'g'ani ma'qul, a, labbay,	taqsir	?</s><s>Imom yerga qarag'an holda boshini qimirlatib qo'ydi.</s><s>Samad bo'qoq xo'mra
doc#4 ir parcha arza bittirsam, deb kelgan edim".</s><s>O'rdaliqqa gap uqdirish qiyomatdan keyin,	taqsir	</s><s>Lekin bu yigitda gavrlik[68] siyoqi yo'q.</s><s>Shukur so'fi.</s><s>Gapka qonib olil
doc#4 Javobini ham ikki kundan keyin o'zidan olar emishman...</s><s>— Manavi osonliqni qarang,	taqsir	.</s><s>Duo qilib qalam uchisiga bo'zni berdim, olmaydi.</s><s>"Yopiray, ozsindiyov, qurg'u
doc#4 rzaga haq olmayman, bo'zingni olib ketavur" – degani ekan.</s><s>Baraka topqurni qarang,	taqsir	.</s><s>A, shundaqa odam ham bo'ladimi, Samad og'a.</s><s>"Oltin olmasang, duo ol" del
doc#4 oladirmi, ya'ni o'shandog' katta ishning uhdasidan chiqadirmi, ustidadir.</s><s>— Barakalla,	taqsir	, — dedi Safar bo'zchi turg'an yeridan, — menga qolsa, inshoollo, uhdasidan chiqadil</s><s
doc#4 n so'radi: — Uning uhda qilishini siz qayoqdan bilasiz?</s><s>— Xudoy bildirsa, bilaman-da	taqsir	?</s><s>— Xo'sh?</s><s>— Axir, ko'bchilikning duosi ko'l-da, taqsir!</s><s>— Ko'bchilikni
doc#4 doy bildirsa, bilaman-da taqsir?</s><s>— Xo'sh?</s><s>— Axir, ko'bchilikning duosi ko'l-da,	taqsir	!</s><s>— Ko'bchilikning duosi ko'l bo'lsa yaxshi, — dedi imom, — ammo bu xizmatka uning
doc#4 i ko'l bo'lsa yaxshi, — dedi imom, — ammo bu xizmatka uning ilmi kofiymi?</s><s>— Kovfiy,	taqsir	, kovfiy!</s><s>— Axir, siz kofiyligini qayoqdan bilasiz?</s><s>— Axir... kovfiy emasligini siz
doc#4 siz kofiyligini qayoqdan bilasiz?</s><s>— Axir... kovfiy emasligini siz ham qayoqdan bilasiz,	taqsir	</s><s>— Kofiy emasligini men shundan bilamanki, — dedi imom achchig'ini ichiga yutgan
doc#4 zoboshingiz savodxonliqdan boshqani bilmaydir.</s><s>— Bilmaganini siz qayoqdan bilasiz,	taqsir	?</s><s>— Chunki men uni o'z qo'limda o'qutqanman... bir yatim bola edi.</s><s>— Xudon
doc#4 – Chunki men uni o'z qo'limda o'qutqanman... bir yatim bola edi.</s><s>— Xudoning berishi,	taqsir	, — dedi Safar, o'ylab-netib turmay, — janobingiz Buxoroyi sharifda necha yil o'qub, nihoyati
doc#4 irifda necha yil o'qub, nihoyati mahallaga imom bo'ldingiz...</s><s>Xudoy bermasa shunday,	taqsir	</s><s>Ul bo'lsa madrasa ko'rmasdan mirzalarning mirzasi bo'lmoqchi, xudoy bersa shund

Rows per page: 20 1–20 of 165 first_page chevron_left 1 chevron_right last_page

6-rasm (Taqsir so‘zi qatnashgan gaplar)

Abdulla Qodiriy asarlarida chastotasi 1 ga teng bo‘lgan tarixiy so‘zlar(Qarang: 6-rasm).

6-jadval(Qodiriy asarlarida qo‘llagan so‘zlar miqdori)

So‘z	Ma’nosi
Moziy	Arab tilidan kirib kelgan bo‘lib, *“o‘tgan zamon”*, *“o‘tgan davr”* yoki *“o‘tgan kunlar”* degan ma’noni bildiradi. U odatda tarixiy voqealar, avvalgi hayot tarzlari yoki shaxsning hayot yo‘lidagi bosib o‘tilgan davrni ifodalashda qo‘llaniladi.
Sabzavor	Bu so‘z sabzavotlar yoki yashil o‘simliklar bilan bezatilgan joylarni anglatadi. O‘zbek adabiy tilida *“sabzavor joy”*, *“yashil landshaft”* kabi birikmalar tarkibida ishlatiladi. Estetik go‘zallik va tabiiy muhit tasvirida qo‘llanadi.
Maslak	Lug‘aviy ma’nosi – e’tiqod, yo‘nalish, fikrlar tizimi. Ko‘pincha siyosiy, ijtimoiy yoki dini qarashlar tizimini ifodalash uchun ishlatiladi. Matndagi “Markaziy Osiyodagi hukmron yoki boshqaruvchi” ma’nosi xato: bu *“xon”* yoki *“bek”* so‘zlariga tegishli.
Ziynat	Bu so‘z aslida *“bezak”*, *“chiroy”*, *“go‘zallikni oshiruvchi narsa”* degan ma’noni bildiradi. U kuyosh, kiyim, ayollar taqinchoqlari yoki me’moriy elementlar haqida so‘z yuritilganda qo‘llaniladi. Jadvalda berilgan izoh esa ***“bek”* yoki *“mulozim”*** so‘zlariga mos tushadi.

XULOSA

Abdulla Qodiriy asarlarida ishlatilgan eng ko‘p tarixiy so‘zlar “Taqsir” (169), “mulla” (159), “xon” (134), “bek” (119) kabi so‘zlar muallifning o‘sha davrdagi muhitini chiroyli qilib tasvirlab berishga erishilgan. Hozirgi kunda bu tarixiy so‘zlarning ko‘plari kundalik tilimizda kamdan-kam uchraydi yoki ular o‘rniga zamonaviy muqobil so‘zlar qo‘llanilmoqda.

Qodiriy asarlari tahlil natijalarida chastotasi 1 ga teng bo‘lgan so‘zlar son jihatdan sezilarli darajada kam uchragani aniqlandi. “Moziy”, “Sabzavor”,

“maslak”, “ziynat” kabi so‘zlar ijodini o‘rganish bu so‘zlarning asl ma’nosini va tilimizdagi qadriyatlarini anglashga yordam beradi.

Tarixiy so‘zlarni o‘rganish va ularning muqobillarini aniqlash shuni ko‘rsatdiki, Qodiriy asarlarida ishlatilgan tarixiy so‘zlar ajoyib qo‘llanganligi va o‘sha davrni chuqur yetkazib berishga erishilgan. Korpus asosida o‘tkazilgan bu tahlil qilish faqat filologik jihatdan emas, balki madaniy, tarixiy va ma’naviy nuqtayi nazardan ham alohida ahamiyat kasb etadi. Ushbu mavzuni chuqur o‘rganish milliy merosimizga nisbatan hurmatni oshirish va uni asrab-avaylashda katta ahamiyatga ega.

Foydalanilgan adabiyotlar:

1. K.Bahodir. ,Abdulla Qodiriy – T 2022, “Yoshlar nashriyot uyi”
2. https://kitobxon.com/oz/kitob/abdulla-qodiriy-va-sanoyi-nafisa
3. https://www.researchgate.net/publication/342337550_abdulla_qodiriyning_o'zbek_adabiy_tili_va_matbuoti_rivojidagi_xizmatlari
4. https://www.researchgate.net/publication/342337550_abdulla_qodiriyning_o'zbek_adabiy_tili_va_matbuoti_rivojidagi_xizmatlari
5. https://hilolnashr.uz/abdulla-qodiriy-asarlari-jamlanmasi

KORPUS VOSITASIDA HISSIY TAHLIL
(Oybekning “Bolalik” qissasi misolida)

Xushmuratova Madina Baxtiyor qizi,
Kompyuter lingvistikasi talabasi
xushmurodovamadina209@gmail.com
ToshDO‘TAU

Annotatsiya: Ushbu maqola Oybekning “Bolalik” qissasi korpusi asosida hissiy tahlilni aniqlashga bag‘ishlangan. Maqolada qissadagi ijobiy va salbiy ma’nodagi so‘zlar aniqlanib, ularning asardagi hissiy muhitni shakllantirishdagi roli o‘rganilgan. Korpus tahlili orqali Oybek qissada qaysi turdagi hissiy so‘zlarni ko‘proq qo‘llaganini aniqlash uchun maxsus qidiruvlar amalga oshirilgan bo‘lib, muallif ijobiy hissiyotlarni ifodalovchi leksik birliklarga nisbatan salbiy ma’nodagi so‘zlarni ko‘proq qo‘llash orqali bolalik davridagi ichki murakkabliklarni va hayotiy sinovlarni aks ettirishga harakat qilgani o‘rganildi. Shu bilan birga maqolada hissiy so‘zlarning Oybek uslubini boyitishi va qissadagi umumiy mazmunni ochib berishdagi ahamiyati ham tahlil qilingan.

Kalit so‘zlar: *Oybek, statistik tahlil, mualliflik korpusi, leksik birlik, ijobiy, salbiy, qissa, badiiy uslub, chastota.*

Kirish

Hissiy so‘zlar matnning kayfiyatini shakllantiradi, o‘quvchi yoki tinglovchida muallifning hissiy holatini tushunish uchun zarur bo‘lgan hissiy aloqani yuzaga keltiradi. Zamonaviy dunyoda hissiy leksikani o‘rganishning ahamiyati ortib bormoqda. Hissiy leksika nafaqat adabiy asarlarni tahlil qilishda, balki ijtimoiy tarmoqlar, marketing, siyosat, psixologiya va boshqa ko‘plab sohalarda ham keng qo‘llaniladi. Sentiment analysis (hissiy tahlil) usullarining rivojlanishi bilan matnlar mazmunidagi ijobiy, salbiy va neytral hissiyotlarni avtomatik aniqlash imkoniyati yuzaga kelgan. Bu esa korxonalar uchun mijozlarning fikrini o‘rganish, ommaviy axborot vositalari uchun xabarlarning ta’sirchanligini baholash, shuningdek, siyosiy sohadagi jamoatchilik fikrini tahlil qilish uchun katta ahamiyatga ega.

Adabiy asarlarni o‘rganishda hissiy leksika muhim vosita sifatida namoyon bo‘ladi. Mualliflarning hissiyotlarga boy til vositalari orqali qanday g‘oyalarni yetkazmoqchi bo‘lgani, o‘quvchiga qanday ta’sir ko‘rsatishni maqsad qilgani aniqlanadi. Bu jarayon asarlar mazmunini chuqurroq tushunishga yordam beradi. Shuningdek, milliy madaniyatni o‘rganishda va tilning hissiy qatlamini tadqiq qilishda hissiy leksika alohida o‘rin tutadi. Shu sababli, hissiy leksika nafaqat lingvistika sohasida, balki turli ijtimoiy va ilmiy yo‘nalishlarda ham muhim

ahamiyat kasb etadi. U til orqali inson va jamiyat o‘rtasida yanada chuqur aloqalarni shakllantirish imkonini beradi.

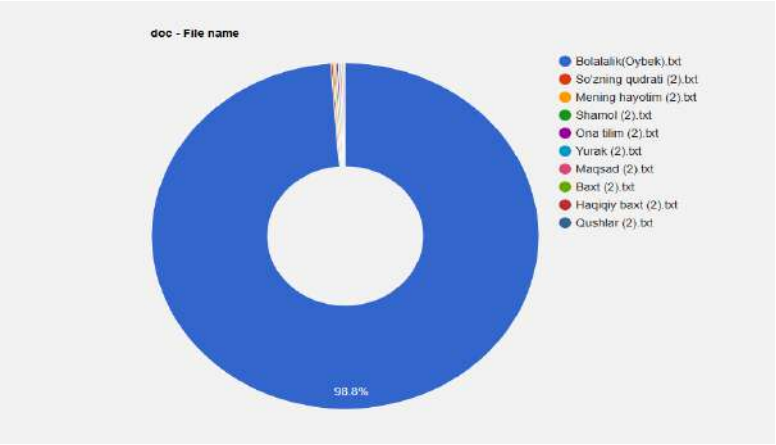
Amalga oshirilgan ilmiy tadqiqotlar

Matnlardagi hissiyotni avtomatik aniqlash bo‘yicha B.Pang o‘z tadqiqotlari bilan tanilgan[4]. Uning ishlari sentiment analysis sohasida asosiy manbalardan biri hisoblanadi. L.Lee tomonidan matnlarni hissiy tahlil qilish va tabiiy tilni qayta ishlash bo‘yicha muhim tadqiqotlar olib borgan[5]. B.Pang bilan hamkorlikda sentiment analysis asosida bir nechta ilmiy ishlar yozgan. P.Turney Matnlarni ijobiy va salbiy deb tasniflash uchun “Thumbs Up or Thumbs Down?” nomli tadqiqoti bilan mashhur[6]. Uning ishlari sentiment analysis sohasida katta ahamiyatga ega. B.Liu hissiy tahlil va fikrlarni qazib olishda yetakchi mutaxassisdir[7]. Uning “Sentiment Analysis and Opinion Mining” nomli kitobi sohadagi muhim manbalardan biridir. O‘zbekistonda ham bu sohada tadqiqotlar olib borilmoqda. Shuningdek, Buxoro davlat universiteti dotsenti N.Ataboyevning korpus lingvistikasi va til leksik qatlamining boyishi bo‘yicha tadqiqotlari mavjud[8].

Tadqiqot metodologiyasi

Sketch Engine korpusi tahlil qilish va korpus yaratish uchun keng qo‘llaniladigan zamonaviy platformalardan biridir[9]. Bu platforma tilshunoslar, tarjimonlar, yozuvchilar va tadqiqotchilar uchun juda qulay vosita hisoblanadi. Sketch Enginega har qanday tilda yaratilgan korpusni joylash imkoni mavjud bo‘lib, so‘zlarning chastotasini hisoblash kabi.

Tadqiqot doirasida Oybekning qissa va she’rlari korpusga aylantirildi, va korpusning 98% hajmini “Bolalik” qissasi egallaganligi uchun bu tadqiqot aynan o‘sha qissa ustida amalga oshirildi. Mazkur tadqiqotda korpus lingvistikasi metodologiyasi asosida “Bolalik” qissasidagi hissiy leksik birliklar tahlil qilinadi. Korpus yaratishda “Skate engine” ga Oybekning asaridan olingan barcha matnlar yig‘ildi, so‘zlarning ma’nosi, shakli, va ularning kontekstdagi o‘rni statistik usulda o‘rganildi, matnlar krill grafikasidan lotin grafikasiga o‘tkazilib, barcha asarlar bir xil formatga keltirildi, matn ortiqcha shovqin va “stop word”lardan tozalandi, undan keyin saytga yuklandi.



1-rasm. Korpusga asarlarni joylash

Tahlil uchun leksik birliklarning chastotasi, ijobiy va salbiy so‘zlar o‘rtasidagi nisbatlar, ularning asaridagi hissiy holatlarni yaratishdagi o‘rni hisobga olinadi. Bu quyidagicha aniqlanadi:

1-qadam: Korpusdagi mavjud bo‘lgan ixtiyoriy birorta salbiy so‘z qidiruvga berildi. Bu bosqichda korpusga “xafa” so‘zi kiritildi. Bu — hissiy tahlilda muhim qadam, chunki dastlabki izlanishlar aynan qaysi hissiy rangdagi leksik birliklar matnda ko‘p ishlatilganini aniqlashdan boshlanadi. “Xafa” so‘zi salbiy emotsiyani bildiradi va bolalikda boshdan kechirilgan iztiroblar, ruhiy siqilish holatlarini anglanadi. Aynan shu qadam orqali salbiy leksik birliklarning mavjudligi isbotlab beriladi va keyingi tahlillar uchun asos yaratiladi. (Qarang:1-rasmga).

2-rasm. Korpusga salbiy so‘z kiritish

2-qadam: Qidiruvga berilgan so‘zning qo‘llanilish soni aniqlandi, va natija taqdim etildi. (Qarang: 2-rasmga).

3-rasm. Salbiy gap tahlili

Tahlil natijasida shuni aytish kerakki, “xafa” so‘zi 35 marta qo‘llangani aniqlangan. Bu raqam – statistik jihatdan sezilarli ko‘rsatkich bo‘lib, muallif salbiy tuyg‘ularni ifodalashda aynan shu so‘zni ko‘p ishlatganini bildiradi. Bu degani, Oybek o‘z bolalik xotiralarida ko‘proq iztirobli, ruhiy siqilishga boy holatlarni tasvirlashga urg‘u bergan.

3-qadam: Korpusdan mavjud ixtiyoriy ijobiy so‘zni qidiruvga berib ko‘rildi. Tahlilning muqarrar davomi sifatida bu qadamda “shodlik” kabi ijobiy hissiyot bildiruvchi so‘zlar korpusga kiritildi. (Qarang:3-rasmga).

4-rasm. Korpusga ijobiy so‘z kiritish

4-qadam: Qidirishga berilgan so‘zning qo‘llanilish soni aniqlanadi va natija taqdim etiladi. Bu metodologik jihatdan muhim bo‘lib, salbiy va ijobiy so‘zlar o‘rtasida nisbat tuzish imkonini beradi. “Shodlik” – bolalikdagi quvonch, orzu-umid, baxtli onlarni ifodalovchi so‘zdir. Bu so‘zning nechalik kam ishlatilgani Oybekning bolalik haqidagi esdaliklarida ijobiy holatlardan ko‘ra, salbiy holatlar ustuvor bo‘lganini ko‘rsatdi. (Qarang: 4-rasmga).

5-rasm. Ijobiy gap tahlili

Qidiruv natijasi – “shodlik” so‘zi bor-yo‘g‘i 2 marotaba ishlatilgan. Bu — juda muhim statistik farq. Salbiy so‘zlar bilan solishtirganda, bu kamlik Oybekning badiiy olami, tasvir uslubi va tematik yondashuvidagi hissiy og‘irlikni anglatadi. Bundan asarda iztirobli ruhiy fon ustuvorligi xulosa qilinadi. Bu raqamlar vositasida muallifning badiiy niyati aniqroq anglashiladi.

Korpusda mavjud ihtiyoriy tarzda salbiy so‘zlarni qidiruvga berib ko‘rildi, chunki tahlil natijasidagi aniqlik yuqori darajada bo‘lishi uchun ko‘proq so‘zni qidiruvga berib ko‘rilishi kerak. Bu natijani quyidagi jadval asosida ko‘rishimiz mumkin:

1-jadval. Salbiy gap statistikasi

Korpusga ixtiyoriy tarzda qidiruvga berilgan **salbiy** so‘zlar:	Natija
“Xafa”	35 ta gapda qo‘llanilgan
“Xunob”	13 ta gapda qo‘llanilgan
“Jahli chiqib”	10 ta gapda qo‘llanilgan
“Jahl bilan”	11 ta gapda qo‘llanilgan

5-qadam: Bu bosqichda “xafa”, “xunob”, “jahli chiqib”, “jahl bilan” so‘zlarining korpusdagi soni aniq jadval asosida taqdim etildi. Jadval — tahlilning vizual dalilidir. Statistik raqamlar asosida salbiy leksika “Bolalik” qissasida ijobiy so‘zlardan ancha ko‘p ishlatilgani yaqqol ko‘rinadi. Ayniqsa, bu so‘zlarning bir xil mavzu atrofida to‘plangani – bolalikdagi ruhiy siqilish, otaning yo‘qligi, ijtimoiy bosim, maktabdagi muammolar kabilarni eslatadi. Har bir so‘z, o‘z navbatida, asardagi voqealikni ochib beruvchi emotsional ko‘zgudir. (Qarang: 1-jadvalga).

Tahlilda har ikkala salbiy va ijobiy so‘zlarni gaplarda qanchalik uchraganligini, bir birini taqqoslash uchun korpusda mavjud ihtiyoriy tarzda ijobiy so‘zlarni qidiruvga berib ko‘rildi. Bu natijani quyidagi jadval asosida ko‘rishimiz mumkin:

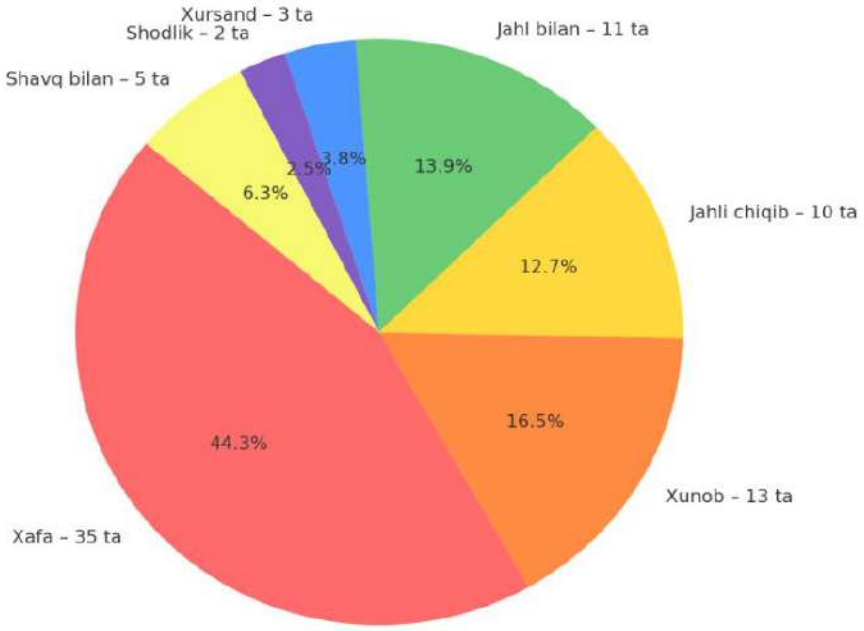
2-jadval. Ijobiy gap statistikasi

*Korpusga ixtiyoriy tarzda qidiruvga berilgan **ijobiy** so‘zlar:*	*Natija*
“Xursand”	*3 ta gapda qo‘llanilgan*

“Shodlik”	*2 ta gapda qo‘llanilgan*
“Shavq bilan”	*5 ta gapda qo‘llanilgan*

6-qadam: Bu yerda “xursand”, “shodlik”, “shavq bilan” so‘zlari kiritilgan. Bu ijobiy leksik birliklarning asarda qanchalik kam ishlatilganini ko‘rsatadi. Bu yerda “shavq bilan” kabi ifodalarning ishlatilishi — bolaning orzulari, umidlari yoki qisqa muddatli quvonchlar sahnalarida yuzaga chiqishini bildiradi. Kam uchrashi esa, Oybekning ruhiy kechinmalarini tasvirlashda asosan og‘riqli, iztirobli holatlarga urg‘u berganini bildiradi. (Qarang: 2-jadvalga).

Bu tahlil natijasini jadval ko‘rinishida ko‘rib oldik, ana endi judayam aniqlik kiritishimiz uchun va barcha soha vakillariga tushunarli bo‘lishi uchun diagramma holatida ham chizildi. Bu diagrammada “Bolalik” qissasidagi barcha hissiy tahlilga xos bo‘lgan so‘zlar jamlanib, birinchi o‘rinda har bitta so‘zni korpusga qidiruvlar asosida nechta gapda qo‘llanilganini aniqlanib, ularni umumiy hisobda 100% qilib darajalaganganda har bitta so‘zning korpusda uchrashlar soniga qarab, Oybekning bu asarida nechi foiz qo‘llanilganini ko‘rishimiz mumkin:



6-rasm. Hissiy tahlil darajalari

Xulosa

Oybekning “Bolalik” qissasidagi hissiy leksika asarining mazmuni va badiiy uslubini tushunishga yordam beradi. Birinchidan, korpus tahliliy natijalari shuni ko‘rsatdiki, Oybekning “Bolalik” qissasida salbiy hissiy leksika ancha keng qo‘llangan. “Xafa” (35 ta), “xunob” (13 ta), “jahli chiqib” (10 ta), “jahl bilan” (11

ta) kabi so‘zlar muallifning asarda iztirobli, murakkab bolalik kechinmalarini tasvirlashga alohida urg‘u berganini ko‘rsatadi. Bu salbiy leksik birliklar asardagi ruhiy muhitni og‘ir va jiddiy tarzda ifodalaydi. Muallif bolalik davridagi psixologik og‘irlik, ijtimoiy bosim va hayot sinovlarini realistik va chuqur ifodalashga erishgan.

Ikkinchidan, tahlil natijalarida ijobiy hissiy leksika son jihatdan sezilarli darajada kam uchragani aniqlandi. “Xursand” (3 ta), “shodlik” (2 ta), “shavq bilan” (5 ta) kabi so‘zlar bolalikdagi ijobiy lahzalar, quvonchli onlarni ifodalovchi birliklardir. Ijobiy so‘zlarning kamligi, muallif tomonidan asardagi ruhiy murakkabliklarga ustuvorlik berilganini anglatadi. Shunga qaramay, mavjud ijobiy leksik birliklar Oybekning umid, orzu va hayotga ishonch kabi tuyg‘ularini ham yoritishga xizmat qilgan.

Uchinchidan, ushbu tahlil shuni ko‘rsatdiki, hissiy leksika Oybekning uslubida muhim stilistik vosita bo‘lib, u orqali muallif o‘zining bolalik davriga xos tuyg‘ularini chuqur yetkazishga erishgan. Korpus asosida o‘tkazilgan bu tahlil, asarning badiiy kuchini, ruhiy qatlamini va muallif tilidagi hissiy tafovutlarni ilmiy asosda yoritib berdi.

Foydalanilgan adabiyotlar:

1. Adabiyot xrestomatiyasi. 8 jildli. 8-jild. – T.: “G’afur G’ulom nashriyoti”, 2018.
2. Oybek. Mukammal asarlar to‘plami. 19 jildlik. –T.: G‘afur G‘ulom nomidagi Adabiyot va san’at nashriyoti, 1975-1980.
3. Oybek. Tanlangan asarlar to‘plami. 9 jild. – Jild 9. – T.: “Adabiyot va san’at”, 1975.
4. https://es.mydramalist.com/people/112239-pang-bo
5. https://shoutoutatlanta.com/meet-lillian-lee-writer-producer-director-ceo/
6. https://alchetron.com/Peter-Turney
7. https://www.imdb.com/es-es/title/tt27843528/
8. https://uniwork.buxdu.uz/ilmiy/resurs.asp?id=566319
9. https://compling.navoiy-uni.uz/index.php/conferences/article/view/303/305
10. https://book.uz/books/details/oybegim-mening-3
11. https://www.imdb.com/es-es/title/tt27843528/

BADIIY ASARLAR KORPUSIDA OMONIM SHAKLLAR TAHLILI

Hamroyeva Shahlo Mirdjonovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
Gavharshodbegim Xudoynazarova,
Kompyuter lingvistikasi yo‘nalishi talabasi
xudoynazarovagavhar@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada badiiy korpus yaratish va uning Sketch Engine tizimida tahlil qilinishi muhokama qilinadi. Maqolada qidiruv tizimida qo‘shimchalar bilan bog‘liq muammolar va omonimik shakllarning turli holatlari aniqlanadi. Xususan, qo‘shimchalar bordek ko‘rinib, aslida qo‘shimchaga ega bo‘lmagan so‘zlar, qo‘shimcha qotib qolgan so‘zlar, so‘z oxirida muammolarga olib keladigan holatlar va kontekstual o‘zgarishlar kabi istisno holatlar tahlil qilinadi. Maqola morfoanalizatorning qidiruv tizimidan foydalanishda uchraydigan muammolarni hal qilish uchun yangi metodologiyalarni ishlab chiqishga yordam beradi, hamda badiiy matnlardagi omonim qo‘shimchalar bilan bog‘liq kompleks holatlarni chuqurroq tushunishga imkon beradi.

Annotation. This article discusses the creation of a literary corpus and its analysis in the Sketch Engine system. It identifies issues related to affixes in the search system and various cases of homonymous forms. In particular, the article examines exceptional cases such as words that appear to have affixes but actually do not, words with fossilized affixes, problematic word endings that cause errors, and contextual changes. The study contributes to the development of new methodologies to address challenges encountered when using a morphological analyzer in the search system and provides a deeper understanding of complex cases involving homonymous affixes in literary texts.

Аннотация. В данной статье рассматривается создание художественного корпуса и его анализ в системе Sketch Engine. В статье выявляются проблемы, связанные с аффиксами в поисковой системе, а также различные случаи омонимичных форм. В частности, анализируются такие исключительные случаи, как слова, которые внешне выглядят как имеющие аффиксы, но фактически их не содержащие, слова с застывшими аффиксами, проблемные окончания слов, приводящие к ошибкам, а также контекстуальные изменения. Статья способствует разработке новых методологий для решения проблем, возникающих при использовании морфоанализатора в поисковой системе, а также позволяет глубже понять сложные случаи омонимичных аффиксов в художественных текстах.

Kalit so‘zlar. *Badiiy korpus, metodologiya, kompleks holat, omonim qo‘shimchalar.*

Badiiy korpus yaratish va ularni tahlil qilish − tilshunoslikda zamonaviy texnologiyalardan foydalanishning muhim qismidir. Korpuslarni tahlil qilishda, lingvistik jihatdan ayniqsa qiyin bo‘lgan omonimlar va qo‘shimchalar o‘zining alohida o‘rniga ega. Bu maqolada biz Sketch Engine dasturi orqali yaratgan badiiy korpusda uchragan omonimik shakllar va qo‘shimchalar bilan bog‘liq tahlilga to‘xtalamiz. O‘tkir Hoshimovning “Ikki karra ikki besh”, “Dunyoning ishlari”, “Hurkitilgan tush”, Tohir Malikning “Shaytanat”, “Jinoyatning uzun yo‘li”, va Omina Shelinko‘g‘lining “Iskanja” asarlarini raqamlashtirib, dasturga yuklash orqali biz turli xil omonim so‘zlarni va ularning qo‘shimchalarini tahlil qilishni amalga oshirdik.

Badiiy asarlar korpusini ihlab chiqish va tahlil qilish

Sketch Engine[8] − bu tilshunoslik tahlili uchun ishlatiladigan onlayn dastur bo‘lib, foydalanuvchilarga matnlarini avtomatik ravishda tahlil qilish imkonini beradi. Sketch Engine − lingvistik korpuslarni avtomatik tahlil qilish uchun ishlab chiqilgan ilg‘or platformalardan biri bo‘lib, turli tillardagi katta hajmdagi matnlarni qayta ishlash imkonini beradi. Ushbu tizim quyidagi afzalliklarga ega:

1) matnlardan avtomatik ravishda so‘z shakllari va ularning kontekstini ajratib olish;

2) omonimik birliklarning real foydalanish chastotasini tahlil qilish;

3) morfologik va sintaktik xususiyatlarni aniqlash orqali til tadqiqotlarini osonlashtirish.

Ayniqsa, Sketch Engine tizimi tilshunoslar va sun’iy intellekt dasturchilari uchun katta ahamiyatga ega bo‘lib, u lingvistik model va algoritmlarni yaratishda qo‘llaniladi.

Tahlil qilinayotgan badiiy korpusda 6 ta asar mavjud:

1. O‘.Hoshimov. Ikki karra ikki besh.
2. O‘.Hoshimov. Dunyoning ishlari.
3. O‘.Hoshimov. Hurkitilgan tush.
4. Tohir Malik. Shaytanat.
5. Tohir Malik. Jinoyatning uzun yo‘li.
6. Omina Shelinko‘g‘li. Iskanja (tarjima asar).

Ushbu asarlar adabiyotda keng qo‘llaniladigan leksik va semantik elementlarga boy bo‘lib, ularning har biri badiiy uslubni o‘rganishda muhim manbadir.

Bugungi kunda lingvistika va sun’iy intellekt tahlillarida badiiy korpus yaratish muhim vosita sifatida ahamiyat kasb etadi. Til birliklarini chuqur o‘rganish va ularning kontekstga bog‘liq holda qanday o‘zgarishini aniqlashda korpus lingvistikasi samarali usul bo‘lib, tahlil jarayonini aniq va tizimli olib borishga imkon beradi. Badiiy korpuslar nafaqat lingvistik tahlil, balki stilistika, semantika va pragmatika bo‘yicha tadqiqotlarda ham muhim ahamiyat kasb etadi. Sun’iy intellekt yordamida korpus tahlilini amalga oshirish, matnlarni avtomatik kategoriyalarga ajratish, omonimik shakllarning kontekstga bog‘liq holda qanday ishlatilishini aniqlash kabi imkoniyatlarni beradi. Bu esa tilshunoslik tadqiqotlarining yanada kengayishiga xizmat qiladi.

Badiiy asarlar korpusining lingvistik tadqiqotlardagi o‘rni

Badiiy asarlar korpusini yaratish bugungi kunda lingvistika va sun’iy intellekt tahlillari uchun muhim vosita ekanligini alohida ta’kidlash lozim. Korpus lingvistikasining asosiy maqsadi til birliklarining real kontekstdagi qo‘llanilishini o‘rganishdir. Badiiy matnlar esa tilning eng tabiiy va stilistik jihatdan boy shakllaridan biri bo‘lib, leksik-semantik tadqiqotlar uchun juda muhim manba hisoblanadi. Korpus lingvistikasida badiiy matnlarning ahamiyati quyidagilarda namoyon bo‘ladi:

Til birliklarining kontekstga bog‘liq holda qanday ishlatilishini tahlil qilish. Til birliklarining kontekstga bog‘liq holda qanday ishlatilishini tahlil qilish – lingvistikaning asosiy yo‘nalishlaridan biri bo‘lib, so‘z va iboralar ma’nosining kontekstga qarab o‘zgarishini o‘rganish bilan bog‘liq. Bunday tahlil ayniqsa, semantika, pragmatika, stilistika, va tabiiy tilni qayta ishlash (NLP) sohalarida muhim o‘rin tutadi.

Stilistik va pragmatik tadqiqotlar uchun real ma’lumotlar bazasini yaratish. Stilistik va pragmatik tadqiqotlar uchun real ma’lumotlar bazasi yaratish – zamonaviy lingvistika va sun’iy intellekt sohalarida muhim yo‘nalishlardan biri hisoblanadi. Bunday bazalar tabiiy tilning murakkab jihatlarini o‘rganish, tahlil qilish va avtomatlashtirish uchun asos bo‘lib xizmat qiladi.

Sun’iy intellektga asoslangan tabiiy tilni qayta ishlash (NLP) tizimlari uchun modellarni shakllantirish. Tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sun’iy intellektning eng ilg‘or yo‘nalishlaridan biri bo‘lib, u inson tilini tushunish, tahlil qilish va yaratish imkonini beradi. NLP modellarini shakllantirish murakkab jarayon bo‘lib, u ma’lumotlarni yig‘ish, tozalash, o‘qitish va optimallashtirish kabi bosqichlarni o‘z ichiga oladi.

1. Omonim so‘zlar va qo‘shimchalar tahlili

Tilshunoslikda omonimiya masalasi eng muhim va murakkab yo‘nalishlardan biri hisoblanadi. Chunki bir xil shakldagi so‘zlar turli kontekstlarda turlicha ma’noga ega bo‘lishi mumkin. Sketch Engine tizimi orqali badiiy korpusda qo‘shimchalar va omonimik shakllarni qidiruv asosida tahlil qilindi. Qidiruv tizimida ba’zi omonim so‘zlar va qo‘shimchalar bilan bog‘liq muammolar aniqlandi. Bu muammolarni quyidagi 3 ta guruhga ajratish mumkin:

Qo‘shimcha bordek ko‘rinadi, lekin qo‘shimchasi bo‘lmagan so‘zlar. Ba’zi so‘zlar tashqi ko‘rinishda qo‘shimchaga o‘xshasa ham, aslida qo‘shimcha bilan bog‘lanmagan. Masalan: Chinni (chin+ni) – bu so‘zda ***-ni*** shakli mavjud, lekin morfem jihatdan qo‘shimcha emas. Pechka (pech+ka) – bu so‘zda ham tashqi ko‘rinishda qo‘shimcha mavjud kabi ko‘rinadi, lekin bu so‘zda ***-ka*** bo‘g‘in, qo‘shimcha emas. Bunga o‘xshash bir qancha so‘zlar mavjud: *(1-2-3-4-jadvallar)*

1-jadval. Tushum kelishigi (-ni) shakli omonimiyasi

№	So‘z	-ni
1.	Chinni	chin+ni
2.	Ayni	ay+ni
3.	Jinni	jin+ni
4.	Mini	
5.	Luqmoni	
6.	Poni	

2-jadval. Jo‘nalish kelishigi (-ka) shakli omonimiyasi

№	So‘z	-ka
1.	Pechka	pech+ka
2.	Yekla	yel+ka
3.	Sirka	sir+ka
4.	Erka	er+ka
5.	Tikka	tik+ka
6.	To‘nka	to‘n+ka

3-jadval. Jo‘nalish kelishigi (-qa) shakli omonimiyasi

№	So‘z	-qa
1.	Halqa	hal+qa
2.	Orqa	or+qa
3.	Qisqa	qis+qa
4.	Buqa	bu+qa
5.	Boshqa	bosh+qa
6.	Aloqa	alo+qa
7.	Yoqa	yo+qa
8.	Loyqa	loy+qa
9.	Antiqa	anti+qa
10.	Manqa	man+qa

4-jadval. Chiqish kelishigi (-dan) shakli omonimiyasi

№	So‘z	-dan
1.	Aftidan	aft+i+dan
2.	Jumladan	jumla+dan
3.	Boshqatdan	boshqa+t+dan
4.	Zimdan	
5.	Obdan	
6.	Qo‘qisdan	

Qo‘shimcha qotib qolgan (leksemalashgan) so‘zlar. Ba’zi so‘zlarda qo‘shimchalar o‘zining asl ma’nosini yo‘qotgan, so‘z tarkibida qotib qolgan. Misol uchun: Boshqacha (boshqa+cha) – bu so‘z semantik va leksik jihatdan o‘zgarib, tarkibiy qismi qotib qolgan so‘zlardan biridir. Bu holatda “-cha” qo‘shimchasi o‘zining asl ma’nosini yo‘qotgan va yangi ma’noga ega bo‘lgan. Asosan ravish so‘z turkumidagi so‘zlarda qo‘shimchalar qotib qoladi va ular ajratilmaydi. Bunga o‘xshash bir qancha so‘zlar mavjud: (Qarang: 5-jadval)

5-jadval. Tarkibida qo‘shimcha qotib qolgan so‘zlar

№	So‘z	qo‘shimcha
1.	Boshqacha	boshqa+cha
2.	Birdan	bir+dan
3.	Ataydan	atay+dan
4.	Nihoyatdan	nihoyat+da
5.	Kunda	kun+da
6.	Birga	bir+ga
7.	Chalqanchasiga	chalqancha+si+ga

8.	Birdaniga	bir+dan+i+ga
9.	Yonlamasiga	yonlama+si+ga
10.	Qatorasiga	qator+a+si+ga
11.	Bekorga	bekor+ga

So‘z qidiruvidagi noaniqlikni keltirib chiqaruvchi shakllar. O‘zbek tilida shunday so‘zlar borki, ularning oxirgi bo‘g‘ini qo‘shimchalar bilan omonimlik hosil qiladi va bu qidiruv tizimida muammoga olib keladi. Masalan: Jonona (jon+ona) – bu so‘zning oxirgi bo‘g‘ini “ona” qo‘shimchasi bilan omonimlik hosil qila oladi va qidiruv tizimida tarkibida “ona” bo‘lgan so‘zlarni ham qo‘shimcha sifatida chiqariladi. Bunga o‘xshash bir qancha so‘zlar mavjud (Qarang: 6-jadval).

6-jadval. So‘z qidiruvida muammoga uchraydigan so‘zlar.

№	So‘z	Ona
1.	Mastona	mast +ona
2.	Kamtarona	kamtar +ona
3.	Ayyorona	ayyor +ona
4.	Ojizona	ojiz +ona
5.	Mardona	mard +ona
6.	Ahmoqona	Ahmoq +ona

2. Tizimdagi muammo va yechimlar

Omonim so‘z va qo‘shimchalarning muammoli holatlari Sketch Engine tizimida ba’zi natijalarning noto‘g‘ri ko‘rsatishiga olib keldi. Bu, ayniqsa, qo‘shimchalar o‘zining leksik shaklini o‘zgartirgan yoki qotib qolgan so‘zlarda uchradi. Agar bu muammolarni hal qilmasak, qidiruv tizimi noto‘g‘ri natijalar ko‘rsatadi. Muammolarni bartaraf etish uchun quyidagi yondashuvlar talab etiladi:

1. So‘zlarni tahlil qilishda semantik qatlamlarni hisobga olish: qo‘shimchalar va omonim so‘zlarning ma’nolarini tushunish va ularning kontekstga qarab qanday ishlashini aniqlash.

2. Qo‘shimcha va so‘z shakllarini tekshirish: korpusdagi har bir so‘zni alohida tahlil qilib, qaysi qo‘shimchalar ishlatilishini aniqlash va ularga asoslangan alohida qidiruv mexanizmlarini ishlab chiqish.

3. Istisno holatlar va muammolar

Istisno holatlar mavjud bo‘lib, ular quyidagi uch guruhda tahlil qilindi:

1. Qo‘shimcha bordek ko‘rinadi lekin qo‘shimchasi bo‘lmagan so‘zlar.
2. Qo‘shimcha qotib qolgan so‘zlar.
3. So‘zni qidirganda oxirida muammoga uchraydigan so‘zlar.

Xulosa. Sketch Engine tizimidan foydalangan holda badiiy korpusda omonimik shakl va qo‘shimchalarni tahlil qilish tilshunoslik tadqiqotlarida muhim o‘rin tutadi. Biroq ba’zi muammolarni, ayniqsa, qo‘shimchalar va omonimlar bilan bog‘liq holatlarni aniqlash va tuzatish zarur. Yangi tahlil metodologiyalarini ishlab chiqish, so‘zning morfem tarkibini to‘g‘ri aniqlash va qidiruv tizimlarini optimallashtirish badiiy matnlar va lingvistik tahlilni yanada samarali qilishga yordam beradi. Ushbu tadqiqot natijalari, ayniqsa, avtomatlashtirilgan tarjima tizimlari, matnlarni avtomatik qayta ishlash, nutqni aniqlash va sun’iy intellekt asosidagi lingvistik dasturlarni rivojlantirishda amaliy ahamiyat kasb etadi. Omonim shakllarning kontekstual o‘zgarishlarini chuqur o‘rganish orqali tilga oid dasturlarni yanada mukammallashtirish imkoniyati yuzaga keladi.

Foydalanilgan adabiyotlar:

1. Hoshimov O. Dunyoning ishlari: Qissa va hikoyalar. – Toshkent: Yangi asr avlodi, 2015.
2. Malik T. Shaytanat. 1-qism. – Islom.uz ma’rifat markazidan olindi. Manba: www.ziyouz.com, 2007.
3. Malik T. Jinoyatning uzun yo‘li. – PDF versiyasi. Manba: www.ziyouz.com.
4. Hoshimov O. Hurkitilgan tush: Hikoyalar. – Toshkent: Ziyo nashr, nashr yili ko‘rsatilmagan.
5. Shenliko‘g‘li O. Iskanja. – Toshkent, 2021. (Nashr tili: o‘zbek, kirill).
6. Hoshimov O. Ikki karra ikki – besh. – Toshkent: Nurli dunyo nashriyot uyi, 2024.
7. Abjalova M. Korpus lingvistikasi: Uslubiy qo‘llanma. – Toshkent: Bookmany Print, 2022.
8. Sketch Engine. Sketch Engine – corpus analysis platform. Manba: https://www.sketchengine.eu.
9. Yusupova R. D. Sketch Engine platformasida korpuslar yaratishning amaliy tadqiqi // *E-Conference Series.* – ISSN: 2835-5326. – 46-bet. Manba: econferenceseries.com.
10. Zaxarov V., Mengliyev B., Xamroyeva Sh. Korpus lingvistikasi: Korpus tuzish va undan foydalanish. – Globe Edit

O‘ZBEK VA FORS TILLARI UCHUN PARALLEL KORPUS: LINGVISTIK VA TEXNOLOGIK YONDASHUVLAR

Xusainova Zilola Yuldashevna,
Filologiya fanlari bo‘yicha falsafa doktori
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Raxmatxonova Nozimaxon Ismoilxon qizi,
Filologiya va tillarni o‘qitish yo‘nalishi talabasi
nozimaxonrahmatxonova@gmail.com
ToshDO‘TAU

Annotatsiya. Parallel korpuslar zamonaviy kompyuter lingvistikasi, mashina tarjimasi, til o‘rgatish va NLP sohalarida muhim manba hisoblanadi. O‘zbek va fors tillari orasidagi madaniy, tarixiy va tilshunoslik aloqalari ushbu tillar uchun parallel korpus yaratish zaruratini yuzaga keltiradi. O‘zbek va fors tillari uchun hali to‘liq va sifatli parallel korpus mavjud emas, mavjud manbalar esa lingvistik va texnologik jihatdan yetarlicha boy emas. Ushbu maqola O‘zbek-Fors parallel korpusini yaratish bosqichlari, duch kelinadigan lingvistik va texnologik muammolar hamda ularning yechimlarini yoritishga qaratilgan.

Abstract. Parallel corpora are an important resource in the fields of modern computational linguistics, machine translation, language teaching, and NLP. The cultural, historical, and linguistic connections between the Uzbek and Persian languages make it necessary to create a parallel corpora for these languages. There is no complete and high-quality parallel corpora for the Uzbek and Persian languages, and the existing sources are not sufficiently rich in linguistic and technological terms. This article aims to highlight the stages of creating an Uzbek-Persian parallel corpus, the linguistic and technological problems encountered, and their solutions.

Аннотация. Параллельные корпуса являются важным ресурсом в областях современной компьютерной лингвистики, машинного перевода, изучения языков и обработки естественного языка. Культурные, исторические и языковые связи между узбекским и персидским языками обусловливают необходимость создания параллельного корпуса для этих языков. До сих пор не создан полный и качественный параллельный корпус для узбекского и персидского языков, а имеющиеся источники недостаточно богаты в лингвистическом и технологическом плане. Целью данной статьи является освещение этапов создания узбекско-персидского параллельного корпуса, возникших лингвистических и технологических проблем и путей их решения.

Kalit so‘zlar. *Korpus, parallel korpus, o‘zbek tili, fors tili, NLP, statitik tahlil, raqamlashtirish, mashinali o‘qitish, lingvistik muvofiqlik*

Kirish. Tabiiy tilni kompyuter yordamida tadqiq qilish va qayta ishlash to‘rt jihatni: formallashtirish, algoritmlash, dasturlash va amaliyotga qo‘llashni o‘z ichiga oladi[1]. Tabiiy tilni qayta ishlash (NLP)ga asoslangan dasturiy ilovalar(axborot tizimlari)ni ishlab chiqish uchun NLP tizimini mavjud ma’lumotlarni o‘rganishini (o‘qitishini) ta’minlash kerak. Ushbu amalni til korpusi vositasida amalga oshirish mumkin. Til korpusi elektron shaklda taqdim etiladigan katta hajmli va strukturlangan matnlar to‘plami sifatida qaraladi. Til korpusi yozma yoki og‘zaki materialni ifodalab, NLP tizimini mavjud resurslarni o‘rganishi uchun lingvistik tahlilni amalga oshirish lozim. Til korpusi tabiiy tilni qayta ishlash tizimining asosidir. Unda gazeta, roman, retsept, radioeshittirishdan tortib teleko‘rsatuv, film va tvitlargacha turli shakldagi ma’lumotlar bo‘lishi mumkin[2].

Korpus ikki yoki undan ortiq til doirasidagi ma’lumotlar bazasi bo‘lib, undan ko‘plab maqsadlarda foydalanish mumkin. Shuningdek, korpus tillar o‘rtasidagi muloqot vositasi hisoblanib, bu tizim yirik elektron ko‘rinishdagi matnlar jamlanmasidir. Korpus lingvistikasining *statistik* va *hisoblash* usullari turli tillarning **grammatik, leksik** va **semantik** xususiyatlarini tizimli ravishda tahlil qilish imkonini beradi. U katta hajmdagi yozma yoki og‘zaki matnlarni tahlil qilishga asoslanadi. Korpus aniq maqsad uchun **yig‘ilgan** va **raqamlashtirilgan** matnlar to‘plami bo‘lib, u turli xil lingvistik tahlillar uchun ishlatiladi[3]. Parallel korpuslar bir tildan ikkinchi tilga tarjima qilingan matnlarni o‘z ichiga olganligi sababli, ular **lingvistik va hisoblash tahlili uchun muhim manba hisoblanadi**.

Shuningdek so‘z turkumlariga xos kategoriyalar, grammatik ma’noning ifodalanishi, so‘z yasalish tizimiga oid ma’lumotlar bera olishi bilan ham e’tiborga molikdir. Parallel matnlar korpusi chog‘ishtirma aspektdagi tadqiqotlarda o‘ziga xos o‘rin tutadi. Parallel korpuslar yordamida bir so‘z, jumla, abzats, supersintaktik butunliklarning turli tillardagi variantlarini bilish mumkin. Parallel korpuslar madaniyatlararo muloqot keng yoyilgan hozirgi davr uchun muhim bo‘lgan ma’lumotlar bazasidir. Parallel korpuslar orqali turli til muhitdagi, madaniyatlardagi universaliyalar hamda tillarning o‘ziga xos mental xususiyatlari, realiya va lakunar birliklarni aniqlash mumkin bo‘ladi. Parallel matnlar korpusi avtomatik tarjima rivoji uchun ham xizmat qiladi, kompyuter leksikografiyasining taraqqiyotini ta’minlaydi[4]. Parallel matnlar korpusi turli tillar asosida xilma-xil madaniyatlarni

solishtirish, madaniy munosabatlarni ifodalovchi leksik birliklar mazmun-mohiyatini o‘zlashtirish imkoniyatini beradi. Parallel korpusning NLPdagi roli turlicha, xususan, NLPning ko‘plab vazifalarini hal qilishda til korpuslaridan foydalaniladi:

– **Mashinali o‘qitish modellarini o‘rgatish.** Turli NLP ilovalari, jumladan hissiyotlarni tahlil qilish, matn tasnifi, mashina tarjimasi va nutqni tanib olish uchun korpuslar mashinali o‘qitish modellarini o‘rgatish va takomillashtirish uchun ishlatiladi. Korpusdagi katta hajmdagi matn ma’lumotlar ushbu modellarga shablonlar (patterns), korrelyatsiyalar (correlations) va murakkabliklar (complexities)ni o‘rgatish uchun ishlatiladi.

– **Tilni tushunish.** Til korpusi tabiiy tilning tuzilishi, grammatikasi, lug‘ati va qo‘llanilishi haqida to‘liq tasavvur beradi. Kontekstda so‘zlar va so‘z birikmalarning qanday ishlatilishini o‘rganish uchun NLP modellarida til korpusidan foydalaniladi. Chet tili o‘qituvchilari va o‘rganuvchilar uchun til korpusi tilning haqiqiy misollarini taqdim etadi va haqiqiy foydalanish shakllarini ochib beradi. U o‘quv dasturlarini ishlab chiqish, materiallarni loyihalash va hatto dars mashg‘ulotlarida foydalanish mumkin. Korpus ko‘plab til resurslarini, jumladan, grammatika tekshiruvi, imlo tekshiruvi va nutqni aniqlash dasturlarini ishlab chiqishga yordam beradi. Ushbu NLP ilovalari bizning kundalik raqamli tajribamizda ajralmas rol o‘ynaydi.

– **Diskurs tahlili.** Korpus nutqni tahlil qilish uchun ham ishlatilishi mumkin, bu tilning muayyan kontekstlarda qanday qo‘llanilishi haqida tushuncha beradi. Bu bizning siyosiy, ijtimoiy va madaniy nutqlar haqidagi tushunchamizni yanada oshirishi mumkin.

– **Tarjimashunoslik.** Til korpusi tarjimashunoslik uchun muhim vosita hisoblanadi. Ikki tilli yoki ko‘p tilli korpus yaratish tilshunoslarga tarjima me’yorlarini o‘rganish, ma’no jihatdan mos tarjimalarni aniqlash va samarali mashina tarjimasi modellarini yaratish imkonini beradi.

– **Qoidalarga asoslangan tizimlar.** Til korpuslari tilshunoslar va NLP mutaxassislari tomonidan til qoidalari va shablonlarini ishlab chiqish va sinab ko‘rish uchun ishlatiladi. Keyinchalik, POS teglash, grammatik qayta ishlash va NERlarni tanib olish kabi NLP vazifalari uchun ushbu qoidalarga asoslangan NLP tizimlarida qo‘llaniladi.

– **Lug‘at va semantika.** Turli lug‘atlar yoki so‘zlardan iborat to‘plamlar til korpuslari yordamida yaratiladi va kengaytiriladi. Sinonimlar, antonimlar va so‘z birikmalari kabi so‘z munosabatlarini ko‘rsatish orqali ular semantik tahlilga yordam beradi. Korpuslar asosida lug‘atlarni yaratish usulini tubdan o‘zgartirib,

leksikografiya sohasini tubdan o‘zgartirdi. Katta hajmli strukturlangan matnlar to‘plamini tahlil qilib, leksikograflar yangi so‘zlarni, sezgilarni va foydalanish shakllarini yanada samaraliroq aniqlashlari mumkin.

– **Statistik tahlil.** Til korpuslari tabiiy tilning statistik tahlili uchun foydalidir. Ular yordamida so‘z chastotasini taqsimlash, birgalikda paydo bo‘lish shablonlari va boshqa statistik xususiyatlarni o‘rganish uchun ehtimollikka asoslangan NLP yondashuvlari uchun zarur bo‘lgan ma’lumotlarni taqdim etadi.

– **Sohaviy bilimlar.** Til korpuslari turli sohagalarga oid bilimlar manbayi hisoblanadi. Chunki ular muayyan mavzular yoki sohalarga xos bo‘lishi mumkin. Huquqiy hujjatlarni o‘rganish, tibbiy ma’lumotlarni qayta ishlash va muayyan sohalar uchun yaratilgan chatbotlar kabi ilovalar bunga yaqqol misol bo‘ladi[5].

O‘zbek va fors tillarining ta’limiy parallel korpusi hozircha yaratilmagan bo‘lib, biroq ikkala tilning ham boshqa tillar bilan aloqadorlikda yaratilgan korpuslari mavjud, masalan, o‘zbek-ingliz, o‘zbek-turk va fors-ingliz, fors-fransuz, fors-rus, fors-nemis, fors-arab kabilar. Ko‘rinib turibdiki, fors tilidagi korpuslar o‘zbek tiliga nisbatan son jihatdan ustun. O‘zbek tili bilan bevosita aloqador korpuslar sekin-asta rivojlanib bormoqda. O‘zbek-fors tillari o‘rtasidagi korpus ham ana shu harakatlarning namunalaridandir. Bu shakllantirilayotgan korpusda ikkala tilning barcha xususiyatlari, ya’ni leksik, sintaktik, semantik, morfologik tomonlari parallel ravishda o‘rganildi. Umuman olganda, fors tili IX-XIV asrlarda turkiy davlatlarda, ayniqsa, Xuroson, Samarqand, Buxoro, Xorazm kabi Markaziy Osiyoning yetakchi shahar va viloyatlarida keng iste’molda bo‘lgan. Ana shu davrda turkiy tilga fors tilidan ko‘plab so‘zlar o‘zlashgan bo‘lib, bugungi kunda o‘zbek tilining lug‘at tarkibida forscha o‘zlashmalar taxminan 30-40% atrofida bo‘lishi mumkin.

Korpusni yaratish davomida bir nechta qiyinchiliklarga duch kelinadi. Chunki kam resursli tillar bilan bog‘liq korpuslarda ma’lumotlar bazasini yig‘ish biroz murakkab, tarjima jarayonida ko‘proq inson omilidan foydalanishga to‘g‘ri keladi. Yana bir masalani sabab qilib ko‘rsatish mumkinki, u ham bo‘lsa, o‘zbek tili agglyutinativ tillar oilasiga, fors tili esa flektiv tillar oilasiga kiradi. Quyidagi jadval orqali batafsilroq ma’lumotga ega bo‘lamiz.

Til tizimi xususiyati	**O‘zbek tili**	**Fors tili**
Alifbo	Lotin (ba’zi manbalarda Kirill)	Arab yozuvi

Fonetika	Unli tovushlar: 6 ta: a, e, i, o, u, o‘. Undosh tovushlar: 24 ta.	Unli tovushlar: 6 ta. Yozuvda qisqa unlilar ifodalanmaydi (e, a, o‘): کتاب, بهار، شتر kabi. Undosh tovushlar 23 tadir.
Morfologiya	Agglyutinativ	Flektiv
Fe’l zamonlari	3 ta (Hozirgi, O‘tgan, Kelasi)	Asosan 2 ta: o‘tgan zamon va hozirgi-kelasi zamon. (O‘tgan zamonning o‘zi ham ingliz tili kabi bir nechta ichki turlarga bo‘linadi, masalan, o‘tgan zamon davom fe’li, o‘tgan zamon natijali fe’li, uzoq o‘tgan zamon fe’li kabi)
So‘z tartibi	SOV: Men ertaga universitetga boraman.	SOV: من فردا به دانشگاه می روم
Fe’l shakllari	Fe’l + yordamchi so‘zlar: sifatdosh(kelgan,ishlaydigan), ravishdosh(kulib, borgach), harakat nomi(o‘qimoq, salomlashish)	Fe’l va uning o‘zgaruvchan shakllari: sifatdosh(امده، نشسته), ravishdosh(maxsus qo‘shimchasi yo‘q) – sifatdosh shaklidagi fe’l bilan ifodalanadi, gapda ifodalangan mazmun orqali ajratib olinadi
So‘z yasash	Qo‘shimchalar va kompozitsiya usuli orqali: tekislamoq, terimchi, hozirjavob	Qo‘shimchalar va birikish usuli orqali: کتابخانه ، همراه، کارگر ،دوربین.
Kelishiklar	6 ta (bosh kelishik shakli yo‘q, qaratqich, tushum, jo‘nalish, o‘rin-payt va chiqish)	Kelishik kategoriyasi yo‘q, ma’lum bir predloglar orqali ifodalanadi: به، در، از، با، تا، برای، بر، را
Predlog/yordamchi birliklar	Pregloglar ham yordamchi birliklar ham ko‘p	Predloglar ko‘p
Yozuv tizimi	Chapdan o‘ngga: → Mening uyim Pushkin ko‘chasida joylashgan.	O‘ngdan chapga: ← خانه ی ما در خیابان پوشکین است

Til korpusini yig‘ishdagi eng muhim omillardan biri sifatida manba, ya’ni ma’lumotlar bazasini to‘plashni olishimiz mumkin. Korpusning hajmi qancha yirik bo‘lsa, undan foydalanish koeffitsiyenti shuncha baland bo‘ladi. Faqatgina hajm bilan cheklanib qolmaslik zarur, chunki korpusga kiritiladigan ma’lumotlar turli xil bazalardan to‘planadi, masalan, rasmiy qonun yoki hujjatlar, badiiy asarlar, ilmiy tadqiqotlar va hokazo. Manbalarni to‘plash jarayonida tarjimaga ham e’tibor berib boriladi, matnlar sifatining yuqori bo‘lishi ishning tizimli ravishda o‘sishiga yordam

beradi. Yetarlicha manba to‘planib bo‘lgandan so‘ng, ishning ikkinchi bosqichiga o‘tiladi. Tozalash (dataset) bosqichi yordamida matnlar UTF-8 formatida shakllantiriladi. Ma’lumotlar to‘plamini juda ko‘p xatolari bo‘ladi va qayta-qayta tozalashdan o‘tadi. Avtomatik tarjima yordamida yig‘ilgan ma’lumotlar inson omili yordamida ham bir necha bor tekshiriladi. Shundan so‘ng matnlar keraksiz belgi va teglardan tozalanadi. Bu jarayonni Python kodlari yordamida avtomatik ravishda amalga oshirish mumkin. Keyingi muhim qadamlardan biri parallel bo‘lmagan jumlalarni transliteratsiya qilishdir. Korpusdagi har bir jumlalarni kerakli tilga transliteratsiya qilish lozim bo‘ladi. Maxsus dastur va algoritmlar yordamida avtomatik transliteratsiya usulidan foydalaniladi, chunki qo‘lda transliteratsiya qilish juda ko‘p vaqtni oladi. Sifat nazoratini tekshirish orqali korpusimizning qiymatini belgilab olamiz. Ba’zida keraksiz belgilar qolib ketgan bo‘lishi mumkin yoki tarjima jarayonida matnlar bir-biriga aralashib ketish hollari kuzatiladi. Bunday holda keraksiz belgilar o‘chiriladi, aralash matnni ajratib olinadi va to‘g‘rilanadi. Bir-biriga ma’nosi to‘g‘ri kelmaydigan yoki umuman xato tarjimalar bor yoki yo‘qligini ko‘rib chiqish kerak. Masalan: من مدرسه رفتم

✓ Men maktabga bordim.

× Men universitetdan qaytdim.

Sifat nazorati to‘liq tekshirilib bo‘lingandan so‘ng, korpusga biroz qo‘shimchalar kiritish mumkin, masalan, bironta gap keltirilganda uning aynan tarjimasi bilan birgalikda unga teng keluvchi muqobil variantlarni ham kiritiladi va shu tariqa korpusning ichki manbasi boyishiga sezilarli hissa qo‘shiladi. Buning uchun esa ko‘proq badiiy asarlardan namunalar kiritilishi lozim.

Xulosa. Parallel korpus tilshunoslik, tarjima va sun’iy intellekt sohalarida muhim ahamiyat kasb etadi. Tarjimonlar yoki til o‘rganmoqchi bo‘lgan havaskorlar uchun manfaatli bo‘lib, ular tarjima jarayonida nafaqat gap tarjimalarini balki o‘sha so‘zga to‘g‘ri kelishi mumkin bo‘lgan muqobil variantlarni ham ko‘rib chiqish imkoniga ega bo‘ladilar. Tilshunoslar esa o‘zbek va fors tillari o‘rtasidagi grammatik va leksik munosabatlarni chuqur tahlil qila olishadi. Dasturchilar yoki AI muhandislari uchun avtomatik tarjima tizimlarini rivojlantirishda asosiy ma’lumotlar bazasi sifatida ishlatiladi. Keyinchalik uni rivojlantirib borishimiz natijasida avtomatlashtirilgan tarjima tizimi yaratilib, forscha kontentlar uchun o‘zbekcha subtitrlar ishlab chiqilishi ham maqsad qilingan. Parallel ko‘rinishda taqdim etiladigan subtitrlar fors tilini o‘rganayotganlar uchun qulaylik yaratadi.

Foydalanilgan adabiyotlar:

1. Elov B.B., Xusainova Z.Y. Katta til modellari / “O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi”. 2024. Vol 2(6). – P. 78-92.
2. Xusainova, Z., & Yangibayeva, S. (2024). Til korpusi turlari. Uzbekistan: Language and Culture, 2024. Vol 2(6). – P. 54-64.
3. Elov, B., & Xusainova, Z. (2024). Til korpuslarini lingvistik teglash bosqichlari. // Computer Linguistics: problems, solutions, prospects, 1(1).
4. Akramdjanova, M. (2024). Parallel matnlar korpusini yaratishning amaliy va nazariy ahamiyati. *Наука и инновация*, *2*(34), 34-37.
5. .B.Elov, A.U.Yuldashev, U.S.Yodgorov. Til korpusi turlari va umumiy xususiyatlari / Xorazm Ma’mun akademiyasi axborotnomasi: ilmiy jurnal.- 2024. №12/4 (121). – P.118-125.
6. Elov, B. B., Hamroyeva, S., & Xusainova, Z. (2022). NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar. TerDU, filologik tadqiqotlar: til, adabiyot, ta’lim, 5-6.
7. B. Elov, Sh. Hamroyeva, R. Alayev, Z. Xusainova, U. Yodgorov – O‘zbek tili korpusi matnlarini qayta ishlash usullari – Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali, October 2023 – Volume 1, Issue 3.
8. Ro‘Ziyev H. B., Xodjayeva N. T. Matnlarning parallel korpuslarini tuzish muammolari //Academic research in educational sciences. – 2021. – T. 2. – №. 9. – C. 515-522.
9. Sobirova N. G. Q. Korpus lingvistikasi va parallel korpuslar tavsifi //Academic research in educational sciences. – 2022. – T. 3. – №. 5. – C. 1442-1450.

XORAZM SHEVASIDAGI MATNLARNI O‘ZBEK ADABIY TILIGA O‘GIRISH DASTURINI YARATISHNING AMALIY AHAMIYATI

Xusainova Zilola Yuldashevna,
Filologiya fanlari bo‘yicha falsafa doktori
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU
Bekchanova Munira Oybek qizi,
Kompyuter lingvistikasi yo‘nalishi 3-kurs talabasi
bekchanovamunir@gmail.com
ToshDO‘TAU

Annotatsiya. Bugungi kunda tilshunoslikda hududiy shevalarning tizimli o‘rganilishi va elektron resurslar shaklida saqlanishi katta ahamiyat kasb etmoqda. Xorazm shevasi o‘zining fonetik, morfologik va leksik xususiyatlari bilan umumiy adabiy o‘zbek tilidan sezilarli farq qiladi. Xorazm shevasi korpusining yaratilishi o‘zbek tilining dialektologik xilma-xilligini saqlash va o‘rganishda muhim qadam hisoblanadi. Korpus yaratish jarayonida audio yozuvlar, transkripsiyalar va matnli ma’lumotlar tizimli ravishda yig‘ilib, ularni annotatsiya qilish muhim ahamiyat kasb etadi. Matnlar turli manbalardan (suhbatlar, folklor namunalari, she’riy va nasriy asarlar, audio yozuvlar) to‘planadi, ularning lingvistik va etnografik jihatlari belgilanadi. Ushbu maqolada Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga avtomatik tarzda o‘girish imkoniyatlarini tadqiq etish va bu yo‘nalishda dasturiy ta’minot yaratish masalasi ko‘rib chiqiladi.

Аннотация. Сегодня систематическое изучение региональных диалектов и сохранение их в виде электронных ресурсов имеет большое значение в языкознании. Хорезмский диалект существенно отличается от общелитературного узбекского языка по своим фонетическим, морфологическим и лексическим признакам. Создание корпуса хорезмских диалектов является важным шагом в сохранении и изучении диалектологического многообразия узбекского языка. В процессе создания корпуса важно систематически собирать и аннотировать аудиозаписи, транскрипции и текстовые данные. Тексты собираются из различных источников (интервью, образцы фольклора, поэтические и прозаические произведения, аудиозаписи), определяются их лингвистические и этнографические аспекты. В данной статье рассматривается возможность автоматического перевода текстов на хорезмском диалекте на узбекский литературный язык и разработка программного обеспечения в этом направлении.

Abstract. Today, the systematic study of regional dialects and their preservation in the form of electronic resources is of great importance in linguistics. The Khorezm dialect is significantly different from the general literary Uzbek language with its phonetic, morphological and lexical features. The creation of a corpus of the Khorezm dialect is an important step in preserving and studying the dialectological diversity of the Uzbek language. In the process of creating a corpus, audio recordings, transcriptions and textual data are systematically collected and annotated. Texts are collected from various sources (interviews, folklore samples, poetic and prose works, audio recordings), their linguistic and ethnographic aspects are determined. This article considers the issue of researching the possibilities of automatically translating texts in the Khorezm dialect into the Uzbek literary language and creating software in this direction.

Kalit so‘zlar. *Dialektal korpus, Xorazm shevasi, NLP, mashinali o‘qitish, lingvistik o‘zgartirish.*

Dialektal nutqni adabiy tilga moslashtirish kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasida dolzarb yo‘nalishlardan biri hisoblanadi. Dialektal nutqning adabiy til bilan muvofiqlashtirilishi til birligini mustahkamlash, raqamli til resurslarini kengaytirish va mashinali tarjima tizimlarining sifatini oshirish uchun muhim ahamiyat kasb etadi. Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga o‘girish dasturini yaratish o‘zbek tilining dialektal xilma-xilligini saqlash va kengroq auditoriyaga yetkazishda muhim ahamiyatga ega. Xorazm she’vasining o‘ziga xos fonetik, morfologik va sintaktik xususiyatlari uning adabiy tilga ta’sirini va o‘ziga xosligini yanada kuchaytiradi. Xorazm viloyatining turli hududlarida yashovchi xalqning tiliga oid to‘plangan ma’lumotlarni tizimli ravishda tahlil qilish, o‘ziga xos dialektal korpus yaratish zarurati paydo bo‘lgan. Xorazm shevasida yaratilgan matnlarni o‘zbek adabiy tiliga o‘girish uchun zamonaviy texnologiyalar va lingvistik yondashuvlarni uyg‘unlashtirish muhim hisoblanadi. Ushbu tadqiqot natijasida amaliy dastur ishlab chiqiladi. Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga o‘girish dasturini yaratish sheva matnlarini avtomatik ravishda adabiy tilga o‘tkazish orqali ta’lim materiallarini tayyorlashni osonlashtiradi. Shevadagi matnlarni adabiy tilga o‘girish matnlarning kengroq kontekstda tushunarli bo‘lishini ta’minlaydi. Xorazm aholisi, ayniqsa yoshlar, adabiy til bilan muloqotda bo‘lishda qiyinchiliklarga duch kelishi, ushbu dastur ularga kommunikatsiya vositasini yengillashtiradi.

Shuningdek, maqolada Xorazm she’vasining fonetik, morfologik va sintaktik xususiyatlari, bu xususiyatlarning dialektal korpusga qanday aks etishi va uning Xorazm shevasida qanday qo‘llanishi haqida so‘z boradi. Xorazm she’vasining

dialektal korpusini yaratish, o‘g‘uz she’valarining o‘ziga xosligini aniqlash va adabiy tilda ularning rolini tahlil qilishga yordam beradi.

Dunyo tilshunosligida XXI asrga kelib, korpus muammolarini o‘rganish harakati kuchaydi. Hozirda til korpuslarining ijtimoiy ahamiyati, tuzish tamoyillari, lingvististik teglar tizimini ishlab chiqish, avtomatik tahlil qilish masalalari bo‘yicha maqsadli tadqiqotlar olib borilmoqda. Shu nuqtai nazardan olib qaraganda globallashuv jarayonida ayrim tillar qo‘llanishining kengayishi boshqa tillarning saqlanib qolishiga xavf solmoqda, shuning uchun mazkur korpuslar tilning yashab qolishi uchun imkoniyatlardan biri sifatida qarasak bo‘ladi. Korpus borasida fanda o‘ziga xos ilmiy ishlar qilinmoqda. Jahon tilshunosligida raqamli axborot vositalaridan foydalanish qulayligi hamda internet orqali milliy tillarni o‘zlashtirishga bo‘lgan harakat ta’sirida milliy korpuslarda shevalar bazasini yaratish talabi o‘smoqda. Natijada tilda mavjud dialektlarni nafaqat lingvistik, balki amaliy jihatdan tadqiq etish, bu orqali sheva birliklarini saqlab qolish, avloddan-avlodga yetkazish imkoniyatining ko‘lami kengaymoqda.

Shevadagi so‘zlar jamlanmasi va ularning elektronlargan varianti dialektal korpus deyiladi. Dialektal korpusning arxitekturasi, umuman olganda, bir nechta asosiy qism va bosqichlardan tashkil topgan bo‘lib, ularning har biri ma’lumotlar yig‘ish, tasniflash va tahlil qilish jarayonlarini o‘z ichiga oladi. Dialektal korpus bu tilning turli mintaqaviy yoki ijtimoiy turlarini ifodalovchi matnlar va og‘zaki til namunalari to‘plami. U lingvistik tadqiqotlar, til hujjatlari va madaniy merosni saqlash uchun muhim manba hisoblanadi. Dialektal korpusni boshqarish va tahlil qilish uchun yaxshi ishlangan arxitektura, ma’lumotlar bazasi tizimi zarur. Dialektal yozuvlari to‘plami keng ko‘lamli fonologik, morfologik, leksik va sintaktik tadqiqotlar uchun ham qimmatli ma’lumotlar manbayi hisoblanadi.

Ushbu maqolada Xorazm shevasiga xos matnlar o‘zbek adabiy tiliga o‘girish dasturini yaratish uchun shevadagi so‘zlar yozma manbalardan to‘plandi. Matnlar raqamlashtirildi, ortiqcha belgilardan tozalandi hamda morfosintaktik tahlil uchun tayyorlandi. Sheva-adabiy til o‘rtasidagi morfologik farqlarni aniqlash uchun qoidaga asoslangan yondashuvlar ham modelga integratsiya qilindi. Annotatsiya jarayonida matnlar shevaga xos fonetik va morfologik xususiyatlarga asoslanib belgilandi. O‘zbek adabiy tiliga o‘girish uchun dastlab qoidaviy (rule-based) yondashuv asosida o‘girish qoidalari ishlab chiqildi.

Dialektal korpusning amaliy ahamiyati: umumiy qilib olganda, shevaga oid so‘zlarni saqlab qolish uchun elektron bazaga qo‘shish hisoblanadi. Ham so‘z saqlanib qolinadi, ham leksemalar soni bittaga ko‘payadi. O‘zbek xalq shevalari juda ko‘p va xilma-xil bo‘lib, ularning ayrim qismlarigina fonetik va grammatik jihatdan o‘rganilgan. O‘tgan asrlarda yaratilgan ilmiy asarlar bunga yaqqol misol

bo‘la oladi. Jumladan, F.Abdullayev “O‘zbek tilining Xorazm she’valari”, N.Rajabov “O‘zbek she’vashunosligi”, V.V.Reshetov “Тюркологический сборник” va boshqa olimlarning ilmiy ishlarini aytishimiz mumkin. Shu jumladan, F.Abdullayev Xorazm shevalariga oid namunalarni xalq orasidan, ko‘chalarda yurib jonli she’valardan tuzgan. O‘g‘uz lahjasiga oid Xorazm shevalari ham yana bir qancha mayda sheva guruhlariga bo‘lingan. Bunga turlicha sabablar bo‘lishi mumkin. Ular bir-biridan nafaqat geografik joylashuv jihatidan, balki sheva tomondan ham birmuncha farq qiladi. Demak, shevalar umumiy bir lahja tarkibida birlashsa ham, fonetik, morfologik jihatlaridan turlicha bo‘lishi mumkin. Na’muna qilib Urganch va Xiva hududlari shevalarini olishimiz mumkin. Professor E.D. Polivanov tasniflashicha, “o‘g‘uz lahjasi” o‘z ichiga ikki guruh shevalarini olar ekan:

1. Janubiy Xorazm guruhi shevalari (Xiva, Urganch, G‘azavot, Shovot, Xazorasp, Yangiariq, Xonqa kabi) kiritilgan.

2. Shimoliy o‘g‘uz guruhi shevalari (Iqon-Qorabuloq kabi) kiritilgan ekan.

Yuklamalar Xorazm shevasida tubdan farq qiladi. Misol uchun ***“-da”*** yuklamasi kuchaytiruv-ta’kid ma’nosini bildiradi. Xorazm shevasida ***“don”*** qo‘shimchasiga to‘g‘ri keladi: *Sovg‘a bermaganingdan keyin xafa bo‘ladi-da. Sovg‘a barmaganingdan keyin xofo bo‘ladidon*.

“-ku” qo‘shimchasi ham kuchaytiruv-ta’kid yuklamasi turiga kiradi, Xorazmda bu qo‘shimcha ***“-qu”*** shaklida keladi: *Hamma gaplarimni senga aytdim-ku! Hamma gaplarimi songo oytdm-qu.*

Ko‘makchilarga keladigan bo‘lsak, ular ham Xorazm she’vasida turlicha ifodalanadi: 1) ***“so‘ngra”*** ko‘makchisi vazifadosh ko‘makchilar sirasiga kirib, Xorazm she’vasida ***“o‘sin”*** shaklida keladi: *Tushda ishdan kelaman, so‘ng uyda dam olaman. Abetda ishdan galaman, o‘sin uyda dam olomon.*

Xorazm shevasini korpusini yaratish uchun avvalo ma’lumot yig‘ildi ya’ni 600ta shevaga oid so‘zlar yig‘ib chiqilgan. Ushbu jadvalda to‘rtta tumanda so‘zlarning qanday aytilishi alohida ajratilib ko‘rsatilgan. Korpusda foydalanilgan dataset turli sohalarga doir va turli so‘z turkumlari oid so‘zlardan foydalanib yig‘ilgan.

№	Shevadagi so`z	Adabiy shakl	so'z turkumi	Urganch	Xiva	Qo'shko'pir	Gurlan
1	yimirta	tuxum	ot	tuxum	yimirta	yimirta	tuxum
2	gashir	sabzi	ot	gashir	gashir	gashir	gashir
3	timittin	binoyiday	sifat	timittin	timittin	timittin	timittin
4	parahatchiliq	tinchlik	ot	parahatchiliq	parahatchiliq	parahatchiliq	parahatchiliq
5	navlin	bilmadim	fe'l	navlin	navlin	navlin	navlin
6	zangi	narvon	ot	zangi	zangi	zangi	zangi
7	opo	ona	ot	opo	opo	opo	opes
8	aka	ota	ot	aka	aka	aka	akes
9	ejashdi	hazillashdi	fe'l	ejashdi	ejashdi	ejashdi	ejashdi
10	chapak	qarsak	ot	chapak	chapak	chapak	chapak
11	nerda	qayerda	olmosh	nerda	nerda	nerda	nerda
12	man	men	olmosh	man	man	man	men
13	eshik	eshik	ot	eshik	qopi	qopi	eshik
14	qarinja	chumoli	ot	qarinja	qarinja	qarinja	qarinja
15	pitta	ozgina	ravish	pitta	pitta	pitta	pitta
16	borosonmi	borasanmi	fe'l	borosonmi	borosommi	borosomo	borosomo
17	go'sht	go'sht	ot	go'sht	et	go'sh	go'sht

Yuqoridagi ma’lumotlardan kelib chiqib Pythonda kod yozilgan.

```
print("\tO'zbek shevalari lug'ati(Xorazm shevasi)")
soz =  input("So'z kiriting=").lower()
soz1 =["angak", "aka", "apka", "arka","adik","alashtirma", "arik", "asirak", "barak", "bajik", "bangi",
"biyi",  "bon", "bidirdamoq", "bordi-galdi",  "bo'kchimoq", "bol", "bug'oz", "chapak", "chitmak", "chiya",
"cholvur","chug'ilchi","digirjak", "doyi","dim","do'ddi","duvalang", "dish", "dingan", "dapa",  "domar",
"deshon", "diri", "dobbi", "do'shik", "do'shak", "dik", "elot", "ejashmoq", "guch", "galishmoq", "garak",
"galin","gashir", "gal", "guyov", "guppi", "get","gechchi","gizlanmajak", "go'z", "go'rmoq", "go'shanga",
"gurung", "guldirmomo", "halpa", "horvuz", "hovor", "hurkak",  "ijjon",  "ich qismoq", "ilik", "ino",
"inna","ishu", "indi", "ichon", "jakka-jukka","jinnoqdin","jo'nqi","jo'ro", "jo'gi", "jolli",  "katman",
"ko'pir", "kichchi", "kirpi", "kisa", "kadi", "kat", "loppa at", "lo'ppi", "mongloyi qoro","mo'chchi",
"mijina", "navlin", "novvi","nishatin", "nera", "nerda", "nichik", "ojji", "opo", "oston", "orva",
"ongiyo'q", "pishik", "potya","polos",  "po'm", "qorinja", "qumursqa", "qopoq","qopshirma", "sollanmoq",
```

```
        O'zbek shevalari lug'ati(Xorazm shevasi)
So'z kiriting=barak
Adabiy tarjimasi: chuchvara
So'z turkumi: ot
```

Yuqoridagi kodni ishga tushirganimizdagi Xorazm shevasiga oid so‘z kiritiladi natijada shevadagi so‘zning adabiy tilda qaysi so‘zga teng kelishi va uning qaysi so‘z turkumiga tegishli ekanini chiqarib beradi. Olingan natijalar Xorazm shevasidan o‘zbek adabiy tiliga avtomatik o‘girish mumkinligini isbotladi.

Shunga qaramay, tadqiqot davomida ba’zi cheklovlar ham aniqlangan: shevada mavjud bo‘lgan leksik arxaizmlar va lokal semantik o‘zgarishlar modelni chalkashtirishi mumkin. Xorazm viloyatining o‘zida ham turli shevalar mavjud. Bunga Gurlan va Yangibozor shevalarini misol qilishimiz mumkin. Gurlan tumanining Amudaryo bilan chegaradosh bo‘lgan joylarida ***“j”*** lashishlarga duch kelinadi. Shimoliy Xorazmdan Xiva tomonga qarab ***“j”***lashish kamayib boradi. Bu

shevada ayrim leksikologik o‘ziga xosliklar ham mavjud. Xorazm shevasida asosan olmoshlarning variantlari ko‘p va adabiy tildan tubdan farq qiladi.

Shevalarda ham omonimlikni kuzatishimiz mumkin. Masalan *“kadi”* so‘zini ikki xil ma’nosi bor. Birinchi ma’nosi *“qovoq”* bo‘lsa, ikkinchi ma’nosi *“sevgan qizi(yigiti)”* ma’nosida keladi. Korpusni yaratishda frazeologiyalarni ham kiritish muhim. Adabiy tildagi frazeologiyalarning ma’nosi Xorazm shevasida umuman boshqa ma’noda keladigan holatlarni kuzatishimiz mumkin.

bunga	binga
bunga	minga
unga	vunga
undan	vundan
ana shu	onoshu
mana shu	inashu
ana	ono
mana	ina
menga	monga
senga	songa

Og‘zaki nutqda juda charchamoq ma’nosini bildiruvchi *“Og‘zini suvi oqdi”* iborasi mavjud bo‘lib, bu ibora o‘zbek adabiy tilidagi *“Holdan toymoq”* iborasiga teng keladi. Shuningdek, ushbu iboraning Xorazm shevasida *“Oyoqinnan o‘t chiqdi”* sinonim iborasi ham uchrab turadi. Bu iboraning birinchi iboradan deyarli farqi yo‘q. Misol uchun, *“to‘y o‘yda ish atavarip, oyoqimnon o‘t chiqdi”* yoki “og‘zimni suvi oqdi” birikmasini ham qo‘llash mumkin.

Jumladan, Qo‘shko‘pir aholi shevasi nutqida bu iboralar bilan yana bitta sinonimlik hosil qiluvchi ibora borki, bu ibora *“Go‘zzini go‘ki so‘kildi”* tarzida qo‘llanadi. Bu ibora ham “*qattiq charchadi”* ma’nosini bildiradi. Garchi uning ma’nosi ham yuqoridagi iboralar bilan deyarli bir xil bo‘lsa-da, bu iboraning ahamiyatli jihati ham borki, ibora sal tor ma’noda ishlatilib, faqat ko‘z orqali qilinadigan jismoniy harakatlarga nisbatangina(kompyuterda ishlash, tikuv mashinasida ishlash) ishlatiladi. Misol uchun, *Dinmin moshinda tikavarip go‘zzimni go‘ki so‘kildi* (Qo‘shko‘pir sheva og‘zaki nutqidan).

Shuningdek ushbu shevalarda *“Borg‘ono boldiz, galgana yengga”* kerakli va keraksiz joylarga ham boraveradigan, hamma joyda topiladigan, odamga nisbatan qo‘llaniladi.

“*O‘lmagay bi bolo hammeda bo bi borg‘ono boldiz, galgana yenggachilik atip.”*

Xulosa.

Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga avtomatik tarzda o‘girish dasturini ishlab chiqish bo‘yicha olib borilgan tadqiqot natijalari dialektal korpuslar va tabiiy tilni qayta ishlash sohasida muhim ilmiy va amaliy ahamiyatga ega ekanligini ko‘rsatdi. Yaratilgan dastur dialektal farqlarni sezilarli darajada aniqlay olgan va adabiy til me’yorlariga moslasha olgan. Biroq, yuqori aniqlikka erishish uchun kengroq, sifatli parallel korpuslar va tilga xos xususiyatlarni yanada chuqurroq o‘rganishga ehtiyoj mavjud. Tadqiqot natijalari O‘zbek tilining hududiy variantlarini standartlashtirish, madaniy merosni raqamlashtirish va mashinali tarjima tizimlarini takomillashtirish yo‘lida mustahkam asos yaratdi.

Foydalanilgan adabiyotlar:

1. Elov B.B., Xusainova Z.Y. Katta til modellari / “O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi”. 2024. Vol 2(6). – P. 78-92.
2. Elov B., Abdullayeva O. Dialektal korpusni qurish va uning ahamiyati. // O‘zbek folklori va shevalari tadqiqotlari: amaliyot, metodologiya, yangicha yondashuv.
3. Palvonnazarova H., Shamuratova S. O‘g‘uz shevalarida yordamchi so‘z turkumlarini qo‘llanilishi. Eurasian Journal of Social Sciences, Philosophy and Culture.
4. Ibragimov Y., Ibragimova Zamira. O‘zbek tilining dialektal korpusini yaratish tamoyillari. // O‘zbek Tilshunosligi va Adabiyotshunosligi Taraqqiyoti, Xalqaro Hamkorlik Masalalari.
5. Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari / Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali, October 2023. Volume 1, Issue 3.
6. Xusainova, Z., & Yangibayeva, S. (2024). Til korpusi turlari. / Uzbekistan: Language and Culture, 2(2).
7. Abdullayev F. O‘zbek tilining Xorazm she’valari.
8. Elov, B. B., Hamroyeva, S., & Xusainova, Z. (2022). NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar. / TerDU, filologik tadqiqotlar: til, adabiyot, ta’lim, 5-6.
9. Paquot M., Th.Gries S. Corpus Architecture. A Practical Handbook of Corpus Linguistics.
10. Boberg Ch., Nerbonne J., Watt D. Corpus-Based Approaches to Dialect Study. The Handbook of Dialectology.

III SHO‘BA. TABIIY TILNI QAYTA ISHLASH (NLP)

UDK: 004.912

MATNNI TAHLIL QILISH BO‘YICHA ILG‘OR TAJRIBALAR

Mamatov Narzillo Solidjonovich,
Texnika fanlari doktori, professor
m_narzullo@mail.ru
“Toshkent irrigasiya va qishloq xo‘jaligini mexanizasiyalash muhandislari”Milliy tadqiqot universiteti

Nuritdinov Nurbek Davlataliyevich,
nuritdinovnurbek85@gmail.com
Namangan muhandislik-qurilish instituti, tayanch-doktorant

Annotatsiya: Ushbu maqolada matnlarni tahlil qilishning turli fan sohalarida qo‘llaniladigan asosiy usullari tahlil qilinadi va ularning kompyuterlashtirish darajasi bo‘yicha tasnifi keltiriladi. Tadqiqotda tematik tahlil, kontent tahlili, lug‘at yondashuvlari, so‘zlar sumkasi (BOW) modeli, boshqariladigan va boshqarilmaydigan o‘rganish usullari, shuningdek tabiiy tilni qayta ishlash (NLP) texnikalari qamrab olingan. Har bir yondashuvning nazariy asoslari, amaliy qo‘llanilishi, afzalliklari va cheklovlari muhokama qilinadi. Maqolaning maqsadi – matn tahlilining turli metodologik yondashuvlarini tizimlashtirish va tadqiqotchilar uchun ularning imkoniyatlari hamda chegaralarini aniqlab berishdan iborat. Tadqiqot natijalari matn tahlil usullarini tanlashda asos bo‘lishi, ilmiy, ijtimoiy va texnologik sohalarda matnga asoslangan tadqiqotlar samaradorligini oshirishga xizmat qilishi mumkin.

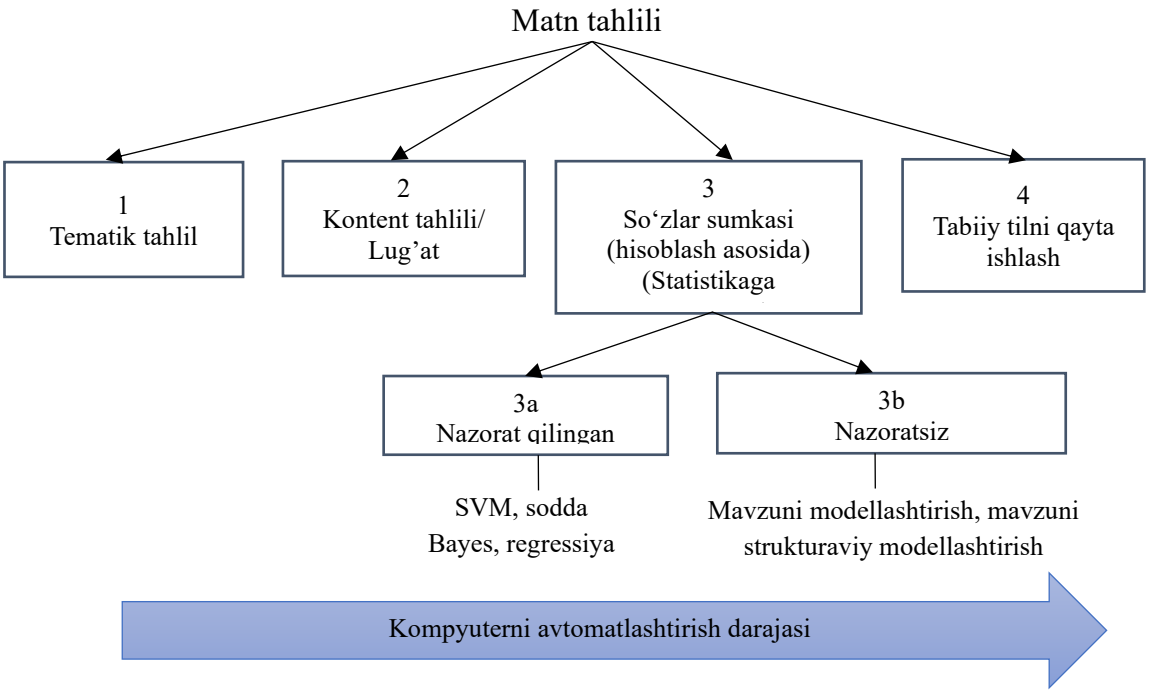
Аннотация: В данной статье рассматриваются основные методы анализа текстов, применяемые в различных научных дисциплинах, и приводится их классификация в зависимости от степени компьютеризации. В исследование включены тематический анализ, контент-анализ, лексические подходы, модель «мешка слов» (BoW), методы контролируемого и неконтролируемого обучения, а также технологии обработки естественного языка (NLP). Обсуждаются теоретические основы, практическое применение, преимущества и ограничения каждого подхода. Цель статьи — систематизировать различные методологические подходы к анализу текста и определить их возможности и ограничения для исследователей. Результаты исследования могут послужить основой для выбора методов анализа текста, а также способствовать повышению эффективности текстоориентированных исследований в научной, социальной и технологической сферах.

Abstract: This article examines the main methods of text analysis used across various academic disciplines and provides their classification based on the degree of

computerization. The study includes thematic analysis, content analysis, lexical approaches, the bag-of-words (BoW) model, supervised and unsupervised learning methods, as well as natural language processing (NLP) technologies. Theoretical foundations, practical applications, advantages, and limitations of each approach are discussed. The aim of the article is to systematize different methodological approaches to text analysis and to identify their capabilities and limitations for researchers. The research findings can serve as a basis for selecting appropriate text analysis methods and contribute to enhancing the effectiveness of text-based research in scientific, social, and technological domains.

Kalit so‘zlar: *Matn tahlili, Tematik tahlil, Kontent tahlili, So‘zlar sumkasi modeli, Tabiiy tilni qayta ishlash, Lug‘atga asoslangan yondashuv, Fanlararo tahlil, Semantik tahlil, Statistik yondashuvlar, Sun’iy intellekt.*

Ushbu maqolada matnlarni tahlil qilishning turli fan sohalarida qo‘llaniladigan turli xil matn tahlil usullarini tizimlashtiradi va muhokama qiladi. Matnni tahlil qilish ko‘p shakllarni oladi, ularning har biri o‘z taxminlari, afzalliklari va cheklovlariga ega[1]. Muayyan fan doirasida qo‘llaniladigan o‘ziga xos texnikalar asosan noma’lum yoki boshqa fanlar tomonidan qo‘llanilmaydi, bu esa har bir texnikaning o‘ziga xos xususiyatlaridan xabardor bo‘lmasligiga olib kelishi mumkin. 1-rasmda kompyuterni avtomatlashtirish darajasi bo‘yicha guruhlangan turli fanlarda qo‘llaniladigan eng mashhur matn tahlil usullarining tasnifi keltirilgan. Ushbu turkumlash tadqiqotchilarga matn ma’lumotlarini tahlil qilish uchun mavjud vositalarni ko‘rish imkonini beruvchi tashkiliy tuzilmani taqdim etadi [2].



1-rasm. Matnni tahlil qilishning umumiy usullarining tasnifi tematik tahlil.

1-rasmning 1-bandiga nazar tashlaydigan bo‘lsak, tematik tahlil odatda asosli nazariya metodologiyalari bilan bog‘liq bo‘lgan matn tahliliga sharhlovchi yondashuvdir. Asoslangan nazariya va tematik tahlilning maqsadi odamlarning o‘z dunyolariga xos bo‘lgan ma’no va ma’nolarga asoslangan nazariyani yaratish va takomillashtirishdir. Binobarin, ma’lumotlar ko‘pincha intervyu stenogrammasi, ishtirokchilar kuzatuvlari qaydlari va arxiv matni, jumladan hujjatlar, veb-saytlar va elektron pochta xabarlari shaklida taqdim etiladi. Tematik tahlil ko‘pincha tashkiliy fanlar va aloqa fanlarida qo‘llaniladi [3].

Tematik tahlil takrorlanuvchi jarayonni o‘z ichiga oladi, bunda tadqiqotchi matndan kelib chiqadigan bir qator kodlar va toifalarni ishlab chiqadi. Umuman olganda, toifalar tahlil boshlanishidan oldin noma’lum, nazariyani takomillashtirish zarur bo‘lgan hollar bundan mustasno; bunday hollarda ma’lumotlarni tahlil qilish adabiyot va ma’lumotlarni doimiy ravishda taqqoslashni o‘z ichiga oladi. Tadqiqotchi ishtirokchilarning o‘z tilidan ("birinchi tartib kodlari" yoki "ochiq kodlash" deb ataladi) boshlanadi va keyin o‘xshash kodlarni toifalarga ("ikkinchi tartib kodlari" yoki "eksenli kodlash" deb ataladi). NVivo va ATLAS kabi kompyuter dasturlari bunday kodlar va toifalarni tashkil qilishni osonlashtirishga yordam beradi, garchi matn toifalari odatda inson kodlashidan olingan toifalarning operasion ta’riflariga bog‘liq bo‘lsada, shuning uchun ko‘pincha kompyuterni avtomatlashtirish darajasi juda kam yoki umuman yo‘q [4].

Garchi tematik tahlil tadqiqotchilarga unchalik ma’lum bo‘lmagan hodisalar va tushunchalarni tushunish imkonini beradi, ammo bu usul bilan bog‘liq bir qator xarajatlar mavjud. Kodlash yarim yoki to‘liq avtomatlashtirilmaganligi va toifalar noma’lum bo‘lganligi sababli matnni o‘qish, turkumlashtirish va talqin qilish ko‘p vaqt talab etadi, bu esa katta hajmdagi ma’lumotlarni (masalan, 100 000 varaq transkript) tahlil qilishni deyarli imkonsiz qiladi. Bundan tashqari, tematik tahlil xatolar va tadqiqotchilarning noto‘g‘riligiga nisbatan zaif bo‘lishi mumkin. Bundan tashqari, tematik tahlil odatda boshqa yondashuvlarga qaraganda qiziqish sohasi haqida ko‘proq bilim talab qiladi.

Kontentni tahlil qilish lug‘atga asoslangan usullar. Kontentni tahlil qilish va boshqa lug‘at usullari (1-rasm, 2-band) ko‘pincha ma’lum bir matndagi so‘zlar yoki iboralar chastotasini hisoblash orqali amalga oshiriladi. SHuning uchun, lug‘at usullari yordamida sifatli ma’lumotlar miqdoriy yo‘naltirilgan tadqiqot savollariga javob berish uchun ishlatilishi mumkin, chunki matn ko‘pincha so‘z chastotalarini hisoblash uchun qisqartiriladi. Oldindan talab qilinadigan inson bilimlari miqdori, shuningdek, kompyuterni avtomatlashtirish, so‘zlar ro‘yxati ma’lumotlardan kelib chiqadimi yoki so‘zlar ro‘yxati nazariya yordamida apriori aniqlanganiga qarab o‘zgaradi. Tematik tahlilga o‘xshab, kompyuter dasturlari kontentni tahlil qilish jarayonida yordam berishi mumkin. DICTION kabi dasturlar turkumlash lug‘atidan foydalangan holda matnni avtomatik ravishda baholaydi (ya’ni, operativ ta’riflar o‘rniga so‘zlar yoki n-grammalarga asoslangan mavzularni aniqlash). NVivo yoki ATLAS kabi boshqa dasturlar tematik tahlilga o‘xshab qo‘llanilishi mumkin, bunda kodlash va toifalash ma’lumotlarni tashkil etish dasturi yordamida qo‘lda amalga oshiriladi [5].

So‘zlar sumkasi va tabiiy tilni qayta ishlash.

Oxirgi ikkita usul - matnni tahlil qilishda so‘zlar sumkasi (BOW; ya’ni sanashga asoslangan; 1-rasm, 3-blok) va tabiiy tilni qayta ishlash (NLP; 1-rasm, 4-blok) yondashuvlari. Ushbu matnni tahlil qilish vositalari kompyuter dasturlari turli matnlarni tasniflash va toifalarga ajratish vazifalarini o‘rganadigan va ularni yarim yoki to‘liq avtomatik ravishda bajarishi mumkin bo‘lgan mashinani o‘rganishni o‘z ichiga olishi mumkin. Sun’iy intellekt inson tajribasi darajasiga (yoki undan ham yaxshiroq) erishishi mumkin. SHu bilan birga, inson tajribasini sun’iy intellekt bilan birlashtirish sinergetik ta’sir ko‘rsatishi mumkin.

BOW yondashuvlari matnni soddalashtirish va qisqartirish uchun asosan informatika fanlarida qo‘llaniladigan texnikalar guruhidir[6]. Bu usullar hujjatdagi so‘zlarning tartibi muhim emasligini nazarda tutadi; ya’ni hujjatlar "so‘zlar sumkasi" sifatida qaraladi. So‘z tartibini e’tiborsiz qoldirish hujjatda so‘zlarning necha marta paydo bo‘lishini hisoblash orqali ma’lumotlarga statistik xususiyatlarni qo‘shadi. Olingan ma’lumotlar strukturasi hujjat atamasi matrisasi deb ataladi, unda

fayl ma’lumotlar matrisasining har bir satri hujjatni, ustunlar esa hujjatda qo‘llanilgan alohida atamalarni ifodalaydi. BOW usullari so‘z tartibini e’tiborsiz qoldirganligi sababli, ular statistik xususiyatlarni (ya’ni, o‘rnini bosish) ta’minlaydi. Natijada, bu usullar hujjatlarni toifalarga bo‘lish kabi makro vazifalarni bajarishda yaxshi va katta hajmdagi ma’lumotlarni qayta ishlashga qodir. Misol uchun, spam-filtrlar ba’zi elektron pochta xabarlarini mazmuniga ko‘ra spam sifatida tasniflash uchun Baes ehtimollik modellariga asoslangan BOW yondashuvidan foydalanadi. Biroq, BOW usullari matnning semantik ma’nosini aniqlash kabi mikro darajadagi vazifalar uchun mos emas.

BOW yondashuvlarining ikki turi mavjud: nazorat ostida (1-rasm, 3a blok) va nazoratsiz
(1-rasm, 3b blok). Ushbu usullar va ularni qo‘llash tadqiqot kontekstiga qarab farq qiladi. Boshqariladigan usullarda tadqiqotchi nimani qidirayotganini oldindan biladi. Aniqrog‘i, tadqiqotchi dasturga kirish ma’lumotlarini (bu holda matn) ham, chiqish ma’lumotlarini ham (masalan, matn muallifini aniqlash) beradi va tizim ular orasidagi munosabatni ko‘rsatish uchun algoritm yaratadi. Tadqiqotchilar Mosteller va Uolles 12 ta munozarali federalistik hujjatlarning (Jeyms Madison yoki Aleksandr Hamilton) muallifligini bashorat qilish uchun oddiy Baescha so‘z ehtimollaridan foydalangan holda ushbu yondashuvning eng dastlabki misollaridan birini taqdim etdilar. Bugungi kunda Naive Baes va Support Vector Machines (SVMs) kabi usullar matn tahlili uchun ishlatiladigan mashhur nazorat ostidagi algoritmlardir [7].

Shu bilan bir qatorda, nazoratsiz algoritmlar tematik tahlilga o‘xshash so‘z klasterlari va ma’lumotlardan kelib chiqadigan mavzularni aniqlaydi. Biroq, mavzu tahlilidan farqli o‘laroq, mavzuni modellashtirish muhim mavzularni aniqlash uchun yuqori darajada (to‘liq bo‘lmasa ham) avtomatlashtirilgan yondashuvdan foydalanadi. Tahlil qilish uchun kamroq vaqt va matn haqida kamroq ma’lumot talab etiladi. Shu sababli, mavzuni modellashtirish katta hajmdagi ma’lumotlarni tahlil qilish uchun javob beradi, garchi inson tushunchasi paydo bo‘lgan mavzularni sharhlash uchun hali ham muhim. Mavzuni modellashtirish mavzuni tahlil qilish (ya’ni, inson tushunchasi) va mashinani o‘rganish (ya’ni, katta hajmdagi matnni tezda tahlil qilish) afzalliklaridan foydalanadi.

Nihoyat, NLP odatda matn tahlilining eng avtomatlashtirilgan shaklidir. Bu usul odamlarning tilni qanday tushunishi va qayta ishlashini modellashtiradi. Masalan, NLP usullari jumladagi so‘zlarning nutq qismlarini (masalan, otlar, sifatlar va boshqalar) belgilash imkonini beradi. Hujjatlarni bir tildan boshqa tilga tarjima qiling va hatto so‘zlarning ma’nosini aniqlashtirish uchun jumla kontekstidan foydalaning. Shuning uchun, BOW yondashuvidan farqli o‘laroq, NLP so‘z tartibini muhim deb hisoblaydi. CHuqur o‘rganish va multimodallik (ya’ni, matn va tasvirlarni birlashtirish) kabi ilg‘or usullardan foydalangan holda his-tuyg‘ularni

tahlil qilish o‘quv to‘plamlaridan foydalanishda NLPning mashhur shakllaridan biridir. Ushbu maxsus tahlil matnga nisbatan umumiy munosabat, hissiyot yoki fikrni ijobiy, salbiy yoki neytral deb tasniflaydi. Mavzuni tahlil qilishdan farqli o‘laroq, NLP to‘liq avtomatlashtirilgan jarayon bo‘lib, shuning uchun inson tushunishi va talqinini kam yoki umuman talab qilmaydi. Bundan tashqari, insonni kodlashni talab qiladigan usullar (masalan, tematik tahlil) bilan solishtirganda, NLP boshqa yondashuvlarga qaraganda ancha tez va tizimli. Masalan, kompyuter fanlari, informatika, tilshunoslik va psixologiya sohasidagi tadqiqotchilar NLP dan matnni qazib olish vositasi sifatida foydalanadilar [8].

Tekshirish, baholash va izohlash uchun talab qilinadigan mavzuni modellashtirish bosqichlari [9]

1-bosqich: Gipoteza va savollarni shakllantirish

(a) Gipotezalar tadqiqot muammolari. Mavzuni modellashtirishda gipotezalar va tadqiqot muammolari boshqa har qanday tadqiqotga o‘xshash ishlab chiqiladi.

2-qadam: Dizayn va ma’lumotlarni yig‘ish

(a) ma’lumotlar turi. Ma’lumotlar turlariga misollar orasida ochiq so‘rov javoblari va arxiv ma’lumotlari, masalan, veb-sayt matni, bosh direktorlarning xatlari, hamkasblar o‘rtasidagi elektron pochta xabarlari va ijtimoiy media (masalan, Twitter, Facebook) kiradi.

(b) Namuna hajmi. Namuna hajmi farazlar va tadqiqot savollarining nazariy va uslubiy mulohazalari bilan bog‘liq. Masalan, nazariya ma’lum bir qiziqish populyasiyasini ko‘rsatishi mumkin yoki korrelyasion tahlil o‘tkazish istagi namuna hajmini aniqlash uchun quvvat tahlilini talab qilishi mumkin. Tadqiqotchilar hujjat darajasidagi tahlilni ham ko‘rib chiqishlari kerak.

(c) yozish sifati. Yaxshiroq yozish odatda tahlilda yordam beradi. Biroq, hatto sifatsiz matnni ham tahlil qilish mumkin (masalan, Twitter ma’lumotlari).

(d) javoblar davomiyligi. Ko‘proq matn odatda kamroqdan yaxshiroqdir. Minimal so‘zlar soni bo‘lmasada, har bir hujjat uchun kamida 200 belgi tavsiya etiladi.

(e) kovariasiyalar. Turli mavzular paydo bo‘ladimi yoki quyi toifaga qarab har xil muhokama qilinadimi yoki yo‘qligini tushunish uchun "kovariatlar" sifatida sinovdan o‘tkazilishi mumkin bo‘lgan nazariy va uslubiy moderatorlarni hisobga olish kerak.

3-qadam: Oldindan davolash

(a) noto‘g‘ri yozuvlar. Minimal so‘zlar soniga to‘g‘ri kelmaydigan yoki tegishli matnni o‘z ichiga olmagan javoblar kabi noto‘g‘ri yozuvlarni o‘chirish haqida o‘ylashingiz mumkin.

(b) tokenizasiya. Tokenizasiya bosqichi jumlalarni alohida so‘zlarga yoki "tokenlarga" qisqartirishni o‘z ichiga oladi.

(c) tozalash. Tozalash bosqichi kichik harf belgilarini yaratish, bo‘shliqlar va tinish belgilarini olib tashlashni o‘z ichiga oladi.

(d) so‘zlarni to‘xtating. To‘xtash so‘zlarini olib tashlash tavsiya etiladi. To‘xtash so‘zlarini tadqiqot kontekstida juda tez-tez uchraydigan so‘zlar deb hisoblash mumkin, ular mavzularni aniqlashda qo‘shimcha qiymat bermaydi. Tasodifiy shovqin bo‘lishi mumkin bo‘lgan tez-tez uchraydigan so‘zlarni olib tashlash tavsiya etiladi.

(e) kamdan-kam hollarda. Hujjatda kamdan-kam uchraydigan atama bo‘lishiga yo‘l qo‘ymaslik uchun so‘zning minimal sonini belgilashingiz kerak bo‘lishi mumkin. Mavzuni bir necha yuzlab hujjatlarni modellashtirishda, agar natijalar shovqinli bo‘lsa yoki tahlil juda uzoq davom esa, kamida ikkita hujjatdan boshlash va ularni mos ravishda moslashtirish foydali bo‘ladi. Namuna minglab hujjatlardan iborat bo‘lgan hollarda, beshta hujjat kabi yuqori minimumdan boshlash yaxshidir. Ko‘pgina hollarda shovqinni kamaytirish va hisoblash tezligini oshirish uchun kamida beshta hujjat qo‘llaniladi.

(f) Stemming lemmatizasiya. Tahlilni osonlashtirish uchun stemming yoki lemmatizasiya qilish kerak. Masalan, “Imkoniyat” va “imkoniyat”lar “imkoniyat” bilan almashtiriladi. Ushbu bosqichdan oldin bir nechta dastlabki tahlillarni o‘tkazish tavsiya etiladi. Stemming keyingi tahlil uchun tavsiya etiladi, ammo tadqiqotchilar stemming chiqarilgan xulosalarga ta’sir qiladimi yoki yo‘qligini tekshirishlari kerak.

(g) Uni-/bi-/trigramlar. Bigramlardan foydalanish odatda so‘zlar sumkasini yengib o‘tishga yordam berib, tahlil qilishda yordam beradi. Misol uchun, faqat "SHimoliy" va "Karolina" o‘rniga "SHimoliy Karolina" dan foydalanishingiz mumkin. Biroq, uni, bi yoki trigramlardan foydalanganda topilmalar o‘zgarishini aniqlash uchun bu taxminni har bir tahlilda sinab ko‘rish mumkin.

4-qadam: Mavzuni modellashtirish

a) tez-tez ishlatiladigan so‘zlar. Matndagi "muhim" va tez-tez uchraydigan so‘zlar takrorlanuvchi jarayon orqali aniqlanishi kerak, bunda to‘xtash so‘zlarini qo‘shish va olib tashlash va bunday qadamlar paydo bo‘ladigan so‘z turlarini o‘zgartirish yoki o‘zgartirishni baholash uchun boshqa dastlabki ishlov berish bosqichlari.

(b) Mavzular soni. 1 dan 100 tagacha mavzuni ko‘rib chiqish tavsiya etiladi; paydo bo‘ladigan mavzular soni oxir-oqibatda mavzularning talqin qilinishi va parsimonlik zarurati bilan ta’sir qilishi kerak. Bir qator mavzular bo‘yicha tadqiqotlar iterativ tarzda amalga oshirilishi kerak.

(v) tarmoq strukturasini o‘rganish. Mavzular tarmog‘i muayyan mavzular qanchalik bog‘liqligini ko‘rish imkonini beradi. Bu paydo bo‘lgan konstruksiyalarning o‘lchovliligini baholash uchun foydalidir. Mavzu tarmog‘ini xaritalashda minimal korrelyasiya 0,10 dan boshlash tavsiya etiladi. Biroq, chegarani, ayniqsa mavzular soniga qarab sozlash kerak bo‘ladi (ko‘proq mavzular pastroq qiymatni talab qiladi va aksincha).

(d) ish ta’rifi. Mavjud adabiyotlardan konstruktiv ta’riflarni aniqlash yoki so‘zlar ro‘yxati uchun so‘zlarni tanlashni osonlashtirish uchun ishchi ta’rifni ishlab chiqish kerak.

Matndagi xatolarni tuzatish algoritmlari

Tabiiy tilni qayta ishlash (NLP) algoritmlari inson tilidagi ma’lumotlarni, shu jumladan tuzilmagan matn ma’lumotlarini qayta ishlash uchun ishlatiladi. Bugungi kunda NLP algoritmlari til qoidalari, statistik yondashuvlar va sun’iy intellekt yondashuvlari asosida ishlab chiqilgan. Til qoidalariga asoslangan yondashuv asosida NLP vazifalari va til korpusidagi tasniflash operasiyalarining lingvistik asoslarini shakllantirish amalga oshiriladi. Statistik algoritmlar mashinalarga inson tillarini o‘qish, tushunish va ma’no chiqarish imkonini beradi va katta hajmdagi matnlarni (katta ma’lumotlarni) qayta ishlashga asoslangan. Statistik algoritmlar nutqni aniqlash, mashina tarjimasi, hissiyotlarni tahlil qilish, tasniflash va matnni qazib olish kabi ko‘plab NLP vazifalarida qo‘llaniladi. Bugungi kunda CNN va RNN texnologiyalariga asoslangan mashinani o‘rganish (ML) algoritmlarini chuqur o‘rganish modellari mavjud NLP tizimlarini o‘rganishi va katta hajmdagi tuzilmagan matnni aniqroq qayta ishlashni ta’minlashi mumkin.

Bugungi kunda ma’lumotlarni qazib olish tabiiy tilni qayta ishlashning mashhur vazifalaridan biridir. Matnning semantikasini (mazmunini) tushunish, uni matn sifatida ifodalash va qayta ishlash uchun tasniflash, qismlarga ajratish kabi harakatlar bajarilishi kerak.

Kompyuterlar va odamlarga tabiiy til yordamida ma’lumot almashish imkonini beruvchi mexanizmlarni ishlab chiqadilar. Kompyuterlar NLP tufayli inson tilini o‘qishi, sharhlashi, tushunishi va javob berishi mumkin, ishlov berish odatda mashinaning aql darajasiga bog‘liq. NLP algoritmlari odamning xabarlarini ular uchun mazmunli raqamli ma’lumotlarga aylantiradi. Jarayonlar va ularning algoritmlari qanday ishlashini tushunish muhimdir. Zamon bilan hamnafas bo‘lish, bu texnologiyalar imkoniyatlaridan unumli foydalanish zarur.

NLPda hal qilinayotgan muammolar orasida eng muhimlari quyidagilardir (1-rasm) [10],[11]:

- NLP texnologiyalarini ishlab chiquvchilar uchun belgilangan birinchi klassik vazifa;

- grammatika va imloni tekshirish (grammatika va afsun checking) – o‘zbekcha matnlarning imlosini avtomatik tekshirish va tuzatish;
- matn tasnifi (matn klassifikatsiyasi) – matn semantikasini aniqlash;
- ma’lum ob’ektlarni identifikasiya qilish (nomli ob’ekt recognition, NER) - muhim ob’ektlarni aniqlash;
- umumlashtirish - matnni soddalashtirilgan shaklda umumlashtirish;
- matn yaratish (matn avlod) AI tizimlarini yaratishda foydalaniladigan vazifalardan biridir;
- mavzuni modellashtirish (mavzuni modellashtirish) - katta matnlardan yashirin mavzularni ajratib olish.

2-rasm. NLP vazifalarining murakkabligi[12]

Shuni ta’kidlash kerakki, barcha zamonaviy va tegishli NLP vazifalari ko‘pincha interaktiv AI tizimlarini yaratishga to‘g‘ri keladi: chatbotlar. Bu inson so‘rovlarini dasturiy ta’minot bilan birlashtirishga yordam beradigan muhit (tizim).

Bundan tashqari, mobil qurilmalarda matn xatolarini tuzatish jarayonlari paydo bo‘ladi. Bunday holda, foydalanuvchi kursorni xato joyiga ko‘chirish orqali xatoni tuzatishi mumkin. Bunday jarayonlar ko‘p matnli so‘zlarga sarflanadigan vaqtni ko‘paytirishga olib keladi. Tekshiruv jarayonini yaxshilashga yordam beradigan uchta yangi matnni tuzatish usuli ko‘rib chiqiladi: Drag-n-Drop, Drag-n-Throw va Magic Key.

- **Drag-n-Drop**. Bu usul foydalanuvchiga kursorni xato ustiga olib borish va uni tuzatish imkonini beradi.
- **Drag-n-Throw.** Ushbu usul foydalanuvchiga xato matnining dastlabki ro‘yxatini klaviatura so‘rovlari ro‘yxatidan olish imkonini beradi.
- **Sehrli kalit.** Ushbu usul foydalanuvchiga tuzatish kiritish imkonini beradi va neyron tarmoq tomonidan aniqlangan xatolar ro‘yxatini ko‘rsatadi.

Mobil qurilma foydalanuvchilarga yozgan matnni xato holatiga qaytarish uchun Backspace tugmachasini surish imkonini berdi va xatoni tuzatgandan so‘ng,

Backspace tugmasini yana bosish orqali ushbu matnni tiklang. Xato joylarini aniqlash uchun ushbu usul lug‘atdagi matn va so‘zlarni tahrirlash masofasini solishtirish uchun eugene Mers algoritmidan foydalanishi mumkin. Xatolarni aniqlash algoritmi bu ishning asosiy cheklovidir: u faqat imlo xatolarini aniqlaydi. U grammatik xatolar yoki so‘zlarning noto‘g‘ri ishlatilishini aniqlamasligi mumkin.

Xatolarni tuzatish uchun NLP algoritmlari

Drag-n-Throw va Magic Key qo‘llaniladigan tuzatishlar asosida yuzaga kelishi mumkin bo‘lgan xatolarni topish uchun tabiiy tilni qayta ishlash (NLP) neyron tarmoqlaridan foydalanadi. SHuning uchun biz tegishli usullar haqida qisqacha ma’lumot beramiz. An’anaviy xatolarni tuzatish algoritmlari tuzatish bo‘yicha tavsiyalar berish uchun N-gramm masi va qatorni tahrirlash masofasidan foydalanadi. Asl satrdagi har bir so‘z uchun avvalo lug‘atdan mosliklarni qidiradi va har bir mumkin bo‘lgan moslikni N-grammli ma’lumotlar to‘plamidagi chastotasi va qatorni moslashtirish algoritmi asosida almashtiradi. Eng yuqori matn tuzatish sifatida tavsiya etiladi. So‘nggi paytlarda chuqur neyron tarmoqlari NLP tadqiqotlarida ularning umumlashtirilishi va an’anaviy algoritmlarga qaraganda ancha yaxshi ishlashi tufayli mashhurlikka erishdi. Konvolyusion neyron tarmoqlari (CNN) va takroriy neyron tarmoqlari (RNN) NLP vazifalari uchun keng qo‘llaniladi va ko‘pincha kodlovchi-dekoder deb ataladigan tuzilishga amal qiladi, bu yerda modelning bir qismi kirish matnini xususiyat vektoriga kodlaydi va keyin vektorni dekodlaydi.

Foydalanilgan adabiyotlar:

1. Mamatov, N. S., Turg‘unova, N. M., Turg‘unov, B. X., & Samijonov, B. N. (2025). Algorithm for analysis of texts in Uzbek language. In Artificial Intelligence and Information Technologies (pp. 511-515). CRC Press.
2. Cassia Valentini-Botinhao, Xin Vang, Shinji Takaki, Junichi Yamagishi. Speech enhancement for a noise-tolerant text-to-speech synthesis system using deep recurrent neural networks [Onlayn]
3. Agiomyrgiannakis, Rob Klark, Rif A. Saurous. Tacotron: towards end-to-end speech synthesis [Onlayn].
4. Vey Ping, Kainan Peng, Endryu Gibianskiy, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonatan Rayman, Jon Miller. Deep Voice 3: Convolutional Sequence Learning (Onlayn) with text-to-speech conversion.
5. Jonatan Shen, Ruoming Pang, Ron J. Vayss, Mayk Shuster, Navdip Jaytli, Zongheng Yang, Zhifeng Chen, Yu Chjan, Yuxuan Vang, RJ Skerri-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, Yongxui Vu. Natural TTS Synthesis by WaveNet Conditioning on Mel Spectrogram Projections [Onlayn].

6. Quinn, K. M., Monro, B. L., Colaresi, M., Crespin, M. H. va Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. American Journal of Politology, 54, 209–228.

7. Niyozmatova, N. A., Mamatov, N. S., Dusonov, X. T., Samijonov, B. N., & Samijonov, A. N. MFCC-GMM Method for Speaker Identification by Voice.

8. Baumer, E. P., Mimno, D., Guha, S., Quan, E. va Gay, G. K. (2017). Comparing Grounded Theory and Subject Modeling: Extreme Divergence or Unlikely Convergence? Journal of the Association for Information Science and Technology, 68, 1397–1410.

9. Strauss, A. va Corbin, J. (1998). Foundations of qualitative research: Techniques and procedures for developing grounded theory (2nd ed.). A Thousand Oaks: Sage.

10. Reynard, J. C. (2008). Introduction to Communication Studies (4th ed.). Boston: McGraw-Hill.

11. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34, 1–47.

12. Zilola Xusainova Yuldashevna. NLPning zamonaviy algoritmlari va konsepsiyalari // журнал искусство слова №1 (2023) https://doi.org/10.5281/7679911

SINTAKTIK PARSING TURLARI

Elov Botir Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Abdullayeva Oqila Xolmo‘minovna,
doktorant DSc
abdullayevaoqila@gmail.com
ToshDO‘TAU

Annotatsiya. NLPning muhim yo‘nalishlaridan biri matnlarni sintaktik tahlil qilishdir. So‘nggi tadqiqotlar natijasida tillarda sintaktik tahlilni amalga oshiruvchi parser dasturlar tuzildi. Gap darajasidagi sintaktik ma’lumotlarni ifodalash uchun ko‘plab murakkab qoidalarni tahlil qilish kerak. Ushbu maqolada sintaktik tahlil turlari ichida keng tarqalgan dependency (tobelik munosabati tahlili) va constituency (tarkibiy tuzilma tahlili) parsing turlari muhokama qilindi. Kontekstdan mustaqil grammatika (CFG), ya’ni constituency tahlil (yoki birikma tuzilmasi asosidagi tahlil), gapning sintaktik tuzilmasidagi iyerarxik so‘z birikmasini treebank orqali tasvirlashga asoslanadi va bu yondashuv ilk tadqiqot obyekti bo‘lib kelmoqda. Bog‘liqlik (yoki dependency) grammatikasi esa sintaktik va semantik tahlil uchun keng qo‘llaniladigan yana bir mashhur model bo‘lib, unda so‘zlar bevosita bog‘liqlik munosabatlari orqali birikadi va ushbu bog‘liqliklarga ular orasidagi sintaktik yoki semantik munosabatlarni ko‘rsatuvchi teglar belgilanadi. Har ikkala parsing turi tillarda gaplarni sintaktik tahliliga tatbiq qilinib, korpuslarda qo‘llanilmoqda, lekin resurs va grammatik tuzilishiga ko‘ra olinadigan natijalar farqliligi kuzatilmoqda. Maqolada parsing turlarining grammatik xususiyatlari o‘zaro taqqoslandi.

Abstract. One of the important areas of NLP is the syntactic analysis of texts. As a result of recent research, parser programs that perform syntactic analysis in languages have been created. To express syntactic information at the sentence level, it is necessary to analyze many complex rules. This article discusses the most common types of syntactic analysis: dependency (dependence relation analysis) and constituency (structural analysis) parsing. Context free grammar (CFG), that is, constituency analysis (or analysis based on the structure of a phrase), is based on describing a hierarchical phrase in the syntactic structure of a sentence through a tree bank, and this approach is the first object of research. The grammar of dependence (or dependency) is another popular model widely used for syntactic and semantic analysis, in which words are connected through direct dependency relations, and tags are assigned to these connections, indicating the syntactic or semantic relations between them. Both types of parsing are applied to the syntactic analysis of sentences in languages and are used in corpora, but there is a difference in the results

obtained according to the resource and grammatical structure. In the article, the grammatical features of the types of parsing were compared.

Аннотация. Одним из важных направлений НЛП является синтаксический анализ текстов. В результате последних исследований были созданы программы-параметры, выполняющие синтаксический анализ в языках. Для выражения синтаксической информации на уровне предложения необходимо проанализировать многие сложные правила. В данной статье обсуждаются наиболее распространенные типы синтаксического анализа: анализ зависимости (анализ зависимости отношений) и синтаксический анализ (структурный анализ). Контекст-свободная грамматика (KFG), то есть анализ составляющих (или анализ, основанный на структуре словосочетания), основан на описании иерархического словосочетания в синтаксической структуре предложения через банк деревьев, и этот подход является первым объектом исследования. Грамматика зависимости (или зависимости) - это еще одна популярная модель, широко используемая для синтаксического и семантического анализа, в которой слова связаны посредством прямых зависимых отношений, и этим связям присваиваются теги, указывающие на синтаксические или семантические отношения между ними. Оба типа синтаксического анализа применяются к синтаксическому анализу предложений в языках и используются в корпусах, но есть разница в результатах, полученных в соответствии с ресурсом и грамматической структурой. В статье сравнивались грамматические особенности типов синтаксиса.

Kalit so‘zlar: *sintaktik parsing, parser, dependency parsing, constituency parsing, teg.*

NLPda parsing – matnning grammatik tuzilishini, gap bo‘laklarining vazifasini va matn tarkibidagi so‘zlar o‘rtasidagi munosabatlarni aniqlash uchun tahlil qilish jarayoni hisoblanadi. Tilshunoslik nuqtayi nazaridan qaralganda, tahlilning asosiy vazifasi so‘zlarning qanday birikib gap hosil qilishi va gap tuzilishini belgilovchi qoidalarni tahlil qilish bo‘lsa, boshqa tomondan, NLP amaliyotlari nuqtayi nazaridan tahlil mashina tarjimasi, savol-javob tizimlari, axborot ajratib olish, sentiment tahlil va matn yaratish kabi ko‘plab vazifalar uchun muhim hisoblanadi.

Har bir tabiiy tilning asosiy tuzilishi mavjud bo‘lib, odatda bu “grammatika” deb ataladi. Grammatika o‘z navbatida morfologiya va sintaksis kabi bo‘limlarga ajratilib o‘rganiladi. Sintaksisning asosiy g‘oyasi shundan iboratki, so‘zlar birlashib, so‘zlar guruhi yoki so‘z birikmasi kabi tarkibiy qismlarni hosil qiladi va ular bir

butun birlik sifatida harakat qiladi. Ushbu tarkibiy qismlar o‘zaro birlashib, yirikroq birliklarni va oxir-oqibatda gaplarni hosil qilishi mumkin. Tahlil jarayoni ko‘pincha maxsus grammatik qoidalarga tayangan holda amalga oshiriladi, bu grammatik qonunlar sintaktik tuzilmalarni aniq shakllantirishga xizmat qiladi.

Sintaktik parsing turlari bo‘yicha amalga oshirilgan tadqiqot natijalarini tahlil qilganimizda, dependency parsing, constituency parsing, kombinator kategoriyali grammatika (CCG), bosh so‘zga asoslangan birikma tuzilmasi grammatikasi (HPSG), leksik-funksional grammatika (LFG), mavhum ma’no tasviri (AMR), minimal rekursiv semantika (MRS), universal tushunchaviy-kognitiv annotatsiya (UCCA) hamda mantiqiy tahlilga yo‘naltirilgan bir nechta tahlil turlarini o‘rganish mumkin [1]. Ushbu grammatik yo‘nalishlarning barchasi so‘nggi yillarda tadqiq qilinib, ulardan ayrimlari kuchli neyron tarmoqlar va kontekstual oldindan o‘rgatilgan so‘z tasvirlarining (representations) paydo bo‘lishi sababli tez sur’atlar bilan rivojlanmoqda. Mazkur tadqiqotimizda faqatgina dependency va constituency parsing masalasiga to‘xtalamiz.

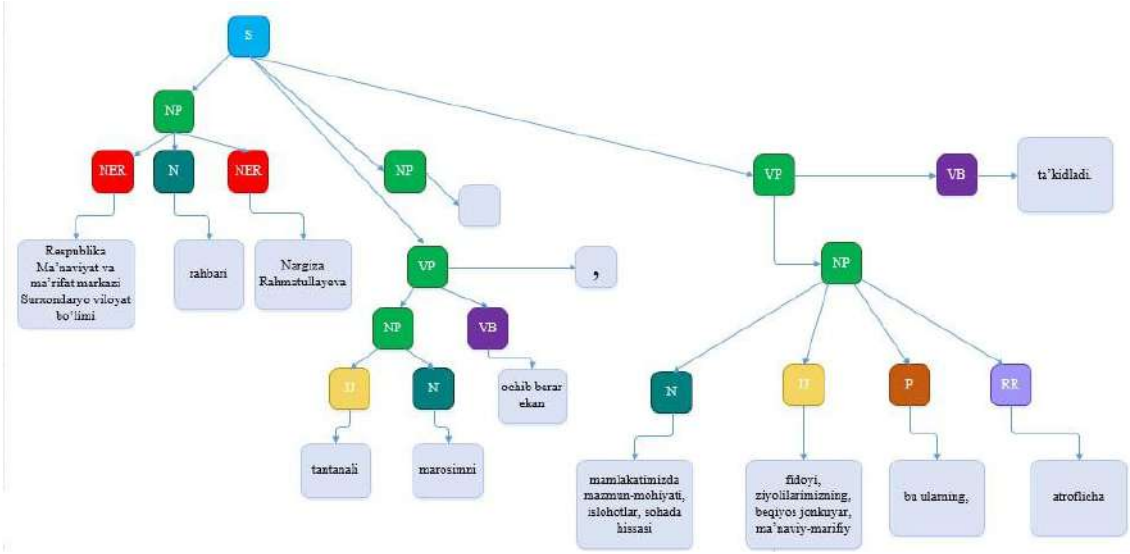
Dependency parsing - bu tilshunoslik va tabiiy tilni qayta ishlash sohasidagi muhim tadqiqot mavzusidir, chunki u so‘zlar orasidagi grammatik bog‘lanishlarni aniqlashga yordam beradi. Bu yondashuvning rivoji, tilshunoslikni chuqurroq tushunishga imkon beradi va tabiiy tilni qayta ishlashda (NLP) samarali foydalaniladi. Ilgari ilmiy ishlar qaramlikni yoki bog‘liqlikni aniq tahlil qilishda grammatika qoidalariga asoslangan metodlarga tayangan bo‘lsa, hozirgi kunda statistik va sun’iy intellekt asosidagi usullar va modellar ko‘plab tillarda yuqori samarali natijalar ko‘rsatmoqda. Ayniqsa, transformer arxitekturalari (masalan, BERT va GPT) tilning sintaktik va semantik strukturalarini tushunishda katta yordam beradi.

Konstituent tahlil (constituency parsing) sintaktik tahlilning asosiy vazifalaridan biri hisoblanadi. 1-rasmda konstituent daraxtining namunasi ko‘rsatilgan bo‘lib, bu yerda tugunlar konstituent segmentlari, ya’ni frazalar sifatida namoyon bo‘lgan. Konstituent tahlilning asosiy maqsadi ushbu frazalarni belgilash va ularning o‘zaro munosabatlarini aniqlashdan iboratdir. Bu parsing turi bo‘yicha amalga oshirilgan tadqiqotlarni tahlil qilganimizda, gap tuzilishida qat’iy so‘z tartibi mavjud tillarda ko‘proq mavjudligini kuzatish mumkin. Erkin so‘z tartibli tuzilmalarga ega tillarda frazalarni ajratish muammolidir. Chunki turkiy tillarda erkin so‘z tartibi mavjud (SOV asosiy bo‘lsa-da), bu esa fraza tuzilmasini qat’iy daraxtlarda ifodalashni qiyinlashtiradi. Dependency tahlil ko‘proq mos keladi: chunki u har bir so‘z orasidagi bog‘lanishni ko‘rsatadi va agglyutinativ tillar uchun ancha samaraliroq hisoblanadi. Shu bois, Universal Dependencies (UD) loyihasi ham turkiy tillar uchun aynan dependency parsing asosida korpuslar yaratmoqda. Quyidagi gapni konstituent parse daraxtida tahlil qilamiz: *Respublika Ma’naviyat va ma’rifat markazi Surxondaryo viloyat bo‘limi rahbari Nargiza Rahmatullayeva*

tantanali marosimni ochib berar ekan, mamlakatimizda ma’naviy-ma’rifiy sohada amalga oshirilayotgan islohotlar, ularning mazmun-mohiyati haqida atroflicha to‘xtalib, bu borada jonkuyar, fidoyi ziyolilarimizning hissasi beqiyos ekanini ta’kidladi.

O‘zbek tili morfoanalizatorida (https://uznatcorpara.uz/) POS teglangan:

Respublika Ma’naviyat va ma’rifat markazi[[NER]] Surxondaryo viloyat bo‘limi[[NER]] rahbari[[N]] Nargiza Rahmatullayeva[[NER]] tantanali[[JJ]] marosimni[[N]] ochib berar[[VB]] ekan[[VB]], mamlakatimizda[[N]] ma’naviy-ma’rifiy[[JJ]] sohada[[N]] amalga oshirilayotgan[[IB]] islohotlar[[N]], ularning[[P]] mazmun-mohiyati[[N]] haqida[[II]] atroflicha[[RR]] to‘xtalib[[VB]], bu[[P]] borada[[II]] jonkuyar[[JJ]], fidoyi[[JJ]] ziyolilarimizning[[JJ]] hissasi[[N]] beqiyos[[JJ]] ekanini[[VB]] ta’kidladi[[VB]].



1-rasm. Constituent daraxti namunasi

Parse daraxti misolida ko‘rish mumkinki, gap tuzilmasini tarkibiy qismlarga ajratib tahlil qilish murakkablik tug‘diradi.

Bu sintaktik parsing turlari yondashuviga ko‘ra turli grammatik munosabatlarni aniqlashga xizmat qiladi [4].

Constituency grammar (tarkibiy tuzilma grammatikasi) - Constituency grammatika, shuningdek, birikma strukturasi grammatikasi deb ham ataladi. Buning sababi, u tarkibiy birikma munosabatiga asoslanadi. Bu bog‘lanish grammatikasining qarama-qarshi turidir.

Constituency grammatikada tarkiblar har qanday so‘z, so‘zlar guruhi yoki birikmalar bo‘lishi mumkin. Constituency grammatikadagi asosiy maqsad – har qanday gapni uning tarkiblariga ajratish va ularning xususiyatlari asosida tuzishdir.

Tarkibiy tuzilma grammatikasi va tarkib munosabatining xususiyatlari:

– barcha tegishli ramkalar gap tuzilishini tarkibiy birikma munosabati orqali ko‘radi;

– tarkibiy munosabatni hosil qilish uchun biz lotin va yunon grammatikalaridagi subyekt-predikat bo‘linishini qo‘llaymiz;

– tarkibiy tuzilma grammatikasida gap tuzilishini ot iborasi (NP) va fe’l iborasi (VP) nuqtayi nazaridan o‘rganamiz. Constituency grammatika har qanday gapni uchta tarkibga ajratishi mumkin: subyekt, predikat va obyekt.

– Tarkibiy tuzilma grammatikasi kontekstsiz va dependency grammatikalarining cheklovlarini hal qilishda yaxshiroq ishlaydi:

– bu yondashuv usuli tilga xos emas, bu esa bir xil modelni bir nechta tillarda ishlatishni yoki tillar o‘rtasida almashishni osonlashtiradi, shu bilan boshqa ikki grammatik turdagi ko‘p tillilik muammosini hal qiladi;

– bu tahlil turida tahlil daraxtini ishlatganligi sababli gapning tarkiblari o‘rtasidagi iyerarxik munosabatni ifodalashni osonlashtiradi va bu boshqa tasvirlash grammatikalariga nisbatan ko‘proq tushunarli va intuitivdir;

– tarkibiy tuzilma grammatikasi boshqa usullarga nisbatan soddaroq va amalga oshirish osonroq bo‘lib, masalan, bog‘lanish grammatikasi kabi formalizmga qaraganda tadqiqotchilar va amaliyotchilar uchun qulayroqdir;

– bu tahlil usuli xatolarga chidamli va shovqinli yoki to‘liq bo‘lmagan ma’lumotlarni boshqarishda yaxshi ishlaydi, shuningdek, so‘z birikmalari yoki jumlalarni bog‘lashda qo‘llaniladigan birlashtiruvchi vosita sifatida koordinatsiyani boshqarishda yaxshiroq ishlaydi.

Dependency (Bog‘liqlik) grammatikasi. Bog‘liqlik grammatikasi Constituency grammatikaning qarama-qarshi turidir va bog‘lanish munosabatiga asoslanadi. U Constituency grammatikaga qarama-qarshi, chunki u birikma tugunlarini o‘z ichiga olmaydi. Bog‘lanish grammatikasi gapdagi so‘zlar bir-biriga bog‘lanish munosabatini tahlil qiladi. Bu so‘zlar bog‘lanish grammatikasida yo‘naltirilgan havolalar orqali bir-biriga ulanadi. Kesim gap tuzilmasining markazi sifatida qaraladi. Bog‘lanish grammatikasi gapdagi so‘zlarni ularning bog‘lanishlariga qarab tartiblaydi. Har bir sintaktik birlik kesimga nisbatan yo‘naltirilgan belgilash orqali bog‘lanadi. Bu sintaktik birliklar **bog‘lanishlar** deb ataladi. Gapdagi so‘zlardan biri ildiz sifatida ishlaydi va boshqa barcha so‘zlar, o‘sha so‘zni hisobga olmaganda, to‘g‘ridan-to‘g‘ri yoki bilvosita ildiz bilan o‘z bog‘lanishlari orqali bog‘lanadi. Ushbu bog‘lanishlar gapdagi so‘zlar o‘rtasidagi munosabatlarni ifodalaydi va dependency grammatikasi so‘zlar o‘rtasidagi struktura va semantik bog‘lanishlarni aniqlash uchun ishlatiladi.

Bog‘lanish grammatikasining ba’zi cheklovlari mavjud, ularni quyida tushunib chiqamiz:

- ma’lumotlarni annotatsiya qilish: Bog‘lanish tahlilini o‘rganish modelini o‘qitish uchun belgilangan ma’lumotlar kerak bo‘ladi, bu esa vaqtni talab qiladi va olish qiyin;
- uzoq masofadagi bog‘lanishlarni boshqarish: Ba’zi hollarda, gapdagi so‘zlar o‘rtasidagi bog‘lanishlar juda uzoqda bo‘lishi mumkin, bu esa gapning grammatik tuzilishini aniq olishni qiyinlashtiradi;
- ellipsis va koordinatsiyani boshqarish: Bog‘lanish grammatikasi, shuningdek, so‘zlar o‘rtasidagi to‘g‘ridan-to‘g‘ri bog‘lanishlar bilan tushuntirilmagan hodisalarni, masalan, ellipsis va koordinatsiyani boshqarishda qiyinchiliklarga duch keladi, bu hodisalar odatda Constituency grammatika orqali ko‘rib chiqiladi.

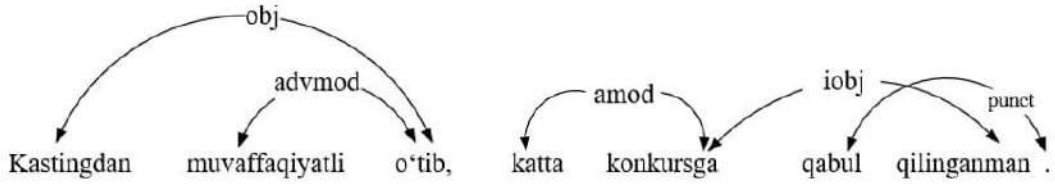
Dependency grammatikasining cheklovlarini yengish uchun Constituency grammatika qo‘llanilishi mumkin, bu esa kamroq kuchli bo‘lsa-da, lekin ko‘proq tushunarli va amalga oshirish osonroqdir. Shuningdek, Constituency va Dependency grammatikalari birgalikda qo‘llanadigan gibrid yondashuvdan foydalanish foydali bo‘lishi mumkin.

Tarkibiy tuzilma grammatikasi va bog‘liqlik grammatikasini taqqoslash: Constituency grammatika so‘zlarni birikma va jumlalarga guruhlashga e’tibor qaratadi, dependency grammatikasi esa har bir so‘z o‘rtasidagi munosabatlarga qaratiladi. Har bir so‘z gapda bog‘lanish grafida tugun sifatida ifodalanadi va tugunlar o‘rtasidagi qirralar so‘zlar o‘rtasidagi grammatik munosabatlarni ifodalaydi. Bog‘lanish grammatikasi ayrim tillar va NLP vazifalari uchun ko‘proq kuchli bo‘lishi mumkin, chunki u so‘zlar o‘rtasidagi munosabatlarni aniqroq tushuntiradi, ammo buni amalga oshirishda yanada murakkab va kamroq intuitivdir. Constituency grammatika esa ko‘proq tushunarli va amalga oshirish osonroq, lekin ifodalashda kamroq kuchli bo‘lishi mumkin. Quyidagi jadvalda ikki parsing turining o‘xshash va farqli jihatlari aniq ko‘rsatilgan.

Dependency va Constituency Parsing taqqoslanishi

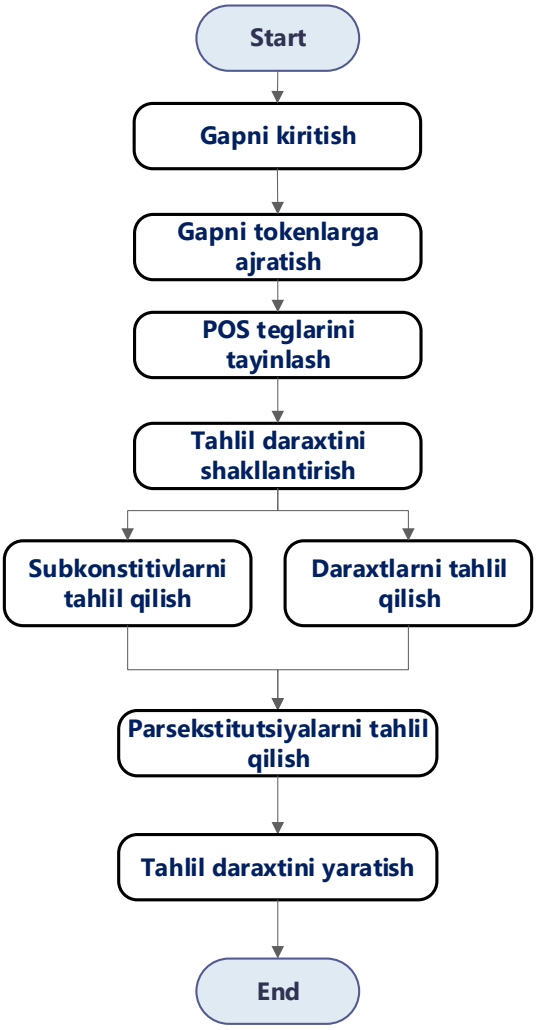
Dependency Parsing		Constituency Parsing
So‘zlar orasidagi bog‘lanishlarni aniqlash		Birikmalar tuzilishini aniqlash
	Sintaktik tahlil uchun ishlatiladi	
Graf asosida tahlil		Daraxt asosida tahlil
	NLP modellari uchun muhim	
Universal Dependencies (UD)		Phrase Structure Grammar (PSG)
	Taglangan korpus talab qiladi	
Ko‘p tillar uchun moslashuvchan		Ko‘pincha ingliz tili uchun ishlatiladi

Bog‘liqlik tahlili sintaktik va semantik tahlil uchun ishlab chiqilgan bo‘lib, u bi-leksikallashgan bog‘liqlik grammatikasiga, ya’ni ikki so‘z munosabatiga asoslanadi. Bu yerda barcha sintaktik va semantik hodisalar bi-leksikallashgan bog‘liqliklar orqali ifodalanadi. 2-rasmda bog‘liqlik tahlilining namunaviy daraxti keltirilgan. Tahlilda UD universal dependency teglaridan foydalanilgan. Turli bog‘liqlik analizatorlarini baholashda asosiy mezon sifatida bog‘liqlik aniqligi qo‘llaniladi, bu esa teglanmagan birikma ko‘rsatkichi (UAS – *Unlabeled Attachment Score*) va teglangan birikma ko‘rsatkichi (LAS – *Labeled Attachment Score*) bilan o‘lchanadi. Boshlang‘ich bosqichlarda bog‘liqlik tahlili daraxt shaklida, ya’ni proyektiv yoki no-proyektiv tuzilmalar doirasida cheklangan edi. So‘nggi yillarda esa bog‘liqlik grafigi asosidagi tahlilga qaratilgan bir qancha tadqiqotlar olib borilmoqda. Bir tomondan, dastlabki bog‘liqlik daraxtlari asosan sintaktik yo‘naltirilgan bo‘lsa, bugungi kunda so‘zlar orasidagi semantik munosabatlarga e’tibor kuchaymoqda.



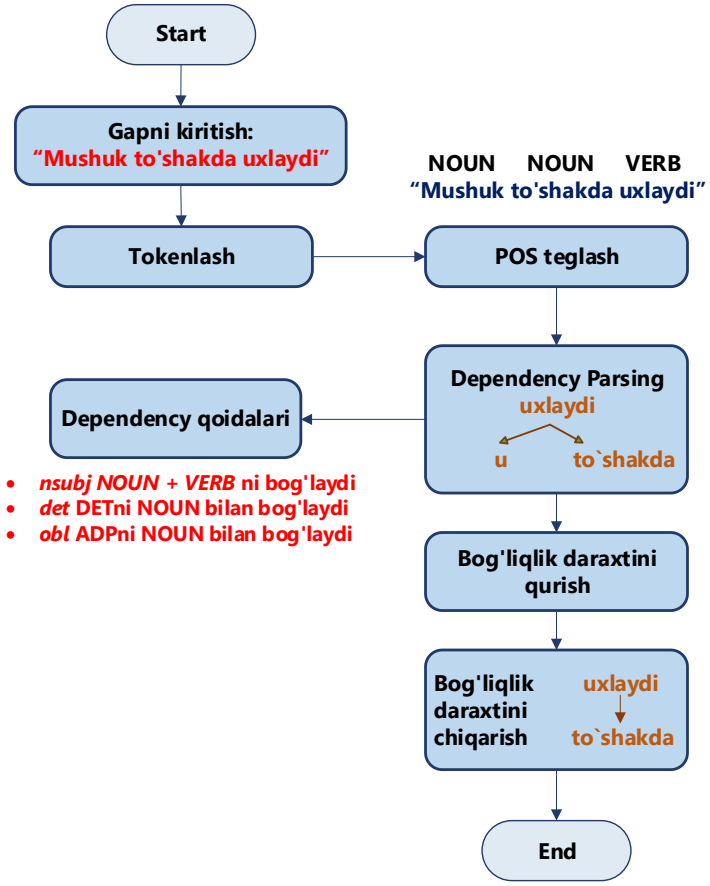
2-rasm. Dependency daraxti namunasi

Tarkibiy tuzilma tahlili Chomsky iyerarxiyasiga binoan qat’iy qoidalarga tayangani bois, qat’iy so‘z tartibli tillarda (masalan, ingliz tilida SVO) yaxshi ishlaydi, chunki grammatik munosabatlar asosan so‘zlarning tartibi bilan ifodalanadi. Ammo erkin so‘z tartibli tillarda bir fikrni turlicha tartibda ifodalash mumkin bo‘lganligi tufayli, an’anaviy tuzilmaviy grammatikalar murakkablashadi. Bog‘liqlik va tarkibiy tuzilma tahlil usullarida gaplarni tahlil qilish jarayoni ham farq qiladi.



3-rasm. Constituency parsing asosida sintaktik tahlil qilish jarayoni

Amaliy jihatdan, bugungi kunda ko‘pchilik zamonaviy parserlar bog‘liqlik paradigmasidan foydalanadi yoki dependency daraxtlarni hosil qiladi. Bunga sabab – Universal Dependencies kabi standartlarning joriy qilinishi va bog‘liqlikning kross-lingvistik (cross-lingual generalization) moslashuvchanligidir. Bu holat esa, ayniqsa, tabiiy tilni qayta ishlash (NLP) sohasida muhim bo‘lib, ko‘p tillilik (multilingualism) va kam resursli tillarni qo‘llab-quvvatlashda katta rol o‘ynaydi. O‘zbek tili erkin so‘z tartibiga ega tillardan biri bo‘lganligi uchun, oxirgi tadqiqotlarda bog‘liqlik tahlili qo‘llash samarali ekanligi xususida yozilmoqda. Chunki constituent daraxtini qurish uchun o‘zbek tilida qat’iy qoidalar asosida birikmalarni aniqlash mushkuldir. Chunki gap bo‘laklari doim bir o‘rinda kelishi yoki kelmasligi ham mumkin. Shuningdek, gaplar egali yoki egasiz bo‘lishi mumkin. Shu sababli ham bog‘liqlik grammatikasi o‘zbek tilini tahlil qilishda qulay usul bo‘lmoqda.



4-rasm. Dependency parsing asosida sintaktik tahlil qilish jarayoni

Xulosa. Grammatika – bu to‘g‘ri tuzilgan jumlalarni hosil qilish uchun qoidalar sifatida aniqlanadi va tabiiy tillarda muloqotda ishlatiladigan sintaktik qoidalarni belgilashda muhim rol o‘ynaydi. Sintaksis grammatikani tashkil etuvchi tarmoq bo‘lib, tilning jumlalar tuzilishini va shakllanishini o‘rganadi, ya’ni so‘zlar qanday qilib birikmalar, bandlar va jumlalarni hosil qilish uchun tartiblanishini belgilaydi.

Context free grammar – bu tilni tanib olish uchun ishlatiladigan formal grammatika bo‘lib, unda ishlab chiqarish qoidalari yordamida terminal va no-terminal belgilarni almashtirish orqali tilning satrlarini aniqlash ko‘rsatiladi. Constituency grammatika - bu tilning jumlalar tuzilishini tahlil qilishda ishlatiladi va jumlalar ma’lum grammatik funksiyaga ega bo‘lgan kichik birliklarga - tarkiblarga ajratilishi mumkinligi g‘oyasiga asoslanadi. Bog‘lanish (Dependency) grammatikasi - bu NLPda so‘zlar o‘rtasidagi bog‘lanishlar yordamida jumlaning grammatik munosabatlarini tahlil qilish uchun ishlatiladigan usul bo‘lib, so‘zlar o‘rtasidagi bog‘lanishlar asosida jumla tuzilishini shakllantiriladi. O‘zbek tilida CFG (kontekstsiz grammatika) va bog‘liqlik grammatikalari asosida, kross-domen va kross-lingvistik muhitlarda parsing masalalarini yechish imkoniyati mavjud. Bundan tashqari, qo‘shma modellar (joint models) va parser qo‘llanmalari (parser

applications) nuqtayi nazaridan, ko‘p vazifali o‘rganish (multi-task learning) va oldindan o‘rgatilgan (pretraining) arxitekturalar asosida ham o‘zbek tili matnlarini sintaktik tahlil qiluvchi dasturlar ishlab chiqilishi va samarali natijalarga erishish mumkin.

Foydalanilgan adabiyotlar:

1. Meishan Zhang. A Survey of Syntactic-Semantic Parsing Based on Constituent and Dependency Structures. Science China Press and Springer -Verlag Berlin Heidelberg, 2017.
2. Zong Chengqing. Statistical Natural Language Processing. - Tsinghua University Press, 2013.
3. Dan Jurafsky and James H. Martin. Speech and Language Processing (3rd (draft) ed.). Prentice Hall PTR, 2019.
4. https://www.scaler.com/topics/nlp/introduction-to-grammar-in-nlp/
5. James Henderson. Discriminative training of a neural network statistical parser. In Proceedings of ACL, pages 95–102, 2004.
6. Jan Hajic, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, and et al. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In Proceedings of CoNLL 2009: Shared Task, pages 1–18, 2009.
7. Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Silvie Cinkova, Dan Flickinger, Jan Ha´jic, and Zdenka Uresova. Semeval 2015 task 18: Broadcoverage semantic dependency parsing. In Proceedings of SemEval, pages 915–926, 2015.

O‘ZBEK TILI MATNLARIDA UNDOV GAPLARNI TAHLIL QILISHNING QOIDAGA ASOSLANGAN ALGORITMI

Sharipov Maksud Siddiqovich,
Texnika fanlari nomzodi, dotsent
maqsbek72@gmail.com
Urganch davlat universiteti “Kompyuter ilmlari” kafedrasi dotsenti

Adinayev Xushnudbek Saylboyevich,
hushnudbek.adinaev@gmail.com
TATU Urganch filiali o‘quv-uslubiy boshqarma boshlig‘i

Yusupova Mukaddas Muratovna,
mukaddas0783@gmail.com
Urganch Ranch texnologiya universiteti 1-kurs magistranti

ANNOTATSIYA. Bu ishda biz o‘zbek tili matnlari undov gaplarini tahlil qilishning qoidaga asoslangan algoritmini ishlab chiqishni ko‘rib chiqamiz. Undov gaplar — gaplovchi shaxsning his-tuyg‘ularini (hayrat, quvonch, g‘azab, achinish, chaqiriq va h.k.) ifodalovchi gaplar bo‘lib, ko‘pincha undov so‘zlar, emotsional ohang yoki maxsus so‘z tartibi bilan ajralib turadi. O‘zbek tili kam resursli til hisoblangaligi sababali hozirgacha bunday algoritmlar ishlab chiqilmagan. O‘zbek tili matnlaridagi gaplarning undov belgilari to‘g‘ri yoki noto‘g‘ri qo‘yilganligini aniqlash masalasini yechishning qoidaga asoslangan algoritmi ishlab chiqilgan va qoidalar bazasi yaratilgan.

ABSTRACT. In this work, we consider the development of a rule-based algorithm for analyzing exclamatory sentences in Uzbek texts. Exclamatory sentences are sentences that express the speaker's emotions (surprise, joy, anger, pity, appeal, etc.) and are often distinguished by exclamatory words, emotional tone, or special word order. Due to the fact that the Uzbek language is considered a low-resource language, such algorithms have not been developed so far. A rule-based algorithm for solving the problem of determining whether exclamation marks are placed correctly or incorrectly in sentences in Uzbek texts has been developed and a rule base has been created.

АННОТАЦИЯ. В данной работе рассматривается разработка алгоритма анализа восклицательных предложений в узбекских текстах на основе правил. Восклицательные предложения — это предложения, которые выражают чувства говорящего (удивление, радость, гнев, жалость, восклицание и т. д.) и часто характеризуются восклицательными словами, эмоциональным тоном или особым порядком слов. В связи с тем, что узбекский язык считается языком с ограниченными ресурсами, такие алгоритмы пока не разработаны. Разработан алгоритм на основе правил и создана база правил для решения

задачи определения правильности или неправильности расстановки восклицательных знаков в предложениях узбекских текстов.

Kalit so‘zlar. *Tinish belgilari, NLP, undov belgisi, undov gaplar.*

KIRISH

O‘zbek tili matnlarida undov belgisining xato yoki to‘g‘ri qo‘yilganligini aniqlash (punktuatsion tahlil) masalasi tabiy tilni qayta ishlash (NLP) masalalaridan biri hisoblanadi. Punktuatsion tahlil tabiiy tilni qayta ishlashda mashinali tarjima, matn tahlili, semantik tahlil, sintaktik tahlil kabi masalalarni hal qilishda ishlatiladi.

Punktatsiya qoidalari va me’yorlarini bilish esa yozma fikrni aniq va to‘g‘ri, mantiqli va ifodali, ravon va tushunarli hamda maqsadga muvofiq bayon qilishda muhimdir. Tinish belgilarining asosiy vazifasi yozuvda ifodalanadigan nutqning fikriy bo‘linishini, gapning qanday qismlarga ajralishini va u bo‘laklarning o‘zaro munosabatlarini ko‘rsatishdan iboratdir. Tinish belgilari gap mazmuni, tuzilishi va ohangiga ko‘ra qo‘llaniladi. Globallashuv davrida axborot texnologiyalari kun sayin hayotimizdan chuqur o‘rin egallamoqda. Bugun yozma nutq deyarli axborot texnologiyalari orqali amalga oshmoqda. Biron matn yoki xabarni elektron holatda yozilar ekan, bevosita tinish belgilariga ham murojaat qilinadi. Undov belgisi o‘zbek yozuvlarida XX asrdan ko‘pchilik tinish belgilari kabioha davrning yagona adabiy til me’yorlariga amal qilingan “Turkiston viloyatining gazeti”da muntazam qo‘llangan.[1].

METODOLOGIYASI

Tarkibida emotsionallikni kuchaytiradigan qanday, qanchalik, qanchalar, qanaqa, naqadar, muncha, bunchalik kabi so‘zlar mavjud bo‘lgan gaplar oxirida undov belgisi qo‘yiladi.

Yashash qanday yaxshi, tag‘in shunday katta, go‘zal shaharda!

Emotsionallik darajasi kuchli bo‘lgan ritorik so‘roq gaplar oxirida ba’zan undov belgisi qo‘yiladi.

Ona o‘z bolasiga yomonlikni ravo ko‘radimi!

1. Undov gapning mazmuni (semantik belgi)

- Gapda **gaplovchining his-tuyg‘usi, emotsiyasi yoki kayfiyati** ifodalanadi: hayrat, quvonch, g‘azab, chaqiriq, og‘riq, achinish va hokazo.

Misollar:

- Voy, qanday go‘zal manzara!
- Oh, belim!

2. Undov so‘zlarning ishtiroki (leksik belgi)

Undov gaplarda ko‘pincha maxsus **undov so‘zlar** ishlatiladi:

Kategoriya	Namuna undov so‘zlar
Hayrat	Voy, opa-ey, eh-ha, hay-hay
Og‘riq	Oh, o‘h
Achinish	Bechora, afsus, voy-bo‘ldim
Chaqiriq	Qani, yur, ana, mana

Qoidasi: Gapda ushbu undov so‘zlardan biri qatnashsa, bu gap undov gap bo‘lishi ehtimoli yuqori.

3. Gap oxiridagi punktuatsiya belgisi (grafik belgi)

Undov gaplar odatda quyidagi belgilar bilan tugaydi:

- ! — Undov belgisi
- ?! — Hayrat, so‘roq va undovni birlashtirgan belgilar

Qoidasi: Gap ! yoki ?! bilan tugasa, bu undov gap bo‘lish ehtimoli yuqori.

4. Gap tuzilmasi (sintaktik belgi)

- Undov gaplar ko‘pincha **egasiz yoki iboraviy** bo‘ladi:
  - "Voy!" – egasiz
  - "Qani yur!" – buyruq + chaqiriq
  - "Bechora bola!" – iboraviy, emotsional

Qoidasi: Gapda ot+ot, undov+ot, yoki buyruq shaklida tuzilgan qisqa gaplar bo‘lsa — undov gap bo‘lishi mumkin.

5. Gapning ohangidan foydalanish (nutq belgi)

Yozma matnda to‘g‘ridan-to‘g‘ri ohangni aniqlash qiyin, lekin:

- ! belgisi
- undov so‘zlar
- qisqa va emotsional tuzilma

— bu belgilar **emotsional ohang** borligini ko‘rsatadi.

6. Chaqiriq va buyruqni ifodalovchi gaplar

- Ko‘pincha harakatga da’vat yoki tezkor amal ifodalangan:
  - "Qani boshladik!"
  - "Yur, ketdik!"

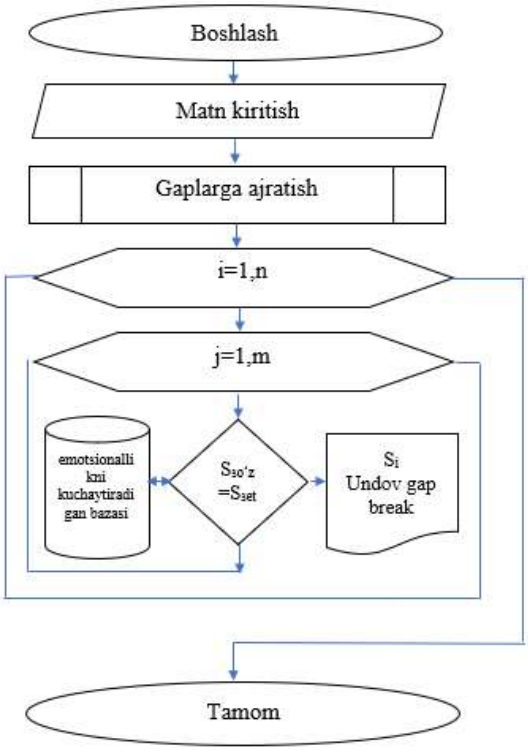
Qoidasi: Gap chaqiriq yoki topshiriq bo‘lsa, lekin emotsional tusda bo‘lsa — undov gap hisoblanadi.

UNDOV GAPLAR TAHLILINING ALGORITMI

Ushbu maqolada matndagi undov gaplarni tahlil qilish aloritmi ishlab chiqildi. Aloritm matndagi undov belgisini to‘g‘ri qo‘yilgan yoki noto‘g‘ri qo‘yilganligini tahlil qiladi. Quyida algortmning ishlash jarayonini ko‘ib chiqamiz.

1. Dasturni boshlash.
2. Dasturga matnni kiritish.

3. Tokenizatsiya jarayoni.
4. Ikkita sikl ochiladi. Dastlab sikl gaplarning sonigacha keying sikl gapdagi so‘zlar sonigacha, Xar bir so‘zni emotsionallikni kuchaytiradigan so‘zlar bor yo‘qligi bilan tekshiradi. Agar shunday so‘z bo‘lsa ekranga undov gap deb chiaradi.
5. Dasturni yakunlash.



1-rasm: O‘zbek tili matnlarida undov gaplarni aniqlash algoritmi.

NATIJA

Dastur 30 turdagi matn, 137 848 gap, 1 642 860 so‘zdan iborat Uzbek korpusidagi matnlar bilan sinovdan o‘tkazildi. Korpusda 3492 ta undov gap mavjud. Dastur 3492 ta undov gapni topdi va 111 tasini undov gap emas deb topdi. Jami 133 945 undov emas gapning 300 tasini undov gap deb topdi. Bu 74.3% aniqlikni beradi.

№	Matn turlari	Gaplar soni	Undov gap		Undov gap emas		F1 Score
			undov gap	undov gap emas	undov gap	undov gap emas	
	Adabiyot	11659	1143	11	14	10491	98,9
	Anatomiya	10565	1	1	1	10562	50,0
	Biologiya	4757	0	0	3	4754	0,0

	Botonika	2717	2	1	1	2713	66,7
	Diniy	6079	20	2	10	6047	76,9
	Fizika	12078	6	1	6	12065	63,2
	Geografiya	6237	0	0	19	6218	0,0
	Huquq	1964	5	1	2	1956	76,9
	Ijtimoiy soha	421	0	0	4	417	0,0
	Informatika	2512	22	3	12	2475	74,6
	Iqtisodiyot	520	1	0	1	518	66,7
	Kimyo	3566	18	2	14	3532	69,2
	Kino	668	3	1	1	663	75,0
	Matematika	7921	22	3	12	7884	74,6
	O‘qish kitobi	3382	193	12	34	3143	89,4
	O‘zbek tili kitob	11020	56	9	26	10929	76,2
	Qishloq xo‘jaligi	1209	0	0	12	1197	0,0
	Qissa	10529	1729	32	26	8742	98,4
	Qonun va qarorlar	4360	1	0	1	4358	66,7
	Rassomchilik	3673	12	1	12	3648	64,9
	Roman	1896	80	8	19	1789	85,6
	San’at	9570	129	12	23	9406	88,1
	Siyosat	1222	7	1	2	1212	82,1
	Sport	2565	18	5	5	2537	78,3
	Tabobat	421	1	1	1	418	51,3
	Tarix	4619	5	1	2	4611	76,9
	Texnologiya	796	7	2	2	785	77,8
	Tibbiyot	5490	0	0	23	5467	0,0
	Urush	1084	10	1	11	1062	62,8
	Zoologiya	4348	1	0	1	4346	66,7
	jami	**137848**	**3492**	**111**	**300**	**133945**	**74,3**

XULOSA

Hozirgi kungacha o‘zbek tili matnlari uchun undov belgilarini tahlil qilish agoritmlari deyarli ishlab chiqilmaganligi sababli ushubu ishda taklif qilingan algoritmlar dolzarb hisoblanadi. O‘zbek tilida tabiiy tilni qayta ishlash masalasi bo‘lgan punktuatsion tahlil (undov belgilari) qilishishni qoidaga asosolangan algoritmlari ishlab chiqildi, punktuatsion tahlil (undov belgilari) uchun qoidalar bazasi yaratildi. Ishlab chiqilgan algoritmlar yordamida o‘zbek tilining 30 ta kategoriyadan iborat 1 642 860 ta so‘z va 137 848 ta gapdan iborat korpus tahlil qilinib natijalar olindi. Ushbu ishning davomi sifatida keljakda o‘zbek tili matnlarini punktuatsion (undov gaplarni) tahlil qilish masalasini sun’iy intellekt modellaridan foydalanib yechish va yaxshiroq aniqlikdagi natijalarni olish maqsad qilingan.

Foyadalanilgan adabiyotlar:

1. M. A., M. N., A.D. Mahmudov N., O‘zbek tili me’yorlari (Punktuatsiya), vol. 1. Tashkent: Zamin nashr, 2021.
2. J. Pokrywka, “Punctuation Prediction for Polish Texts using Transformers,” Feb. 2024. doi: 10.48550/arXiv.2410.04621.
3. P. Żelasko, P. Szymański, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, “Punctuation Prediction Model for Conversational Speech,” Feb. 2018, pp. 2633–2637. doi: 10.21437/Interspeech.2018-1096.
4. M.S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.
5. M. Bayraktar, B. Say, and V. Akman, “An analysis of English punctuation: the special case of comma,” International Journal of Corpus Linguistics, vol. 3, Jul. 1998, doi: 10.1075/ijcl.3.1.03bay.
6. D.Hardt, “Comma checking in Danish,” 2001.
7. U.Salaev, E.Kuriyozov, and C. Gomez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022, pp. 199 – 206. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138700420&partnerID=40&md5=bf476cd74317f06577dd0548c5c600d6
8. A. M.Abdurashetona and I. O. Ismailovich, “Methods of Tagging Part of Speech of Uzbek Language,” in Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021, 2021, pp. 82 – 85. doi: 10.1109/UBMK52708.2021.9558900.
9. A. M.Abdurashetona and U. Mokhiyakon, “Software Features and Linguistic Features of Uzbek Synonymizer,” in Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022, 2022, pp. 171 – 175. doi: 10.1109/UBMK55850.2022.9919447.
10. B. Mengliyev, Sh.Shahabitdinova, Sh.Khamroeva, Sh.Gulyamova, and A.Botirova, “The morphological analysis and synthesis of word forms in the linguistic analyzer,” Journal of Language and Linguistic Studies, vol. 17, no. 1, pp. 558 – 564, 2021, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103797274&partnerID=40&md5=8a4052419f721c3c734f8fe1c48984ec
11. K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of Karakalpak language stop words,” Data Brief, vol. 48, 2023, doi: 10.1016/j.dib.2023.109111.

12. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language”, CEUR Workshop Proceedings, vol. 3315, pp. 93–98, 2022.

13. D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov, and O. Alloyorov, “Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language,” Jan. 2023, pp. 1720–1724. doi: 10.1109/APEIE59731.2023.10347765.

14. K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of stopwords extracted from Uzbek texts”, Data in Brief, vol. 43, 2022.

O‘ZBEK TILI MATNLARIDA SAVOLLARNI ANIQLASHNING AVTOMATLASHTIRILGAN TIZIMI UCHUN MASHINANI O‘RGANISH YONDASHUVI

Sharipov Maksud Siddiqovich,
Texnika fanlari nomzodi, dotsent
maqsbek72@gmail.com
Urganch davlat universiteti

Adinayev Xushnudbek Saylboyevich,
hushnudbek.adinaev@gmail.com
TATU Urganch filiali o‘quv-uslubiy boshqarma boshlig‘i

Yusupova Mukaddas Muratovna,
1-kurs magistranti
mukaddas0783@gmail.com
Urganch Ranch texnologiya universiteti

Qadamboyeva Durdona Dilshod qizi,
1-kurs magistrant
durdona06@gmail.com
Urganch davlat universiteti

ANNOTATSIYA. Bu ishda biz o‘zbek tili matnlari undov gaplarini tahlil qilishning qoidaga asoslangan algoritmini ishlab chiqishni ko‘rib chiqamiz. O‘zbek tili kam resursli til hisoblangaligi sababali hozirgacha bunday algoritmlar ishlab chiqilmagan. O‘zbek tili matnlaridagi gaplarning undov belgilari to‘g‘ri yoki noto‘g‘ri qo‘yilganligini aniqlash masalasini yechishning qoidaga asoslangan algoritmi ishlab chiqilgan va qoidalar bazasi yaratilgan. Til - bu muloqot usuli bo‘lib, uning yordamida biz gapirish, o‘qish va yozishimiz mumkin. Misol uchun, biz o‘ylaymiz, biz tabiiy tilda qarorlar, rejalar va boshqalarni qabul qilamiz; aniq, so‘z bilan. Biroq, ushbu AI davrida bizni duch keladigan katta savol shundaki, biz kompyuterlar bilan xuddi shunday tarzda muloqot qila olamizmi? Boshqacha qilib aytadigan bo‘lsak, inson o‘z tabiiy tilida kompyuterlar bilan muloqot qila oladimi? NLP dasturlarini ishlab chiqish biz uchun qiyin, chunki kompyuterlar tuzilgan ma’lumotlarga muhtoj, lekin inson nutqi tuzilmagan va tabiatan ko‘pincha noaniq.

ABSTRACT. In this work, we consider the development of a rule-based algorithm for analyzing exclamatory sentences in Uzbek texts. Since Uzbek is considered a low-resource language, such algorithms have not been developed so far. A rule-based algorithm for solving the problem of determining whether exclamation marks are placed correctly or incorrectly in sentences in Uzbek texts has been developed and a rule base has been created. Language is a means of communication with which we can speak, read, and write. For example, we think, we make decisions, plans, and so on in natural language; specifically, with words.

However, the big question that faces us in this AI era is whether we can communicate with computers in the same way? In other words, can humans communicate with computers in their natural language? Developing NLP programs is difficult for us because computers need structured information, but human speech is unstructured and often ambiguous by nature.

АННОТАЦИЯ. В данной работе рассматривается разработка алгоритма анализа восклицательных предложений в узбекских текстах на основе правил. В связи с тем, что узбекский язык считается языком с ограниченными ресурсами, такие алгоритмы пока не разработаны. Разработан алгоритм на основе правил и создана база правил для решения задачи определения правильности или неправильности расстановки восклицательных знаков в предложениях узбекских текстов. Язык — это способ общения, посредством которого мы можем говорить, читать и писать. Например, мы думаем, принимаем решения, планируем и т. д. на естественном языке; ясно, словами. Однако главный вопрос, стоящий перед нами в эпоху искусственного интеллекта, заключается в том, сможем ли мы общаться с компьютерами таким же образом? Другими словами, могут ли люди общаться с компьютерами на их естественном языке? Нам сложно разрабатывать программы обработки естественного языка, поскольку компьютерам нужны структурированные данные, а человеческая речь по своей природе неструктурирована и часто неоднозначна.

Shu ma’noda aytishimiz mumkinki, tabiiy tilni qayta ishlash (NLP) bu kompyuter fanining, ayniqsa sun’iy intellektning (AI) kichik sohasi bo‘lib, u kompyuterlarga inson tilini tushunish va qayta ishlashga imkon beradi. Texnik jihatdan, NLP ning asosiy vazifasi tabiiy tildagi katta hajmdagi ma’lumotlarni tahlil qilish va qayta ishlash uchun kompyuterlarni dasturlashdir.

Kalit so‘zlar. *Tinish belgilari, NLP, so‘roq belgisi, so‘roq gaplar, QA, SVM, LSTM, KNN.*

Kirish

Dunyo bo‘ylab har soniyada turli savollar beriladi. Ushbu savollarga javob topish uchun biz ushbu mavzu bo‘yicha ma’lumotga ega bo‘lishimiz kerak. Bu ma’lumotlar kitoblarda, qog‘ozlarda va eng muhimi internetda. Internet va Internetning takomillashuvi bilan tobora ko‘proq ma’lumotlar mavjud bo‘lib, hamma uchun topish osonroq bo‘ldi. Muayyan savolga javob topishimiz kerak bo‘lganda, biz ushbu savolga tegishli ba’zi ma’lumotlardan javob olamiz. Biz ushbu mavzudagi mavjud ma’lumotlarni qo‘lda o‘qiymiz va javob topishga harakat qilamiz. Qo‘shimcha ma’lumot Internet va Internetda mavjud bo‘lgani uchun, agar

biz ba’zi hujjatlarni o‘qish va ularga javob olish jarayonini avtomatlashtira olsak, ko‘proq savollarga avtomatik javob berishimiz mumkin bo‘ladi. Ushbu maqolada biz bengal tilida avtomatlashtirilgan savollarga javob berish tizimini (QA) yaratish uchun qo‘llanilgan dastlabki qadamlar va qo‘llaniladigan usullarni tasvirlab beramiz. Savollarni aniqlash QA tizimida katta maqsadga xizmat qiladi [1]. Odamlar ko‘pincha tabiiy tilda savollar berishadi, ular grammatik qoidalarga rioya qilmaydilar. Shunday qilib, jumla savolmi yoki yo‘qmi, avvalo aniqlanishi kerak. Odamlar tomonidan berilgan savollarning turlicha shakli aniqlashni qiyinlashtiradi [2]. Berilgan savolga hujjatlardan javob olish uchun tizim ushbu savolga ishlov berishi kerak va uni qayta ishlashning birinchi bosqichi savollarni aniqlashdir, shuning uchun bu savolga javob berish tizimining birinchi bosqichidir [3]. Savolni aniqlash qismidan so‘ng tizim savolni oldindan belgilangan toifalar toifasiga ajratishi kerak. Ushbu tasnif kutilgan javob turini taxmin qilishga va qidiruv domenini qisqartirishga yordam beradi. Demak, tasniflash QA tizimining muhim qismidir.

Savollarni aniqlash - bu savollarga javob berish tizimining dastlabki vazifalari. Savollarni aniqlash, QA tizimining birinchi bosqichi SVM, Logistik regressiya,
K-Neighbor klassifikatori, Ko‘p qatlamli Perceptron va LSTMni qo‘llash orqali amalga oshirildi. Biz chiziqli yadro hiylasi bilan SVM uchun eng yaxshi ishlashga erishdik. 2000 xususiyatli so‘zlar uchun biz 1,4% xatolik darajasiga ega bo‘ldik, bu biz Bengal tili uchun ishlatgan yondashuvlar orasida eng yaxshisidir. LSTM algoritmi uchun biz 3,22% xatolik darajasiga ega bo‘ldik. LSTM katta va yaxshiroq ma’lumotlar to‘plami bilan istiqbolli ko‘rinadi.

Ma’lumotlar to‘plamini tayyorlash va tahlil qilish

Biz jami 8000 qatorli o‘zbek tili matnidan iborat ma’lumotlar to‘plamini tayyorladik. Bizning ma’lumotlar to‘plamimizda 3100 ta savol va 4900 ta savol bo‘lmagan. Ushbu ma’lumotlar turli veb-saytlardan to‘plangan. O‘bekistonning mashhur shoirlari haqidagi ba’zi maqolalar Vikipediyadan to‘plangan. Shuningdek, biz matnlardan savollarga javoblar juftligini to‘pladik. Ushbu ma’lumotlar qo‘lda tayyorlangan. Keyin biz ushbu ma’lumotlarni savolni aniqlash maqsadida belgilashimiz kerak edi. Biz ma’lumotlarni ikki toifaga ajratdik: savol va savolsiz. Biz ishimiz uchun butun ma’lumotlar to‘plamini qo‘lda tozaladik. Keyin biz keyingi ishlov berish uchun jumlalarni tokenizatsiya qildik. Keyin biz olgan jumlalarda so‘zlarni leksema qildik. Keyin countvectorizer yordamida so‘zlarni vec ga aylantirdik. Shundan so‘ng, TF/IDF yordamida biz so‘zlarni to‘pladik va eng qimmatli so‘zlarni oldik,

Xususiyatlarni chiqarish

Jumlalarni tokenizatsiya qilish: Birinchidan, biz har bir jumlani keyingi qayta ishlash uchun to‘plangan parchadan ajratdik. Jumlalarni tokenizatsiya qilishda yaxshi natijaga erishish uchun ko‘p narsalar ko‘rib chiqildi.

So‘zlarni tokenizatsiya qilish: jumlalar tokenizatsiyasidan so‘ng biz ushbu jumlalarda so‘z tokenizatsiyasini qo‘lladik. So‘zlarni tokenizatsiya qilish orqali biz jumlalarning har bir so‘zini ajratdik va ularni keyingi qayta ishlash uchun saqladik.

Wh-so‘z: So‘z tokenizatsiyasidan keyin vazifamiz biz uchun osonlashdi. Bengal tilida xuddi ingliz tilidagi wh so‘zi kabi ba’zi so‘zlar mavjud. Bu so‘zlar biz uchun juda foydali bo‘ldi. Ular asosan savollarda uchraydi. Shunday qilib, bu Wh so‘zlari so‘z tokenizatsiyasidan keyin eng ko‘p uchraydigan so‘zlar ekanligi aniq.

Biz bu so‘zlarga boshqa so‘zlarga qaraganda ustunlik berdik. Wh-so‘zga ko‘proq ustunlik berilgandan so‘ng, bu so‘zlar sodir bo‘lsa, gapning savol bo‘lish ehtimoli ortadi. Jadval II Bengal Wh-so‘zlarning ba’zi misollarini ko‘rsatadi.

Trening va tan olish

Trening maqsadida biz SVM, MLP, KNN, Logistic Regression va LSTM algoritmlaridan foydalandik. Ushbu algoritmlar 8000 Bengal matnidan iborat katta ma’lumotlar to‘plamimizda ishga tushirildi. Dastlab biz SVM algoritmini ishga tushirdik va SVM ning 3 xil yadrosidan foydalandik. Yadrolar Lineer, Polynomial va RBF. Biz MLP dan foydalandik, bu erda yashirin qatlamlar o‘lchami 12 va hal qiluvchi lbfg bo‘lgan. KNN biz ishlatgan yana bir mashinani o‘rganish algoritmidir. Barg o‘lchami 30 ta bo‘lib, algoritm avtomatik hisoblanadi. KNN metrikasi minkowski, ishlatilgan og‘irliklar esa masofa edi. Maqsadimiz bo‘yicha matn ma’lumotlar to‘plamini tasniflash uchun biz mashinani o‘rganish algoritmini LR (Logistik regressiya) ishlatdik. LSTM matn ma’lumotlarini tasniflash uchun eng ko‘p ishlatiladigan algoritmlardan biridir. Biz foydalangan oxirgi algoritm LSTM. LSTM modelini amalga oshirishda xususiyatli so‘zlar soni 20000 tani tashkil etdi. Ketma-ket model qo‘llanildi va kirish uzunligi 42. Bizning modelimizda foydalanilgan jami qatlamlar 21 ta, faollashtirish funksiyasi sigmasimon. SVM uchun eng yaxshi natijaga erishiladi. Agar ma’lumotlar to‘plami yaxshilansa, LSTM bilan ishlash yaxshilanadi.

Metodologiyasi

O‘zbek tilida yozilgan matnlarda savol gaplarni avtomatik tarzda aniqlash ko‘plab tizimlar uchun muhim: chatbotlar, matn tahlili, avtomatik javob berish tizimlari, savol-javob platformalari va boshqalar. Biroq, O‘zbek tili uchun NLP resurslar kamligi bu vazifani murakkablashtiradi.

Metodologiya bosqichlari

1. Ma’lumotlar to‘plami tayyorlash

- O‘zbek tilidagi matnlar (forumlar, maqolalar, chatlar) to‘planadi.
- Har bir gap **"savol"** yoki **"savol emas"** deb belgilanadi.
- Agar mavjud bo‘lsa, mavjud datasetlardan (masalan, O‘zbekcha OpenCorpora yoki O‘zNLP) foydalaniladi.

2. Matnni oldindan qayta ishlash

- Gaplarni ajratish (tokenizatsiya).
- Belgilarni tozalash (punktuatsiya, emoji, html taglar).
- Harf o‘lchamlarini bir xil qilish (masalan, kichik harflarga o‘tkazish).
- Lemmatizatsiya yoki stemming (ixtiyoriy).

3. Belgilarni (Feature) ajratish

Matndan quyidagi belgilar chiqariladi:

- Savol so‘zlari (kim, nima, qachon, qayerda, nega, nima uchun, qanday)
- Gap oxirida ? belgisi mavjudligi.
- -mi, -chi kabi savol qo‘shimchalari.
- Gap uzunligi.
- Gapdagi fe’l yoki yordamchi fe’llar.

4. Model tanlash va qurish

Variant A: Qoida asosidagi yondashuv (Rule-based)

- Regex (muntazam ifodalar) orqali savolni aniqlash.
- Foydalanuvchiga tezkor natija berish uchun soddalashtirilgan algoritm.

Variant B: Mashina o‘rganish (ML) asosida

- Klassifikatsiya modellari:
  - Naive Bayes
  - Logistic Regression
  - Random Forest
- Belgilar asosida model o‘rgatiladi va test qilinadi.

Variant C: Transformer modeli (Deep Learning)

- O‘zbek tilida o‘qitilgan BERT model (masalan, uzbekBERT, XLM-R, yoki mBERT) fine-tuning qilinadi.
- Klassifikatsiya uchun BERTning so‘nggi qatlamiga softmax layer qo‘shiladi.

5. Modelni baholash

Modellar quyidagi metrikalar bo‘yicha baholanadi:

- **Aniqlik (Accuracy)**
- **Aniqlovchanlik (Precision)**
- **Qamrov (Recall)**
- **F1-ko‘rsatkich**

Kross-valikatsiya usuli bilan turli matn turlari (chat, maqola, ijtimoiy tarmoq)da sinovdan o‘tkaziladi.

Natija

Ma’lumotlar to‘plamimizni muvaffaqiyatli o‘qitishdan so‘ng biz yuqori aniqlik va kamroq xatolikka erishdik, bu juda qoniqarli edi. LSTM dan tashqari har bir algoritmda bizning ma’lumotlar poezdi testi bo‘linishi 80/20 ni tashkil etdi, bu umumiy ma’lumotlarning 80% ta’lim uchun, 20% esa sinov uchun ishlatilgan. LSTM algoritmi uchun bizning o‘quv va sinov ma’lumotlarimiz mos ravishda 66,67% va 33,33% edi.

Xulosa

Biz bengal tili uchun QA tizimining dastlabki ishlarini bajardik. To‘liq QA tizimi uchun keyingi qadamlarni bajarish kerak. Buning uchun ko‘p kuch va etarlicha yaxshi ma’lumotlar to‘plami kerak. Agar ma’lumotlar to‘plami ko‘paytirilsa, biz LSTM yordamida yanada yaxshi ishlashga umid qilamiz. Yaxshiroq natijaga erishish uchun boshqa usullarni ham sinab ko‘rish mumkin. Savollarni tasniflash, ma’lumot olish va javoblarni olish QA tizimining qismlari bo‘lib, Bengal tabiiy tilini qayta ishlash bo‘yicha katta tadqiqot sohasi bo‘lishi mumkin. Savollarni aniqlash - bu savollarga javob berish tizimining kichik qismi va boshqa ko‘plab muammolar. U boshqa ko‘plab tadqiqot sohalari uchun eshikni ochishi mumkin. Ta’riflangan savollarni aniqlash tizimi yaxshiroq ma’lumotlar to‘plami bilan yaxshi ishlash imkonini beradi.

Foyadalanilgan adabiyotlar:

1. M. A. , M. N. , A. D. Mahmudov N., O*‘zbek tili meyorlari (Punktuatsiya)*, vol. 1. Tashkent: Zamin nashr, 2021.
2. J. Pokrywka, “Punctuation Prediction for Polish Texts using Transformers,” Feb. 2024. doi: 10.48550/arXiv.2410.04621.
3. P. Żelasko, P. Szymański, J. Mizgajski, A. Szymczak, Y. Carmiel, and N. Dehak, “Punctuation Prediction Model for Conversational Speech,” Feb. 2018, pp. 2633–2637. doi: 10.21437/Interspeech.2018-1096.
4. M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in *International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM*, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.
5. M. Bayraktar, B. Say, and V. Akman, “An analysis of English punctuation: the special case of comma,” International Journal of Corpus Linguistics, vol. 3, Jul. 1998, doi: 10.1075/ijcl.3.1.03bay.
6. D. Hardt, “Comma checking in Danish,” 2001.
7. U. Salaev, E. Kuriyozov, and C. Gomez-Rodríguez, “SimRelUz: Similarity and Relatedness scores as a Semantic Evaluation Dataset for Uzbek Language,” in 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages, SIGUL 2022 - held in conjunction with the International Conference on Language Resources and Evaluation, LREC 2022 - Proceedings, 2022, pp. 199 –

206. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85138700420&partnerID=40&md5=bf476cd74317f06577dd0548c5c600d6

8. A. M. Abdurashetona and I. O. Ismailovich, “Methods of Tagging Part of Speech of Uzbek Language,” in Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021, 2021, pp. 82 – 85. doi: 10.1109/UBMK52708.2021.9558900.

9. A. M. Abdurashetona and U. Mokhiyakon, “Software Features and Linguistic Features of Uzbek Synonymizer,” in Proceedings - 7th International Conference on Computer Science and Engineering, UBMK 2022, 2022, pp. 171 – 175. doi: 10.1109/UBMK55850.2022.9919447.

10. B. Mengliyev, S. Shahabitdinova, S. Khamroeva, S. Gulyamova, and A. Botirova, “The morphological analysis and synthesis of word forms in the linguistic analyzer,” Journal of Language and Linguistic Studies, vol. 17, no. 1, pp. 558 – 564, 2021, [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85103797274&partnerID=40&md5=8a4052419f721c3c734f8fe1c48984ec

11. K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of Karakalpak language stop words,” Data Brief, vol. 48, 2023, doi: 10.1016/j.dib.2023.109111.

12. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, ‘Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language’, CEUR Workshop Proceedings, vol. 3315, pp. 93–98, 2022.

13. D. Mengliev, E. Akhmedov, V. Barakhnin, Z. Hakimov, and O. Alloyorov, “Utilizing Lexicographic Resources for Sentiment Classification in Uzbek Language,” Jan. 2023, pp. 1720–1724. doi: 10.1109/APEIE59731.2023.10347765.

14. K. Madatov, S. Bekchanov, and J. Vicic, “Dataset of stopwords extracted from Uzbek texts”, Data in Brief, vol. 43, 2022.

UDC: 004.65, 8, 81’2, 81’28

MARKOV MODEL FOR TEXT GENERATION IN UZBEK: PRACTICAL APPLICATION AND RESULTS

Sharipov Maksud Siddiqovich,
Candidate of Technical Sciences, Docent
maqsbek72@gmail.com
Urgench State University

Kurbonova Ruzikajon Ulug‘bek qizi,
ruzikajonqurbanova0809@gmail.com
PhD student, Urgench State University

Annotatsiya: Ushbu maqolada o‘zbek tilida matn generatsiyasi uchun Markov modeli qo‘llanilishi o‘rganilgan. Tadqiqotda Markov modelining xususiy holati trigram modeli asosida matn yaratish jarayoni va natijalari tahlil qilingan. O‘zbek tilida hali bu borada yetarlicha ishlar amalga oshirilmagan. Eksperimentlar natijasida Markov modelining matn generatsiyasida mantiqiy bog‘liqlikni saqlashda cheklovlari borligi aniqlangan. Modelning aniqlik darajasi 0.73% bo‘lib, generatsiya qilingan matnlarning grammatik sifatini oshirish uchun N-gram uzunligini oshirish yoki chuqur o‘rganish (LSTM, Transformer) modellari bilan taqqoslash zarurligi qayd etilgan. Maqola Markov modelining o‘zbek tilida matn yaratishda imkoniyatlari va cheklovlarini tahlil qiladi.

Аннотация. В данной статье исследуется применение марковской модели для генерации текста на узбекском языке. В рамках исследования был проанализирован процесс создания текста на основе частного случая марковской модели — триграммной модели, а также её результаты. В узбекском языке пока еще недостаточно работы в этом направлении. Эксперименты показали, что марковская модель имеет ограничения в сохранении логической связности при генерации текста. Точность модели составила 0.73%, что указывает на необходимость увеличения длины N-граммы или сравнения с моделями глубокого обучения (LSTM, Transformer) для улучшения грамматического качества сгенерированных текстов. Статья анализирует возможности и ограничения марковской модели в контексте генерации текста на узбекском языке.

Abstract. This article explores the application of the Markov model for text generation in the Uzbek language. The study analyzes the process and outcomes of text generation using a specific case of the Markov model – the trigram model. Not enough work has been done in this regard in the Uzbek language yet. Experimental results revealed limitations of the Markov model in preserving logical coherence during text generation. The model’s accuracy rate was 0.73%, highlighting the need

to either increase the N-gram length or compare it with deep learning architectures (e.g., LSTM, Transformer) to improve the grammatical quality of generated texts. The paper critically examines the potential and limitations of Markov models for text generation in Uzbek.

Kalit so‘zlar: *Markov modeli, matn generatsiyasi, o‘zbek tili, tabiiy tilni qayta ishlash (NLP), N-gram modeli, trigram, perplekslik, aniqlik darajasi.*

Introduction

Today, the widespread use of information systems and the internet across various fields has led to the generation of massive volumes of data. Consequently, the need for automated processing and searching of multilingual data is growing rapidly. Text generation, a critical branch of natural language processing (NLP), involves creating logical and coherent text based on provided data. In recent years, advancements in deep learning technologies have revolutionized text generation, enabling the production of human-like texts with unprecedented accuracy.

Natural language processing (NLP) is one of the key directions in the field of text generation. While numerous studies have been conducted earlier for languages such as English, French, and others, there is still insufficient scientific research on text generation for the Uzbek language. The Markov model is one of the simplest and most effective approaches used in natural language processing, enabling the generation of new text based on a probabilistic chain derived from a given input. The aim of this article is to explore the potential of Markov models for text generation in Uzbek, evaluate the results, and analyze its advantages and limitations.

Literature Review

In many studies involving morphologically rich languages, deep learning models such as LSTM and Transformers have outperformed traditional n-gram approaches in terms of semantic coherence and grammatical accuracy. For example, Transformer-based architectures like BERT and GPT have shown state-of-the-art performance in text generation tasks.

Today, numerous researchers worldwide are conducting studies on automated text generation using natural language processing (NLP) tools. For instance, A.M. Kassenkhan proposed adapting two modern generative models – Diffusion models and Transformers – for text generation in the Kazakh language, demonstrating the effectiveness of Transformer models in generating paraphrased text [1]. Meanwhile, Junyi Li and Tianyi Tang presented a comprehensive analysis of using Pre-trained Language Models (PLMs) for text generation. They highlighted three key aspects of applying PLMs:

1. Encoding input expressions into forms that preserve semantic meaning and integrating them into PLMs;

2. Designing PLMs that function effectively as generative models;
3. Efficiently fine-tuning PLMs based on reference texts to ensure unique characteristics in generated outputs.

They also identified major challenges and solutions related to these aspects [2].

In text generation research, statistical and neural approaches dominate. Statistical methods, such as N-gram models, generate text based on sequential probabilities. While these methods are efficient due to their simplicity, their limited ability to handle long-range context makes achieving accurate results challenging [3], [4]For example, N-gram language models – a straightforward and intuitive approach – have been applied to text generation in Uzbek, though with notable limitations [3].

Neural approaches, on the other hand, leverage deep learning architectures [5], [6]. Transformer models like BERT and GPT, trained on large datasets, produce high-quality text by capturing contextual information. These models excel at learning semantic and syntactic features of language, which is critical for managing complex grammatical structures in Uzbek [6].

Research on the Uzbek language remains in its early stages[7]. Current efforts focus on tasks such as corpus preparation and text generation using N-gram models. Preliminary experiments adapting BERT-like models for Uzbek have shown promising results [4], [8].

Research Methodology

In this study, we analyzed the process of text generation in the Uzbek language using a Markov model.

A Markov model assumes that the next event in a sequence depends only on the current state. In text generation, this means that a word’s occurrence depends on the previous N words.

Markov models are often implemented using N-gram models. Examples include:

- Bigram (n=2): "men bugun" → "kitob"
- Trigram (n=3): "men bugun kitob" → "o‘qidim"

To build a functional Markov model, a large Uzbek text corpus is required.

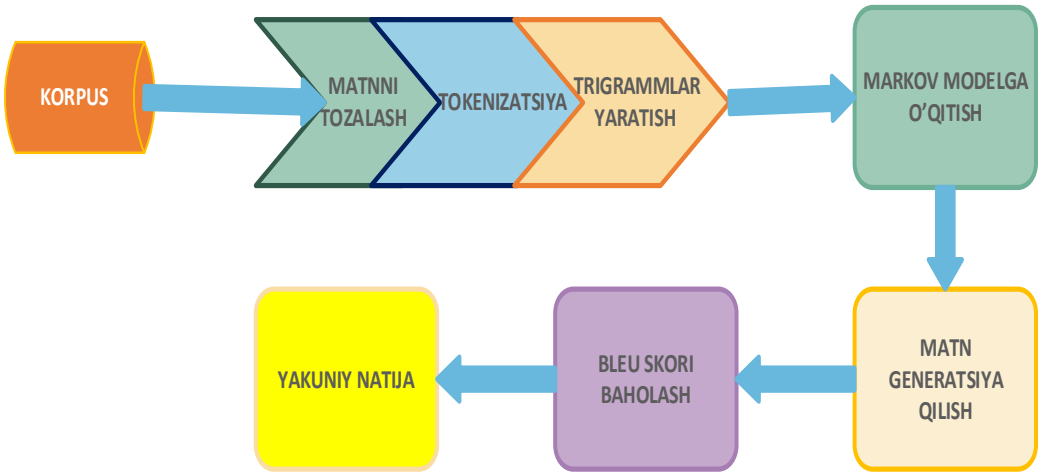


Figure 1: Stages of initial processing of language corpus texts

The following methodological approach was employed:

1. Data Collection and Preparation

- Corpus Selection: The "odilyakubov.txt" text corpus was chosen for experiments.
- Text Cleaning:
  - Removed extra spaces, special characters, and standardized uppercase/lowercase distinctions.
- Tokenization: The text was tokenized (split into words).

2. Model Training

- Model Selection: A trigram model ($n=3$) was implemented.
- Training Data: The entire corpus was used as the training set.
- Training Process: The Markov model was trained on this dataset to predict subsequent words based on probabilistic transitions.

3. Text Generation

- Input: Starting words were provided as input.
- Output: The model generated new text sequentially from the starting words.
- Evaluation: Generated texts were assessed for grammatical correctness and logical coherence.

Key Findings

- Capabilities: The Markov model demonstrated basic feasibility for generating Uzbek text.
- Limitations:
  - Struggled to preserve long-range logical coherence (e.g., topic consistency over long sentences).

○ Limited grammatical accuracy due to the model’s inability to handle complex Uzbek morphological rules.

Importance of research

• Provides a baseline for comparing traditional (Markov) and advanced (e.g., Transformer, LSTM) models in low-resource languages like Uzbek.

• Highlights the need for hybrid approaches (e.g., combining N-gram models with neural architectures) to address grammatical and contextual challenges.

Preparing the Text Corpus

For experimentation, the "odilyakubov.txt" dataset was used. The text underwent:

• Cleaning: Removed unnecessary characters and spaces.

• Normalization: Converted all text to lowercase.

• Tokenization: Split into individual words.

BLEU (Bilingual Evaluation Understudy) score

How does this method work?

• BLEU is an evaluation method used for machine translation (MT) or text generation.

• It measures how similar the text generated by the model is to the ideal text.

• The BLEU score ranges from 0 to 1, with 1 → perfect match.

$$BLUE = BP \times \exp\left(\sum_{n=1}^{N} w_n log p_n\right)$$

• BP (Brevity Penalty) – A penalty based on the length of the text.

• pnp_npn – N-gram matching level.

Disadvantages:

• Does not understand semantic meaning, only compares based on n-grams.

• Due to the large number of synonyms in the Uzbek language, sometimes incorrect evaluations are possible.

Matnni boshlash uchun 3 ta so‘z kiriting: men kitobni o‘qishim

◆ *Yangi generatsiya qilingan matn:*

men kitobni o‘qishim kerakligini esladim . . . . . .

✅ *Modelning BLEU skori bo‘yicha aniqligi: 0.73*

Analysis and Results

Model Performance

1. Text Quality: The Markov model successfully generates Uzbek text while maintaining basic syntactic structures. However, the generated text sometimes lacks semantic coherence, leading to inconsistencies in meaning.

2. Dependence on Local Context: The model makes accurate predictions within a short context but fails to capture long-range dependencies, resulting in disjointed and fragmented sentences over extended texts.

3. Repetitive Patterns: Due to the probabilistic nature of the model, certain words and phrases tend to be repeated, reducing the variability and creativity of the generated text.

Common Issues Observed

1. Logical Inconsistencies: The generated text does not always maintain a coherent structure, leading to sudden topic shifts.

2. Grammatical Errors: The model struggles with Uzbek morphology, particularly case endings and verb conjugations.

3. Word Prediction Limitations: Since the Markov model relies on fixed-length sequences, it cannot infer contextual meanings beyond its immediate N-gram scope.

Model Evaluation and Suggested Improvements

1. Increasing N-gram Size: Expanding n=4 or n=5 could help capture more context and improve text fluency.

2. Hybrid Approaches: Integrating deep learning models (LSTM, Transformer) could enhance grammatical accuracy and logical consistency.

3. Data Augmentation: A larger and more diverse corpus could improve the model’s ability to generate varied and contextually appropriate text.

4. Contextual Awareness Mechanisms: Enhancing the model with memory-based techniques or semantic embeddings could mitigate logical inconsistencies.

One possible way to address the model’s inability to capture long-range dependencies and morphological agreements is to integrate deep learning models, particularly sequence-aware architectures like LSTM and Transformer networks. These models can learn hierarchical representations and capture semantic nuances more effectively, especially in morphologically complex languages like Uzbek.

Conclusions

This study demonstrates that the Markov model offers a simple yet foundational approach for generating text in the Uzbek language. Its effectiveness lies in its ability to model short-range dependencies and maintain local syntactic structure using

probabilistic transitions between N-grams, particularly trigrams in our implementation. However, the results also reveal notable limitations that prevent the model from producing high-quality, coherent, and grammatically sound long-form text.

While the generated sentences exhibit basic fluency and partial contextual relevance, the model fails to maintain semantic consistency and long-range coherence, especially in more complex or abstract expressions. Furthermore, it struggles with morphological nuances of the Uzbek language, such as suffix handling, verb conjugation, and case agreement—factors that are critical for natural-sounding Uzbek text[9].

To address these challenges, the following strategies are recommended:

- Expanding the N-gram window size (e.g., to 4-grams or 5-grams) can improve context awareness by including more prior words in each prediction.
- Integrating deep learning models such as LSTM and Transformer architectures would allow the model to capture deeper contextual relationships and better handle the complex grammatical structure of the language.[10], [11]
- Training on larger, domain-diverse corpora can enhance the model’s ability to produce varied, topic-relevant, and semantically rich text outputs.

In addition, the study underscores the importance of incorporating evaluation metrics like BLEU to quantify output quality and guide model development.

Future Work

Future research should explore hybrid modeling approaches, combining the strengths of traditional statistical methods (like Markov chains) with the capabilities of neural language models. Such approaches can potentially offer greater accuracy, fluency, and contextual richness, particularly for low-resource languages like Uzbek where annotated data is limited. Moreover, expanding this work to domain-specific generation tasks (e.g., education, medicine, legal documents) can provide more practical applications and further refine model adaptability.

References.

1. A. M. Kassenkhan, N. K. Mukazhanov, S. Nuralykyzy, and Z. B. Kalpeyeva, “Text generation models for paraphrase on kazakh language,” *КазУТБ*, vol. 1, no. 22, Mar. 2024, doi: 10.58805/kazutb.v.1.22-249.

2. S. Haque, Z. Eberhart, A. Bansal, and C. McMillan, “Semantic Similarity Metrics for Evaluating Source Code Summarization,” in *IEEE International Conference on Program Comprehension*, IEEE Computer Society, 2022, pp. 36–47. doi: 10.1145/nnnnnnn.nnnnnnn.

3. E.B. Boltayevich. N-gramm til modellari vositasida o‘zbek tilida matn generatsiya qilish. // “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya 2022. [Online]. Available: http://compling.navoiy-uni.uz/

4. B Elov *et al.*, O‘zbek tili korpusi matnlarini qayta ishlash usullari. //Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali. Volume 1, ISSUE 3, October 2023.

5. N. Fatima, A. S. Imran, Z. Kastrati, S. M. Daudpota, and A. Soomro, “A Systematic Literature Review on Text Generation Using Deep Neural Network Models,” *IEEE Access*, vol. 10, pp. 53490–53503, 2022, doi: 10.1109/ACCESS.2022.3174108.

6. T. Iqbal and S. Qureshi, “The survey: Text generation models in deep learning,” Jun. 01, 2022, *King Saud bin Abdulaziz University*. doi: 10.1016/j.jksuci.2020.04.001.

7. K. Makhija, T. N. Ho, and E. S. Chng, “Transfer learning for punctuation prediction,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA ASC 2019*, Institute of Electrical and Electronics Engineers Inc., Nov. 2019, pp. 268–273. doi: 10.1109/APSIPAASC47483.2019.9023200.

8. M. S. Sharipov, H. S. Adinaev, and E. R. Kuriyozov, “Rule-Based Punctuation Algorithm for the Uzbek Language,” in *International Conference of Young Specialists on Micro/Nanotechnologies and Electron Devices, EDM*, 2024, pp. 2410 – 2414. doi: 10.1109/EDM61683.2024.10615061.

9. S. Maksud, K. Elmurod, Y. Ollabergan, and S. Ogabek, “UzbekVerbDetection: Rule-based Detection of Verbs in Uzbek Texts,” in *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 - Main Conference Proceedings*, 2024, pp. 17343 – 17347. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85195974039&partnerID=40&md5=ff3003ea2644833f3dc45277429459f7

10. Z. Y. Peng and P. C. Guo, “A Data Organization Method for LSTM and Transformer When Predicting Chinese Banking Stock Prices,” *Discrete Dyn Nat Soc*, vol. 2022, 2022, doi: 10.1155/2022/7119678.

11. N. R. Bhowmik, M. Arifuzzaman, and M. R. H. Mondal, “Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms,” *Array*, vol. 13, 2022, doi: 10.1016/j.array.2021.100123.

YORDAMCHI SO‘ZLARNING QO‘LLANISH DAVRIGA KO‘RA SARALOVCHI LINGVISTIK TA’MINOT VOSITASINI TUZISH MASALALARI.

Shahlo Hamroyeva Mirdjanovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Qodirova Madinabonu Murodjon qizi,
Tayanch doktorant
begonammmm@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur maqolada o‘zbek tilidagi yordamchi so‘zlarning tarixiy rivojlanishi, ularning turli davrlardagi qo‘llanilishi va o‘zgarish jarayonlari tahlil qilinadi. O‘zbek tilining shakllanishidan boshlab hozirgi kungacha bo‘lgan davr mobaynida yordamchi so‘zlarning ma’lum qismining iste’moldan chiqishi, ayrimlarining esa yangi shakllarda paydo bo‘lishi lingvistik va ijtimoiy omillar bilan bog‘liq ekani ko‘rsatiladi. Maqolada arab, fors va Yevropa tillarining ta’siri ostida yordamchi so‘zlar tizimining qanday o‘zgarishlarga uchragani tarixiy manbalar asosida o‘rganilgan. Ushbu tadqiqot natijalari mashina tarjimasi, tarixiy tilshunoslik hamda zamonaviy leksikografiya sohalarida qo‘llanilishi mumkin.

Abstract. This article analyzes the historical development of auxiliary words in the Uzbek language, their use in different periods and the processes of change. It is shown that the disappearance of a certain part of auxiliary words from use, and the emergence of some in new forms during the period from the formation of the Uzbek language to the present day, is associated with linguistic and social factors. The article studies the changes in the system of auxiliary words under the influence of Arabic, Persian and European languages, based on historical sources. The results of this study can be used in the fields of machine translation, historical linguistics and modern lexicography.

Аннотация. В данной статье анализируется историческое развитие вспомогательных слов в узбекском языке, их использование в разные периоды и процессы изменения. Показано, что за период от формирования узбекского языка до наших дней часть вспомогательных слов вышла из употребления, а часть из них появляется в новых формах вследствие языковых и социальных факторов. В статье на основе исторических источников рассматривается, как претерпевала изменения система вспомогательных слов под влиянием арабского, персидского и европейских языков. Результаты исследования могут быть использованы в области машинного перевода, исторической лингвистики и современной лексикографии.

Kalit so‘zlar: *grammatik shakl, grammatik ma’no, til tarixi, leksik o‘zgarishlar, grammatik evolyutsiya, mashina tarjimasi, lingvistik davrlashtirish.*

KIRISH

Til tarixiy taraqqiyot jarayonida doimiy ravishda o‘zgarib boradi. Bu o‘zgarishlar ayniqsa, lug‘at tarkibida, xususan, yordamchi so‘zlar tizimida yaqqol namoyon bo‘ladi. O‘zbek tilida yordamchi so‘zlar turli davrlarda turlicha qamrov va shakllarga ega bo‘lib, arab, fors va Yevropa tillarining ta’siri ostida muayyan o‘zgarishlarga uchragan. Tarixiy manbalar shuni ko‘rsatadiki, ba’zi yordamchi so‘zlar asta-sekin iste’moldan chiqib ketgan, ba’zilari esa yangi shakllarda tilimizga singib borgan.

Matn tarjimasi jarayonida so‘zlarning tarixiy qo‘llanilish davriga e’tibor qaratish muhim sanaladi. XX asr boshlariga oid matnni bugungi kunda tarjima qilishda, o‘sha davrda qo‘llangan yordamchi so‘zlardan foydalanish matnning asliyatga yaqin bo‘lishini ta’minlaydi. Shu sababli, o‘zbek tilining turli davrlarda ishlatilgan yordamchi so‘zlarini lingvistik jihatdan tadqiq qilish va ularni davrlashtirish dolzarb vazifa hisoblanadi[1]. Ushbu maqolada o‘zbek tilidagi yordamchi so‘zlarning tarixiy rivoji, ularning davrlarga ajratilishi va til tizimidagi o‘rni tahlil qilinadi.

Bir tildagi matnni boshqa tilga tarjima qilishda uning tarkibidagi so‘zlarning qo‘llanilish davriga e’tibor qaratish muhimdir. Deylik, XX asr boshlarida yozilgan matnni boshqa tilga o‘girish uchun maqsadli tildan ham XX asr boshlarida qo‘llangan so‘zlar bazasiga murojaat qilishga to‘g‘ri keladi. O‘zbek tilidagi so‘zlarning ma’lum davrlarda qo‘llanishini aks ettiruvchi lingvistik ta’minot vositasini tuzish shu sababli mashina tarjimasi uchun muhimdir.

O‘zbek tili leksikasi, xususan, yordamchi so‘zlar tizimi taraqqiyotning turli davrlarida turlicha qamrovga ega bo‘lgan. Arab, fors va yevropa tillarining ta’sirida yordamchi so‘zlar tizimida turli o‘zgarishlar kuzatilgan: yangi so‘zlar qabul qilingan, ma’lum so‘zlar iste’moldan chiqqan yoki shakliy o‘zgarishlarga uchragan. Ularni tahlil qilish mashina tarjimasi uchun muhimdir.

O‘zbek tili tarixiyligi jihatidan turlicha tasnif qilinadi.
Tilshunos olima R.Sayfullayeva o‘zbek tilini quyidagicha davrlashtiradi[2]:
V – XIV asrlar: o‘zbek tilining ilk taraqqiyot davri
XIV – XIX asrlar: o‘zbek tilining takomillashish davri
XIX – XX asrlar: milliy adabiy tilning shakllanish davri
XX – XXI asrlar: o‘zbek milliy adabiy tili davri.

Shuningdek, ayrim olimlar o‘zbek tili tarixini quyidagicha davrlashtirishadi:

1. Qadimiy o‘zbek tili: turkiy tildan o‘zbek tilining ajralib chiqishi va o‘zbek atamasining paydo bo‘lishigacha bo‘lgan davr.

2. XV-XIX asrlar: bu bosqichda o‘zbek tilining madaniy, ijtimoiy va siyosiy rivojlanishi o‘rganiladi.

3. XX asrning boshlari: Bu davrda o‘zbek tilining turli jihatlarida o‘zgarishlar va rivojlanishlar ko‘p o‘rganiladi. Rasmiy til, adabiyot til, axborot tilining o‘zgarishlari va modernizatsiyasi bu davrda katta ahamiyatga ega.

4. XX asrning o‘rtasi -1991: Sovet davri

5. 1991-yildan hozirgacha: O‘zbek tilining mustaqil davrida o‘zgarishlar, yangi so‘zlar, yangi ifodalarni yaratish, adabiyotning rivojlanishi davri [3].

Shu va shu kabi manbalarga asoslangan holda va o‘zbek tilidagi grammatik shakllarining o‘ziga xosliklarini hisobga olib, yordamchi til birliklarini davriy jihatdan quyidagicha guruhlab lingvistik bazaga kiritishni taklif etamiz:

- V –XIV asrlar: o‘zbek tilining shakllanish davri
- XV – XX asrlar: o‘zbek tilining takomillashuvi va milliy adabiy tilning shakllanish davri
- XX asr oxiridan – hozirgi kungacha: o‘zbek milliy adabiy tili davri.

Tilimizdagi yordamchi til birliklari sinonimik qatorlaridagi so‘zlar ushbu uchdavrga oid-oid emasligiga ko‘ra ham farqlanadi.

Shunga ko‘ra so‘zlarni davrlashtirishda har bir davrga oid manba tanlab olindi va yordamchi so‘zlarning qo‘llanishi tahlil qilindi. Chunonchi, o‘zbek tilining shakllanish davrida yordamchi so‘zlarning qo‘llanishini o‘rganish uchun Rabg‘uziyning “Qissasi Rabg‘uziy”, “Tafsir” asarlari, XV–XX asrga oid so‘zlarni tekshirish uchun “Boburnoma”, “Shajarayi turk”, Muqimiy va Abdurauf Fitrat asarlari, o‘zbek milliy adabiy tili davri uchun G‘.G‘ulomning “Shum bola”, U.Hamdamning “Muvozanat” asarlari tanlab olindi.

Shart ma’nosini bildirgan agar so‘zi XI–XII asrlarga oid yodgorliklarda ham iste’molda bo‘lgan. Lekin uning qo‘llanishi XIII-XIV asrlardan boshlab kengaya borgan. Uning agarche shakli XV asrdan boshlab keng qo‘llangan bo‘lib, XIII–IV asrlarga oid yodgorliklarning ayrimlarida ham uchrab turadi. Gar shakli XIII–XIV asrlardan XX asrgacha faol, so‘ng faqat she’riy asarlarda ishlatilgan. Shuningdek, XVIII–XIX asrlarda garchand shakli iste’molda bo‘lgan. XVI–XVIII asrlarda harchand shaklida ham ishlatilgan. XIX asrdan boshlab esa agarda va garchi shaklida ham qo‘llana boshlagan. Qolaversa, XI–XII asrlardan boshlab magar shakli ham ishlatilgan va XVIII–XIX asrlargacha iste’molda bo‘lgan. Basharti so‘zi ham XIII–XIX asrlarda shart ma’nosini ifodalagan.

Masalan,

Agar qolaman desalaring, oldi qish, ovqat-oziqlaring bemalol. (G‘.G‘ulom)

Agarche mardanarag‘i bar edi, vale bir nima telbarak va go‘l kishi edi (BN).

Tüynügiydin agarda chiqsa tutun

Munda ber dastmáyalarin butun (Muqimiy).

In’om bo‘ldi sizga garchand Toshkantdin

Bergaymu-dik deganda Shayxantaxur fanus (Muqimiy).

Garchi pochcham bilan katta opamlar o‘qituvchilik qiladigan maktab dadamlar ishlaydigan mahkamadan xiyla narida bo‘lsa-da, avval o‘sha tomonga o‘tdim. (O.Yoqubov)

Gar qilich báshima ham kelsá, degáymen rástini (Muqimiy).

Hozirgi kunda iste’molda bo‘lgan –mi, -chi, -a, -ya so‘roq yuklamalaridan –mi yuklamasi, ham, hatto, hattoki, nahot, nahotki, axir, g‘irt, -(y)oq kuchaytiruv-ta’kid yuklamalaridan ham, hatto, nahotki; faqat, atigi, xolos, -gina ayiruv-chegaralov yuklamalaridan faqat; hech, sira, na…na inkor yuklamalaridan hech; o‘xshatish-qiyoslash yuklamalaridan xuddi o‘rganilayotgan barcha davrlarda mavjud bo‘lgan. Go‘yo(ki), na, bamisli, misoli, axir, hatto yuklamalari XVI – XVII asrlardan boshlab kirib kelgan bo‘lsa, lim, liq, jiqqa, qip, nim kabi belgini kuchliligini bildiruvchi yuklamalar, -chi, -a, -ya so‘roq yuklamalari, sira, also inkor yuklamalari, aynan o‘xshatish yuklamasi tilimizga XIX asrdan boshlab kirib kelgan. A bog‘lovchisi XX asrdan boshlab qo‘llana boshladi.

Qovun tanlashda didim yo‘q. A siz-chi?

Zidlik ma’nosini bildiruvchi bog‘lovchilardan ammo, lekin bog‘lovchilari XI asrdan ishlatila boshlagan bo‘lsa, biroq so‘zi XVIII asr oxiri – XIX asr boshlaridan boshlab asarlarda ko‘zga tashlana boshladi. XVI – XIX asrlargacha balkim shaklida ishlatilgan balki so‘zining bu shaklda ishlatilishi XX asrdan boshlab kuzatiladi[1].

Eski turkiy til davri (XI asr)dan boshlab hozirgacha faol ishlatilib kelinuvchi yordamchi so‘zlarga ammo, atab, avval, ba’zan, beri, bo‘yi, bo‘ylab, -dir, esa, gina, ham, ichida, kabi, keyin, lekin, oldin, orqali, orqasidan, osha, qarab, -sa, sari, uchun, tomon, tepasida, va, yonida so‘zlari hisoblanadi.

Birgalik ma’nosini ifodalagan bilan so‘zi qadimdan mavjud bo‘lib turli variantlarda ishlatilib kelingan.

bila	qadimdan XIII - XIV asrlargacha	Men qolurman dod bila, faryod bila
birla	qadimdan XIV - XV asrlargacha	Kuchuk birla xo‘tukni qancha qilma tarbiyat (Navoiy)

birlan	XVI—XIX asrlar	Siz birlan faxr eturmiz
ila	XV asrdan XVIII asrgacha	Jan hasrat ila kuydi (Lutfiy)
bilan	XIX asrdan boshlab	Sen bilan safimiz yana zich, azim.

Mustasnolik ma’nosini bildirgan bo‘lak, tashqari so‘zlarining ko‘makchi vazifasida qo‘llanishi nisbatan keyingi davrlarga to‘g‘ri keladi. Ushbu ma’noni ifodalash uchun qadimgi turkiy tilda azin, adin so‘zlari ishlatilgan bo‘lsa, XV–XVI asrlardan boshlab ayru yordamchisi ishlatila boshlandi. Boshqa, o‘zga so‘zlari nisbatan avvalroq: XIII–XIV asrlarga oid ayrim manbalarda uchraydi, biroq XV asrdan kengaya boshlagan.

O‘zbek tili taraqqiyoti davomida ayrim yordamchi so‘zlar butunlay iste’moldan chiqqan. Ular:

ashnu	varna	tabaru
taqi	usk	ar
dag‘i	asra	ayru
ma	burun	kezin
o‘tru	burna	tegru
azin	bikin	la
azu	taba	

Ayrimlari taraqqiyot davomida fonetik o‘zgarishlarga uchragan:

Tovush tushishi	**Tovush almashishi**	**Tovush ortishi**	**Ikki so‘zning qo‘shilishi**
magar => agar	tuba =>teba => tepa	uza => uzra	to + ki
agarche => garchi	kezin =>kedin => keyin	chun =>uchun	chun + ki
birlan => bilan	kibi => kabi	bila => birla => birlan	yo + ki
chunkim =>chunki	alt => old	o‘ng => so‘ng	va + gar + na
Balkim => balki	teb => deb		
	harchand => garchand		

Ba’zi yordamchi so‘zlar tilimizga XIX asrdan boshlab kirib kelgan:

Umuman olganda, yordamchi so‘zlarni tadqiq qilish davomida ularning o‘zbek tiliga kirib kelishi, ayrim o‘zgarishlarga uchrashi, iste’moldan chiqishi ma’lum lisoniy va nolisoniy omillarga bog‘liq ekanligi ko‘zga tashlanadi. Bu omillarga quyidagilar kiradi:

XI asrdan Sharq Uyg‘onish davri boshlanib, arab tili va madaniyatining keng yoyilishi natijasida o‘zbek tilining lug‘at sathi boyidi, arabcha o‘zlashmalar kirib

kela boshladi.Qo‘shni mamlakatlar bilan yaqin qo‘shnichilik aloqalari sababli forscha o‘zlashmalar ham o‘zlashtirildi[4].

XV – XVIII asrlarda Temuriylar davridan boshlangan ilm-fan, adabiyot, madaniyat rivoji til rivojiga ulkan hissa qo‘shdi. Tilda yordamchi so‘zlardan foydalanish kengaydi. Keyingi xonliklar davrida esa a Aniq va tabiiy fanlar rivoji susaygan bo‘lsa-da, tarix, adabiyot sohalari rivojlangan. Tilga yangi o‘zlashmalar kirib keldi, ayrimlari iste’moldan chiqdi. Jumlalarda yordamchi so‘zlarning qo‘llanishi kengaydi.

XIX asr – podsho Rossiyasi bilan aloqalar kuchayishi va Markaziy Osiyoning istilo qilinishi bilan Yevropa tillaridan o‘zlashmalar kirib kela boshladi.

XX asr – Yevropa tillari bilan aloqalarning kuchayishi natijasida yordamchi so‘zlarning qo‘llanishi va ma’nolarining kengayishi kuzatiladi.

So‘zlarning o‘zbek tilida qo‘llanishini davrlashtirishda shu omillar hisobga olindi va quyidagicha guruhlandi:

XI -XII	XIII-XIV	XV-XVIII	XIX	XX
		tortib	tortib	tortib
tub	tub	tub	tub	
			tufayli	tufayli
-u (-yu)	-u (-yu)	-u (-yu)	-u (-yu)	-u (-yu)
uchun	uchun	uchun	uchun	uchun
ustida	ustida	ustida	ustida	ustida
		uzra	uzra	uzra
va	va	va	va	va
	valekin	valekin	valekin	valekin
		xoh	xoh	xoh
			xolos	xolos

XULOSA

Xulosa sifatida aytish mumkinki, o‘zbek tilidagi yordamchi so‘zlarning tarixiy taraqqiyoti murakkab va uzoq jarayon bo‘lib, turli davrlarda ularning shaklanishi, qo‘llanishi va iste’moldan chiqishi muayyan lingvistik va nolisoniy omillarga bog‘liq bo‘lgan. Til rivojiga Sharq Uyg‘onish davridan boshlab arab va fors tillarining ta’siri, XV-XVIII asrlardagi adabiy va ilmiy taraqqiyot, XIX asrda Rossiya bilan aloqalarning kuchayishi hamda XX asrdagi modernizatsiya jarayonlari sezilarli darajada ta’sir ko‘rsatgan.

Yordamchi so‘zlar tarixiy manbalar orqali o‘rganilib, ularning davriy tasnifi taklif etildi. O‘zbek tilining turli davrlarida ma’lum yordamchi so‘zlarning paydo bo‘lishi yoki iste’moldan chiqishi aniqlanib, ularning semantik va grammatik jihatdan o‘zgarishlari yoritildi. Ushbu tadqiqot natijalari kelgusida mashina tarjimasi

va tarixiy tilshunoslik sohalarida qo‘llanilishi mumkin. Shu bilan birga, tilning doimiy rivojlanishini hisobga olgan holda, yordamchi so‘zlar tizimidagi o‘zgarishlarni yanada chuqurroq tahlil qilish ham dolzarb masala bo‘lib qoladi.

Foydalanilgan adabiyotlar:

1. Abdurahmonov G‘. O‘zbek tilining tarixiy grammatikasi. – Toshkent: Fan, 2005.
2. Sayfullayeva R. R., Mengliyev B. R., Boqiyeva G. H., Qurbonova M. M., Yunusova Z. Q., Abuzalova M. Q. Hozirgi o‘zbek adabiy tili Toshkent-2010
3. Karimov S. O‘zbek tilining shakllanish tarixi. – Samarqand: Samarqand Davlat universiteti nashriyoti, 2012.
4. Qosimov U. O‘zbek tili tarixiy leksikologiyasi. – Toshkent: Fan, 1987.
5. Rahmatullayev Sh Hozirgi adabiy o‘zbek tili TOSHKENT-2006
6. Navoiy A. Mahbub ul-qulub. – Toshkent: G‘afur G‘ulom nomidagi Adabiyot va san’at nashriyoti, 1993.
7. Bobur Z.M. Boburnoma. – Toshkent: Fan, 2008.
8. Fitrat A. O‘zbek tili grammatikasi. – Toshkent: O‘qituvchi, 1929.
9. G‘ulom G‘. Shum bola. – Toshkent: Yangi asr avlodi, 2015.
10. Yoqubov O. Tilla baldoq. – Toshkent: Sharq, 2001.
11. Hamdam U. Muvozanat. – Toshkent: Ma’naviyat, 2016.

UDK: 81'33

NER: TEGLASH MUOMMOLARIGA YECHIMLAR

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent,
elov@navoiy-uni.uz
ToshDO‘TAU

Madina Samatboyeva To‘lqinjon qizi,
tayanch doktorant
samatboyevamadina@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. O‘zbek tili korpusida NER obyektlarini teglash jarayonidagi asosiy muammolar va ularning mumkin bo‘lgan yechimlari tahlil qilish muhimdir. Maqolada, o‘zbek tilidagi NER tizimini rivojlantirishda duch kelinadigan asosiy muammolar – tegishli korpus yetishmovchiligi, so‘z shakllarining ko‘p variantliligi, ommaviy ochiq NER ma’lumotlar bazasining kamligi va semantik noaniqlik kabi jihatlar yoritiladi.

Shuningdek, maqolada ushbu muammolarni hal qilish uchun zamonaviy texnologiyalar, xususan, mashinani o‘rganish (Machine Learning), chuqur o‘rganish (Deep Learning) va tabiiy tilni qayta ishlashga oid ilg‘or metodlar tahlil qilinadi. Bundan tashqari, o‘zbek tiliga moslashtirilgan, maxsus annotatsiyalangan katta hajmdagi korpus va uni xalqaro standartlarga moslashtirish muhim ekani qayd etiladi.

Abstract. It is important to analyze the main challenges in the process of tagging NER entities in the Uzbek language corpus and their possible solutions. The article highlights the main challenges in developing an NER system for the Uzbek language, including the lack of an appropriate corpus, the high variability of word forms, the scarcity of publicly available NER databases, and semantic ambiguity.

Additionally, the article analyzes modern technologies used to address these challenges, particularly machine learning, deep learning, and advanced natural language processing (NLP) methods. Furthermore, it emphasizes the importance of creating a large, specifically annotated corpus adapted for the Uzbek language and aligning it with international standards.

Аннотация: Важно проанализировать основные проблемы процесса разметки объектов NER в корпусе узбекского языка и возможные пути их решения. В статье рассматриваются основные проблемы, возникающие при развитии системы NER для узбекского языка, такие как нехватка соответствующего корпуса, наличие множества вариантов словоформ,

ограниченное количество общедоступных баз данных NER и семантическая неопределенность.

Кроме того, в статье анализируются современные технологии, используемые для решения этих проблем, в частности, методы машинного обучения (Machine Learning), глубокого обучения (Deep Learning) и обработки естественного языка (NLP). Также подчеркивается важность создания большого аннотированного корпуса, адаптированного для узбекского языка, и его соответствия международным стандартам.

Kalit so‘zlar: *NER, nomlangan obyekt, ot, atoqli ot, NLP, teg, annotatsiya.*

KIRISH

Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida o‘zbek tili korpusida nomli obyektlarni aniqlash (Named Entity Recognition, NER) dolzarb muammolardan biri hisoblanadi. O‘zbek tili morfologik jihatdan murakkab va boy til bo‘lgani uchun NER obyektlarini aniqlash va to‘g‘ri teglash jarayoni qiyinchilik tug‘diradi. Shuningdek, xalqaro tajribalarni o‘rganish va o‘zbek tiliga moslashtirilgan, aniq va mustahkam ishlaydigan NER tizimini yaratish, annotatsiya jarayonida yaratilgan ma’lumotlar to‘plamining sifatini oshirish, tokenizatsiya va morfologik analiz orqali ma’lumotlarni yaxshiroq qayta ishlash kabi metodlar muhim hisoblanadi. Shu bilan birga, o‘zbek tili uchun maxsus tayyorlangan lug‘atlar va semantik tahlil vositalaridan foydalanish NER tizimining aniq va samarali ishlashini ta’minlashi mumkin.

Named Entity Recognition (NER) – bu tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasining bir bo‘limi bo‘lib, matndagi aniq va muhim ma’lumotlarni aniqlash va ularni tasniflashga qaratilgan. NER matndagi **nomlangan obyektlar** (ya’ni shaxs nomlari, joy nomlari, tashkilotlar va h.k) **so‘z, ibora yoki so‘zlar ketma-ketligi** shaklida uchraydigan elementlarni topadi va ularni oldindan belgilangan toifalarga ajratadi [1].

NER texnologiyasi zamonaviy matn tahlili tizimlarining ajralmas qismi bo‘lib, jahon miqyosida chuqur o‘rganilgan. O‘zbek tilida bu boradagi izlanishlar endigina jadallashmoqda va global tajribalarga asoslanib, mahalliy xususiyatlarga moslashtirilgan modellar ishlab chiqilmoqda. Bu esa o‘zbek tilidagi sun’iy intellekt tizimlarini rivojlantirishda muhim qadam hisoblanadi.

#Misol:

“Xalq deputatlari Toshkent viloyati Bekobod tumani Kengashining navbatdan tashqari sessiyasida Zaynilobiddin Shahobiddinovich Nizomiddinov tuman hokimi lavozimiga tayinlandi.” [2]

Bu jumladagi NER orqali quyidagilar aniqlanadi:

- **[Toshkent] (viloyati)** — *Joy nomi* (Location)
- **[Bekobod] (tumani)** — *Joy nomi* (Location)
- **[Zaynilobiddin] [Shahobiddinovich] [Nizomiddinov]** – *Shaxs* (Person)

ADABIYOTLAR SHARHI

NER texnologiyalari jahon ilmiy adabiyotida ko‘plab olimlar tomonidan keng o‘rganilgan va muhokama qilingan. Olimalarning fikriga ko‘ra, NER tizimlarini muvaffaqiyatli amalga oshirish uchun morfologik, sintaktik, semantik va kontekstual tahlilning hammasi o‘rganilishi kerak. Har bir olim NERni turli nuqtai nazardan ko‘rib chiqadi, ammo umumiy fikr shuki, bu tizimlar kelajakda tabiiy tilni qayta ishlash (NLP) sohasida yanada rivojlanib, murakkab tahlil imkoniyatlarini yaratadi.

Jehangir, S. Radhakrishnan, R. Agarwal kabi hind olimlari o‘zlarining 2023-yilda nashr etilgan “**A Survey on Named Entity Recognition — Datasets, Tools, and Challenges”** nomli maqolasida [3] qoidalarga asoslangan yondashuvlar to‘g‘risida kengroq tahlil jarayonlarini o‘tkazadilar va amqolada o‘z fikrlarini berib o‘tadilar. Shuninghdek, ushbu maqolada NER sohasidagi turli metodologiyalar, jumladan, qoidaga asoslangan yondashuvlar, ML (Machine Learning) va DL (Deep Learning) usullari tahlil qilinadi. Mualliflar NERning rivojlanishi uchun mavjud bo‘lgan ma’lumotlar to‘plamlari va vositalarni ko‘rib chiqib, sohadagi asosiy muammolarni muhokama qiladilar.

Xitoylik olimlarning “Learn and Review: Enhancing Continual Named Entity Recognition” (U Xia, Quan Wang, Yajuan Lyu, Yong Zhu, Wenhao Wu, Sujian Li, Dai Dai - 2022) nomli maqolasida NER tizimlarining uzluksiz o‘rganish qobiliyatini oshirish uchun “Learn-and-Review” (L&R) deb nomlangan ikki bosqichli yangi modelni taklif etadilar. Ushbu yondashuv avvalgi qoidalarni yangi ma’lumotlar bilan birlashtirish orqali NER tizimlarining samaradorligini oshirishga qaratilgan [4].

Annotatsiyalangan ma’lumotlar, albatta, juda muhim baza hisoblanadi. Annotatsiya jarayonidan o‘tgan matngini mahsus modellar asosida teglanadi. Dastur ishlash jarayoni mukammal bo‘lishi uchun annotatsiya qilingan matnlar muhimdir. Imed Keraghel, Stanislas Morbieu, Mohamed Nadifning **“A Survey on Recent Advances in Named Entity Recognition”** (2023) nomli tadqiqot ishida NER sohasidagi so‘nggi yutuqlar, xususan, transformer va til modellari (LLM) asosidagi yondashuvlar tahlil qilinadi [5]. Mualliflar, shuningdek, annotatsiyalangan ma’lumotlar yetishmovchiligi sharoitida NER tizimlarini rivojlantirish usullarini ham ko‘rib chiqadilar.

Sakher Khalil Alqaaidi, Elika Bozorgi, Afsaneh Shams, Krzysztof Kochut kabi olimlarning **“A Few-Shot Learning Focused Survey on Recent Named Entity**

Recognition and Relation Classification Methods” (2023) maqolasida NER va munosabatlarni tasniflash sohasidagi so‘nggi chuqur o‘rganish modellari, ayniqsa, “few-shot learning” yondashuvlariga e'tibor qaratiladi. Tadqiqotchilar kam miqdordagi ma’lumotlar bilan samarali NER tizimlarini yaratish usullarini tahlil qiladilar [6].

ASOSIY QISM

“Teg” (ingl. tags) – matndagi ishchi izoh, u matn haqidagi ma’lumotni qamrab oladi. Korpus yordamida statistik hisob jarayonida tilimizda mavjud so‘zlarning faqat chastotasini aniqlash emas, balki yana bir qancha ma’lumotlarni olishimiz mumkin. Masalan, har bir so‘z bilan bilan yonma-yon uning turkumi belgilangan bo‘lsa, tilda turli nutqiy vaziyatda so‘z turkumlarining qo‘llanilish darajasini aniqlash ham mumkin. Lingvistik teglash har bir so‘zning ma’lum kodga ega bo‘lishi bilan xarakterlanadi. Ushbu kod teg, so‘zni kodlash esa tegging (ingl.tagging) [7] deyiladi.

NER tizimlari tomonidan ishlatiladigan teglar (labels) quyidagi asosiy turlarga bo‘linadi (Qarang: 1-rasm):

- **PER** (Person) – Shaxs nomlari uchun
- **LOC** (Location) – Geografik joylar uchun
- **ORG** (Organization) – Tashkilotlar nomlari uchun
- **MISC** – Boshqa nomlangan obyektlar uchun, masalan, mahsulot nomlari yoki boshqa aniq obyektlar.

1-rasm. NERlarni matn tarkibida teglar bilan belgilash jarayoni [8]

Teglash atamasi bir vaqtning o‘zida *annotatsiya* atamasi bilan yonma-yon qo‘llaniladi.

Annotatsiya – bu ma’lumotlarni qayta ishlash uchun kerakli teglashlarni amalga oshirish jarayonidir. O‘zbek tilida annotatsiya qilishda quyidagi amallar bajariladi:

- Matndagi nomlangan obyektlarni izlash.
- Ularni teglar bilan belgilash.

• Teglarni yuqori aniqlik bilan joylashtirish uchun ma’lum metodologiyalarni qo‘llashdir.

Ba’zi adabiyotlarda annotatsiya termini razmetkalash leksikasi bilan yonma-yon ham ishlatiladi. Bir vaqtning o‘zida annotatsiyalash bu – teglash jarayoni deb ham tushuniladi.

Korpusni teglash (annotatsiyalash) dasturlashtirilgan yo‘llar bilan amalga oshiriladi. Bunda, avvalo, vaqtni tejash, mehnatni kamaytirish nazarda tutilsa, ikkinchidin, matnga avtomatik ishlov berish muammosiga yechim topiladi [9].

O‘zbek tilida NER tizimlarini samarali ishlatish uchun, asosan, mashinali o‘qitish (ML) va chuqur o‘rganish (DL) metodlarini qo‘llash, maxsus korpuslar yaratish va annotatsiya qilish jarayonida yetarli darajada sifatli va keng qamrovli ma’lumotlar to‘plash metodologiyalarni qo‘llash tavsiya etiladi.

Korpusda NERlarni teglashda quyidagi jihatlarga alohida e’tibor qaratish lozim:

1. Korpusda tozalangan, turli uslub matnlari (baza) mavjudligiga;
2. Korpus matnlaridan erkin foydalanishga;
3. Korpusdagi matnlarda ishtirok etgan boshqa birliklarning teglariga e’tibor qaratish;
4. Boshqa birliklarga biriktirilgan maxsus teglarning qaytarilmasligi (bir xil bo‘lmasligi);
5. NERni teglashda ularni, avvalo, matndan to‘g‘ri aniqlash;
6. Aniqlangan NERni to‘g‘ri klassifikatsiya qilish;
7. NERlarning omonimiya holatiga e’tibor qaratish lozim.

Shuningdek, NERlarni teglash muammolarini hal qilish uchun bir nechta asosiy yo‘nalishlar taklif etiladi:

1. **Annotatsiyalangan katta hajmdagi korpus yaratish** – O‘zbek tiliga xos NER tizimini rivojlantirish uchun katta hajmdagi, aniq va tuzilmali ma’lumotlar to‘plamini yaratish zarur.

2. **Transformer modellaridan foydalanish** – BERT, CNN, LSTM kabi ilg‘or chuqur o‘rganish (DL) metodlarini o‘zbek tiliga moslashtirish va ushbu modellar yordamida NER obyektlarini aniq ajratish texnikalarini rivojlantirish.

3. **Morfologik va sintaktik tahlilni kuchaytirish** – So‘zlarning turli shakllarda qo‘llanilishi sababli tokenizatsiya va lemmatizatsiya kabi metodlardan foydalanish NER tizimining samaradorligini oshirishga yordam beradi.

4. **Ochiq ma’lumotlar bazalarini kengaytirish** – O‘zbek tilidagi ochiq kodli NER korpuslarini yaratish va ulardan ilmiy hamda amaliy loyihalarda foydalanish orqali tahlillarni takomillashtirish.

5. **O‘zbek tiliga xos maxsus lug‘atlar bazasini ishlab chiqish** – NER tizimining aniq va samarali ishlashi uchun maxsus lug‘at va ma’lumotlar to‘plamlarini ishlab chiqish zarur.

XULOSA

O‘zbek tilida NER teglash tizimlarini ishlab chiqish va rivojlantirish katta imkoniyatlarga ega, lekin bir qancha muammolarni ham o‘z ichiga oladi. Morfologik va sintaktik xususiyatlar, kontekstni to‘g‘ri tahlil qilish, resurslarning yetishmasligi kabi qiyinchiliklarga qaramay, ilg‘or texnologiyalar va innovatsion yondashuvlar yordamida bu masalalarni hal qilish mumkin. NER tizimlarini takomillashtirish va yangi resurslar yaratish o‘zbek tilini kompyuter yordamida qayta ishlash sohasida katta yutuqlarga erishishga imkon beradi.

Maqolada O‘zbek tili korpusida NER obyektlarini teglash jarayonida duch kelinadigan asosiy muammolar va ularni hal etish yo‘llari atroflicha yoritildi. O‘zbek tilining morfologik va sintaktik murakkabligi, tegishli annotatsiyalangan ma’lumotlar bazasining yetishmovchiligi va ochiq manbalardan foydalanish imkoniyatlarining cheklanganligi NER tizimlarini yaratish jarayonini qiyinlashtiruvchi omillar sifatida e’tirof etilgan.

Maqolada ilgari surilgan yechimlar O‘zbek tilida NER tizimlarini rivojlantirish va tabiiy tilni qayta ishlash bo‘yicha global standartlarga mos keluvchi model yaratish imkonini beradi. Ushbu yo‘nalishdagi tadqiqotlar va amaliy tajribalar davom ettirilsa, kelajakda o‘zbek tili uchun yuqori samarali NER tizimi ishlab chiqish imkoniyati yuzaga keladi.

Foydalanilgan adabiyotlar ro‘yhati:

1. https://www.altexsoft.com/blog/named-entity-recognition/
2. https://qalampir.uz/news/prezident-administratsiyasining-sobik-ra%D2%B3bari-tuman-%D2%B3okimi-buldi-117080
3. Jehangir, S. Radhakrishnan, R. Agarwal. **A Survey on Named Entity Recognition – Datasets, Tools, and Challenges. 2023. P-3**
4. U Xia, Quan Wang, Yajuan Lyu, Yong Zhu, Wenhao Wu, Sujian Li, Dai Dai. **Learn and Review: Enhancing Continual Named Entity Recognition.** 2022. P-5
5. Keraghel, Stanislas Morbieu, Mohamed Nadifning. A Survey on Recent Advances in Named Entity Recognition. 2023. P-12

6. Sakher Khalil Alqaaidi, Elika Bozorgi, Afsaneh Shams, Krzysztof Kochut kabi olimlarning. A Few-Shot Learning Focused Survey on Recent Named Entity Recognition and Relation Classification Methods. 2023. P-10

7. https://rykov-cl.narod.ru/c.html

8. https://kun.uz/news/2025/03/13/ozbekiston-va-fransiya-strategik-sherik-boldi-maqsadlar-va-manfaatlar-nima

9. V. Zaxarov, B. Mengliyev, Sh.Hamroyeva. Korpus lingvistikasi: korpus tuzish va undan foydalanish, o‘quv qo‘llanma. Toshkent. 2021. 139

UO‘K 811.512.133'32

O‘ZBEK TILIDAGI MATNLAR KOREFERENSIYASINI AVTOMATIK ANIQLASH TIZIMINING LINGVISTIK TA’MINOTINI SHAKLLANTIRISH XUSUSIDA

Shahlo Abdisalomova Abdimurod qizi
Tayanch doktorant
abdisalomovashahlo@gmail.com
ToshDO‘TAU

Annotatsiya. Tabiiy tilga ishlov berish uchun ishlab chiqilgan tizimlar, birinchi navbatda, ma’lumotlar bazasi bilan ta’minlanishi kerak. Ma’lumotlar bazasidagi ma’lumotlarni tez fursatda qayta ishlash, ulardan qulay tarzda foydalanish maqsadida lingvistik bazalar ma’lumotlar bazasini boshqarish tizimlari (MBBT)da saqlanadi. Ushbu maqolada ma’lumotlar bazasini yaratishda e’tibor qaratish zarur bo‘lgan jihatlar, o‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimining lingvistik ta’minotini tuzishda, ularni annotatsiyalashda tayanilgan tamoyillar, ma’lumotlar bazasini SQL Serverga kiritish jarayoni haqida so‘z yuritiladi. Ma’lumotlarni MBBTda saqlash nafaqat tahlilni soddalashtiradi, balki NLP modellari bilan ishlashda aniq va izchil ma’lumot almashinuvini ta’minlaydi. Tadqiqot o‘zbek tilida koreferensiyani modellashtirishga doir dastlabki amaliy yondashuvlardan biri bo‘lib, koreferensiyani avtomatik aniqlash tizimi, o‘zbek tilidagi matnlarni tahlil qiluvchi algoritmlar va ma’lumotlar bazasi asosida ishlovchi lingvistik vositalarni yaratishda muhim poydevor bo‘lib xizmat qiladi.

Abstract. Systems designed for Natural Language Processing must first be equipped with a database. In order to quickly process and conveniently use the data in the database, linguistic databases are stored in database management systems (DBMS). This article discusses the aspects that need to be paid attention to when creating a database, the principles used in creating the linguistic support of the automatic coreference detection system for Uzbek texts, their annotation, and the process of entering the database into SQL Server. Storing data in DBMS not only simplifies analysis, but also ensures accurate and consistent information exchange when working with NLP models. The study is one of the first practical approaches to modeling coreference in the Uzbek language and serves as an important foundation for creating an automatic coreference detection system, algorithms for analyzing Uzbek texts, and linguistic tools based on the database.

Абстракт. Системы, предназначенные для обработки естественного языка, должны быть в первую очередь оснащены базой данных. Лингвистические базы данных хранятся в системах управления базами данных

(СУБД) для быстрой обработки и удобного использования данных базы данных. В статье рассматриваются аспекты, которые необходимо учитывать при создании базы данных, принципы, используемые при создании лингвистического обеспечения системы автоматического определения корреспонденции узбекских текстов, их аннотирование, а также процесс ввода базы данных в SQL Server. Хранение данных в MBBT не только упрощает анализ, но и обеспечивает точный и последовательный обмен информацией при работе с моделями NLP. Исследование является одним из первых практических подходов к моделированию кореференции в узбекском языке и служит важной основой для создания системы автоматического определения кореференции, алгоритмов анализа узбекских текстов и лингвистических инструментов на основе баз данных.

Kalit so‘zlar: *koreferensiya, NLP, ma’lumotlar bazasi, annotatsiyalash, SQL Server, korpus, tizim, o‘zbek tili*

KIRISH

Hozirgi kunda sun’iy intellekt va tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasidagi rivojlanishlar inson tilini tushunish, uni tahlil qilish va undan samarali foydalanish imkoniyatlarini kengaytirmoqda. Ayniqsa, matnlarda uchraydigan **koreferensiya** hodisasi, ya’ni biror so‘z yoki ifodaning boshqa bir so‘z yoki ifodani nazarda tutishi – tabiiy til semantikasi va kontekstini tushunishda muhim ahamiyatga ega. Masalan, *“Kelajak uni kashf etganlarga nasib qiladi”* jumlasidagi *“u”* olmoshining *“kelajak”* so‘ziga tegishli ekanini aniqlash koreferensiya hodisasiga misoldir. Bu hodisani aniqlash nafaqat tilshunoslikda, balki mashina tarjimasi, savol-javob tizimlari, matnni umumlashtirish kabi ko‘plab NLP vazifalarida ham muhim o‘rin tutadi.

Koreferensiyani avtomatik aniqlashdagi birinchi qadam modelni o‘qitish va baholash uchun lingvistik ta’minotni shakllantirishdan iborat. Jahon tajribasiga nazar tashlasak, ko‘pgina tillarda koreferensiyani avtomatik aniqlash tizimlari ishlab chiqilgan va ular maxsus ma’lumot bazalarida sinovdan o‘tkazilgan [1, 2, 3]. O‘zbek tilida koreferensiyani avtomatik aniqlash tizimini yaratishdagi asosiy muammo o‘zbek tilida annotatsiyalangan katta hajmli ma’lumotlar bazasiga bo‘lgan ehtiyoj bilan bog‘liq. O‘zbek tilida turli maqsadga yo‘naltirilgan korpuslar mavjud bo‘lsa-da [4, 5, 6, 7] ularni koreferensiyani avtomatik aniqlash tizimiga to‘g‘ridan to‘g‘ri tatbiq etish kutilgan natijani bermaydi. Demak, o‘zbek tilida koreferensiya hodisasi kuzatiladigan matnlarni aniqlash va ularni strukturaviy ma’lumot shaklida bazaga joylashtirish bo‘yicha tadqiqotlarga ehtiyoj katta. O‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimi uchun lingvistik baza shakllantirilgan va № BGU 1914 raqami bilan ro‘yxatdan o‘tkazilgan. Maqolada ushbu jarayonda amalga oshirilgan ishlar ketma-ketligi batafsil yoritib beriladi.

O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasini shakllantirish

O‘zbek tilidagi matnlar koreferensiyasini yaratishda quyidagi bosqichlar amalga oshirilishi zarur:

1. Keng ko‘lamda koreferent birliklar ishtirok etgan o‘zbekcha matn fragmentlarini to‘plash.

Ushbu bosqichda **matnni tanlash mezonlari, uning hajmi va tadqiqot manbalarini** aniq belgilash maqsadga muvofiqdir.

a) Ma’lumki, koreferensiya istalgan matnda uchraydigan hodisa emas. Matnni tanlash mezonlari orqali ma’lumotlar bazasi uchun qanday xususiyatga ega bo‘lgan matnlar zarurligi oydinlashadi. Biz lingvistik bazani yaratishda amal qilgan matnni tanlash mezonlari 1-jadvalda havola qilinadi:

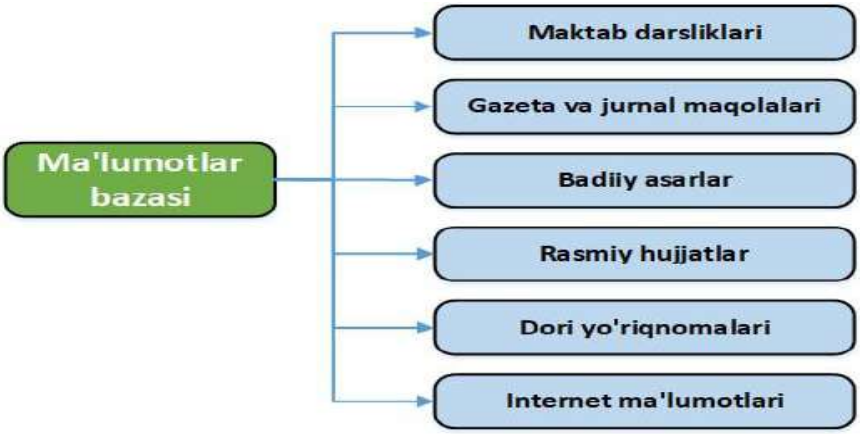
1-jadval. Ma’lumotlar bazasi uchun matnni tanlash mezonlari

Mezon	Tavsif
Turli nutq uslubiga oid matnlar	Badiiy, ilmiy-ommabop, publitsistik, so‘zlashuv va rasmiy uslubdagi matnlar qamrab olinishi maqsadga muvofiq.
Referentlar xilma-xilligi	Matnda koreferentlikni hosil qiluvchi turli leksik birliklar: olmoshlar, sinonimlar, murojaat birliklari, atoqli ot va otli birikmalar ishtiroki
Murakkab sintaktik tuzilmalar	Matnda katafora, anafora hodisalarining mavjudligi
O‘zaro bog‘liqlik	Matn qismlarining mazmunan yaxlit va izchil bo‘lishi muhim.

b) Matnlarni unda ifodalangan axborotning hajm belgisiga ko‘ra, minimal va maksimal matn tiplariga ajratish mumkin. Shuningdek, ma’lum bir matn tarkibidagi murakkab sintaktik butunlikka nisbatan mikromatn, yaxlit matnga nisbatan esa makromatn atamasi qo‘llaniladi [8]. Ayrim adabiyotlarda matn hajm jihatidan uch turga bo‘linadi: kichik, o‘rta va katta hajmli matnlar [9]. Ma’lumotlar bazasining mashinalar uchun xizmat qilishini va matn tasniflarini hisobga olgan holda maxsus koreferensiya korpusini tuzishda minimal, kichik va o‘rta hajmli matnlarga tayanish tavsiya qilinadi.

c) Ma’lumotlar bazasini tuzishda matnlarning qaysi manbaga oid ekanligini ham inobatga olish zarur. Koreferentlikning turli ko‘rinishlarini [10] aks ettirgan matnlarni yig‘ish turli shakldagi manbalarga murojaat qilishni talab qiladi. Mavjud koreferensiya korpuslari tahlili shuni ko‘rsatadiki, koreferensiyani avtomatik aniqlash tizimining ma’lumotlar bazasi uchun manbalardan foydalanish ixtiyoriydir. Ya’ni foydalaniladigan adabiyotlar ro‘yxati, ularning muayyan sohaga oid yoki sohalararo ekanligi, janri bo‘yicha cheklov o‘rnatilmagan. Jumladan, MUC

korpusida [11] “Wall Street” jurnalining 318 ta annotatsiyalangan maqolasi mavjud bo‘lsa, GUM korpusidan [12] muloqot, ta’lim va yangiliklarga oid matnlar o‘rin olgan. WikiCoref korpusi esa asosan 30 ta annotatsiyalangan Vikipediya maqolalaridan iborat [13]. O‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimining lingvistik ta’minoti quyidagi manbalar asosida shakllantirildi (1-rasm):



1-rasm. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasi tarkibi

2. Ma’lumotlar bazasini annotatsiyalash.

Annotatsiyalash deganda korpusga izohlovchi lingvistik ma’lumotlarni qo‘shish [14] nazarda tutiladi. Koreferensiya korpuslarini tuzishda matnlar ichidagi referent birliklar va ularning koreferentlarini qo‘lda yoki yarim avtomatik usulda belgilash, ya’ni annotatsiyalash zarur. Bu jarayonda quyidagi maxsus terminlardan foydalaniladi (2-jadval):

2-jadval. Koreferensiya hodisasining asosiy tushunchalari

Termin	Izoh
Eslatma	Koreferentlikni hosil qiluvchi birlik
Antisedent	Asosiy referent; unga ishora qiluvchi birlikdan oldinda joylashadi.
Anafora	Asosiy referentdan keyin kelib, unga ishora qiluvchi birlik
Postsedent	Asosiy referent; unga ishora qiluvchi birlikdan keyin joylashadi.
Katafora	Asosiy referentdan oldin kelib, unga ishora qiluvchi birlik
Klaster	Bitta shaxs/narsa-hodisani anglatuvchi eslatmalar guruhi
Singleton	Koreferenti mavjud bo‘lmagan referent

Avvalo, koreferensiya korpusini annotatsiyalashda har bir tilning o‘z xususiyatlari asosida qoidalar tizimi ishlab chiqiladi va korpus shu qoidalar yordamida teglanadi. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasini annotatsiyalashda quyidagi qoidalar va ko‘rsatmalarga rioya qilindi:

1. Faqat ot, otli birikma, harakat nomi va olmosh o‘rtasida yuzaga kelgan koreferentlik hodisasi inobatga olinsin. Boshqa holatlar e’tibordan chetda qoldirilsin:

a) Antisedent/postsedent gapga teng kelgan holatlar: *Yaxshi odatingiz* ***shuki,*** *barvaqt turasiz.*

b) To‘pdan ajratilgan shaxs yoki narsani ifodalovchi belgilash olmoshiga ishoralar: ***Har kim o‘zi*** *istagan axborotni izlash, olish va uni tarqatish huquqiga ega;* ***Har kim o‘z*** *og‘zining qorovuli bo‘lsa,* ***uning*** *nafasi hech qachon bo‘g‘ilmaydi.*

d) Noaniqlikni ifodalagan referentlar o‘rtasidagi munosabat: ***Kim*** *izlansa,* ***u****, albatta, maqsadiga yetadi.*

e) Son so‘z turkumiga ishoralar: *Kimyoviy bog‘ –* ***ikki*** *yoki* ***undan*** *ortiq atomlarning o‘zaro ta’sirlashuvi.*

f) Anafora hodisasining pronominal anaforadan boshqa turlari.

2. Koreferentlar bitta shaxs/narsa-hodisani bildirgan taqdirda bitta klasterga joylashtiriladi:

Matn: *Kitob bilim manbayi, shuning uchun biz uni sevamiz.*

Koreferent juftlik: *[Kitob, uni]*

3. Ega va ot-kesim munosabatini koreferentlik sifatida belgilamang. Bu munozarali holat bo‘lib, ma’lumotlar bazasida koreferentlar juftligi safiga kiritilmadi: ***Toshkent – O‘zbekistonning poytaxti.***

4. Izohlovchilarni koreferent juftligiga kiritish-kiritmaslik yuzasidan ikki xil qarash bor. Bu faqat **bitta gap tarkibida ro‘y bergan koreferentlik** hodisasiga taalluqlidir:

Bobokalonimiz Amir Temur *haqida so‘z borganda uning buyuk davlatchilik asoschisi bo‘lganligi ko‘p bor ta’kidlanadi.*

O‘g‘lim, qo‘zichog‘im, *orom olyapti.*

Ayrim tadqiqotlarda izohlovchi-izohlanmish munosabatining anaforik va kataforik holatlari koreferentlik sifatida aniqlangan bo‘lsa, ayrim olimlar bu fikrni rad etadilar. Biz Stenford CoreNLP platformasidagi [15] o‘rganishlarimiz natijasida izohlovchilarni izohlanmish bilan birgalikda **span** sifatida bitta birlik deb qabul qildik. Span – yaxlit ma’nosi bir nechta tokenlar orqali anglashiladigan birlik, otli birikma [16].

5. O‘z-o‘ziga havolalar koreferent guruhiga kiritilsin: *Seton-Tompson* ***tabiatni*** *ardoqladi,* ***tabiat*** *ham uni ardoqladi...*

Yuqori aniqlikka erishish maqsadida o‘zbek tilidagi matnlar koreferensiyasi ma’lumotlar bazasi keltirilgan qoidalar asosida murakkab jarayon hisoblangan usulda – qo‘lda teglab chiqildi. Namuna 3-jadvalda aks ettiriladi:

3-jadval. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasini teglash

Matn	Koreferent juftligi	Klaster	Eslatma tegi	Izoh
Farg‘ona fojiasining sabablari, uni harakatga keltirgan kuchlar kim edi?	[Farg‘ona, uni]	1	[NER, olmosh]	[Antisedent, anafora]
Ona – Quyosh, uning mehri tafti hech qachon sovimaydi.	[Ona, uning]	1	[ot, olmosh]	[Antisedent, anafora]
Bobur she’riyatining o‘ziga xos xususiyati unda shoir shaxsiyatining bo‘rtib turishidir.	[Bobur she’riyatining, unda] [Bobur, shoir]	1 2	[NP, olmosh] [NER, ot]	[Antisedent, anafora] [Antisedent, anafora]
Seton-Tompson tabiatni ardoqladi, tabiat ham uni ardoqladi...	[Seton-Tompson, uni] [tabiatni, tabiat]	1 2	[NER, olmosh] [ot, ot]	[Antisedent, anafora] [Antisedent, anafora]
Ana u! Hamma narsa birdan unutildi: u yerda, oldinda, Issiqko‘lning ko‘m-ko‘k sathida oppoq kema paydo bo‘ldi.	[kema, u] [Issiqko‘lning ko‘m-ko‘k sathida, u yerda]	1 2	[ot, olmosh] [NP, ravish]	[postsedent, katafora] [postsedent, katafora]
“Nima deding, Fitna? Nega dilgir bo‘libdilar?” – so‘radi so‘fi Qurvonbibidan.	[Qurvonbibidan, Fitna]	1	[NER, NER]	[postsedent, katafora]

Ma’lumotlarni SQL Serverda saqlash

SQL Server – ma’lumotlarni saqlaydigan va ularga so‘rov qilish imkonini beruvchi ma’lumotlar bazasi mexanizmi [17]. Ushbu tizimda matnlar va koreferent birliklar alohida-alohida o‘zaro bog‘langan jadvallarda saqlanadi va so‘rov orqali tegishli matndagi koreferent bo‘lgan barcha birliklar taqdim qilinadi.

1. SQL Serverda matnlar uchun jadval yaratish.

```
CREATE TABLE documents (
   id INT PRIMARY KEY IDENTITY (1,1),
   title TEXT,   -- Matn sarlavhasi (ixtiyoriy)
   content TEXT  -- Asl matn
);
```

Yuqoridagi kod asosida to‘plangan matnlar ketma-ket tizimga kiritiladi va matnlardan iborat jadval yaratiladi:

id	title	content
1	1A	Har kim o'zi istagan axborotni izlash, olish va uni tarqatish huquqiga ega.
2	2A	Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishib chiqish darkor.
3	3A	Bobur she'riyatining o'ziga xos xususiyati unda shoir shaxsiyatining bo'rtib turis...
4	4A	Kelajak uni kashf etganlarga nasib qiladi.
5	5A	Tuxumning eng qimmatli qismi uning sarig'idir.
6	6A	Qadimda rossiyaliklar marvaridni tozalash uchun avval uni tovuqqa yutdirishgan.
7	7A	Buyuk shaxslar go'zal nutqlar vositasida emas, o'z mehnatlari va ularning natijal...
8	8A	Biz muammoni yaratgan fikrlash tarzimiz bilan uni hal eta olmaymiz.
9	9A	Ommaga ergashgan odam undan o'zib ketolmaydi.
10	10A	Lava otilishidan keyin uning tarqalish tezligi itning yugurish tezligiga yaqindir.

2-rasm. SQL Serverda matn ma’lumotlarining aks ettirilishi

2. Koreferent guruhlar jadvalini hosil qilish.

Koreferent guruhlar jadvalini tuzishda quyidagi kod qo‘llanildi:

```
CREATE TABLE coreference_groups (
   id SERIAL PRIMARY KEY,
   text_id INTEGER REFERENCES texts(id) ON DELETE CASCADE,
   group_number INTEGER NOT NULL,
   phrase TEXT NOT NULL,
   start_index INTEGER,
   end_index INTEGER
);
```

Koreferent klasterlari jadvalida har bir referent va u bilan bog‘liq koreferent ifodalar jadvalda aniq koordinatalar (boshlanish va tugash indekslari) bilan ko‘rsatiladi. Jadval namunasi bilan quyida tanishish mumkin (3-rasm):

console | coreference_groups

WHERE | ORDER BY

	id	text_id	group_number	phrase	start_index	end_index
1	1	1	1	Asarni	0	6
2	2	1	1	uning	36	41
3	3	2	1	Kelajak	0	7
4	4	2	1	uni	8	11
5	5	3	1	Bobur she'riyatining	0	21
6	6	3	1	o'ziga	22	27
7	7	3	1	unda	47	51
8	8	4	1	pand-nasihatlarini	25	44
9	9	4	1	ular	54	58
10	10	5	1	bazm	75	79
11	11	5	1	o'sha	42	46

3-rasm. SQL Serverda koreferent klasterlari jadvali

3. SQL Serverga so‘rov yuborish va ma’lumot olish.

SQL Serverda so‘rov berish orqali o‘zaro bog‘langan jadvallardagi umumiy ID ga ega bo‘lgan ma’lumotlar birlashtiriladi. SQL Serverda so‘rov yuborish uchun kod:

```
SELECT
   t.id AS text_id,
   t.content AS text_content,
   cg.group_number,
   cg.phrase,
   cg.start_index,
   cg.end_index
FROM texts t
LEFT JOIN coreference_groups cg ON t.id = cg.text_id
WHERE t.id = 1;
```

So‘rov yuborilgach, matn va unga tegishli koreferent birliklar natija sifatida berilishi kerak. 4-rasmda *“Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishib chiqish darkor”* gapi misolida so‘rov natijalari tasvirlangan:

Output | Result 5

	text_id	text_content	group_number	phrase	start_index	end_index
1	1	Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishi…	1	Asarni	0	6
2	1	Asarni tushunish uchun, avvalo, uning mazmuni bilan tanishi…	1	uning	36	41

4-rasm. SQL Serverda so‘rov natijalari

XULOSA

NLP da tabiiy tilga ishlov berishning so‘nggi bosqichi koreferensiyani avtomatik aniqlashdir. O‘zbek tilida bunday tizimni, uning lingvistik ta’minotini shakllantirishdagi kuzatishlar keying ishlarimiz uchun quyidagi xulosalarni berdi.

1. O‘zbek tilidagi matnlar koreferensiyasini aniqlovchi tizimning lingvistik ta’minoti hajmini yanada kengaytirish lozim. Chunki ma’lumot bazasining hajmi modelning aniqlik ko‘rsatkichiga ta’sir etadi.

2. O‘zbek tilidagi matnlar koreferensiyasi lingvistik bazasida gapga ishoralarni ham annotatsiyash foydalanuvchilar uchun yanada qulaylikni ta’minlaydi.

3. Lingvistik ta’minotni ma’lumotlarni boshqarish tizimida saqlash uchun aniq formatni (JSON, CoNLL) tanlash zarur.

Foydalanilgan adabiyotlar:

1. Büyüktekin F., Özge U. A coreference corpus of Turkish situated dialogs / Proceedings of the First Workshop on Natural Language Processing for Turkic Languages (SIGTURK 2024), August 15, 2024. – pages 42–52.

2. Dobrovolskii V., Michurina M., Ivoylova A. RuCoCo: a new Russian corpus with coreference annotation. / https://doi.org/10.48550/arXiv.2206.04925.

3. Poesio M., Camilleri M., Garcia1 P.C., Yul J., Müller M. The ARRAU 3.0. / CorpusProceedings of the 5th Workshop on Computational Approaches to Discourse (CODI 2024), March 21, 2024. – pages 127–138.

4. https://uznatcorpara.uz/

5. http://alishernavoicorpus.uz/uz/about

6. https://uzschoolcorpara.uz/

7. https://uzbekcorpus.uz/newIndex

8. Yo‘ldoshev M. Badiiy matnning lisoniy tahlili. / O‘quv qo‘llanma. – Toshkent, 2007. – 150 b.

9. Qilichev E. Matnning lingvistik tahlili. – Buxoro, 2000. – 36 b.

10. Abdisalomova Sh. NLPda koreferensiyani hal etish vazifasining o‘rni. / Guliston davlat universiteti axborotnomasi, Gumanitar-ijtimoiy fanlar seriyasi, № 4, 2024. – 166-169-betlar.

11. Hirshman, L. and Chinchor, N. MUC-7 coreference task definition. version 3.0. / In Proceedings of the Seventh Message Understanding Conference (MUC-7), 1998. – pages 127-138.

12. https://gucorpling.org//gum/

13. Ghaddar A., Langlais. WikiCoref: An English Coreference-annotated Corpus of Wikipedia Articles. https://aclanthology.org/L16-1021.pdf; https://github.com/victoriasovereigne/WikiCoref-CoNLL

14. Primova M. Til korpuslarida matnlarni annotatsiyalash: afzallik va kamchiliklari. / “O‘zbekiston: Til va madaniyat” jurnali, Vol. 4 (6), 2023. – 6-18-betlar.

15. https://corenlp.run/

16. Elova D. O‘zbek tili matnlarida koreferensiyani hal qilish bosqichlari. / “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Xalqaro ilmiy-amaliy konferensiya materiallari, Vol. 1, № 01 (2023). – 156-161-betlar.

17. https://docs.dot-net.uz/database/ms-sql-server

AVTOMATLASHTIRILGAN TIL TAHLILI: SLENGNI ANIQLASH VA TEGLASH MUAMMOLARI

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Iroda Bekmuradova Zokir qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
irodabekmuradova9@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada zamonaviy matnlarda keng tarqalgan sleng birliklarini avtomatik aniqlash va teglash muammolari yoritilgan. Tilshunoslik va sun’iy intellekt kesishmasida olib borilayotgan tadqiqotlar tahlil qilinib, amaliy misollar keltiriladi. Tadqiqot natijalari til texnologiyalari rivojida muhim ahamiyat kasb etishi mumkinligi ta’kidlanadi. Muammoning dolzarbligi, mavjud yondashuvlar, duch kelinayotgan qiyinchiliklar hamda kelajakdagi imkoniyatlar ko‘rib chiqiladi. Maqola matnlarda slenglarni avtomatik tahlil qilish uchun zamonaviy metodlar va sun’iy intellekt tizimlarining imkoniyatlarini o‘rganishga qaratilgan.

Abstract. This article discusses the problems of automatic detection and labeling of common slang units in modern texts. Research conducted at the intersection of linguistics and artificial intelligence is analyzed and practical examples are given. It is emphasized that the results of the research can be of great importance in the development of language technologies. The relevance of the problem, existing approaches, challenges encountered, and future opportunities are considered. The article is aimed at studying the capabilities of modern methods and artificial intelligence systems for automatic analysis of slang in texts.

Аннотация. В статье рассматриваются проблемы автоматического обнаружения и разметки общеупотребительных сленговых единиц в современных текстах. Будут проанализированы исследования на стыке лингвистики и искусственного интеллекта, а также приведены практические примеры. Подчеркивается, что результаты исследования могут иметь существенное значение в развитии языковых технологий. Обсуждаются актуальность проблемы, существующие подходы, стоящие перед ней проблемы и будущие возможности. Целью статьи является исследование возможностей современных методов и систем искусственного интеллекта для автоматического анализа сленга в текстах.

Kalit so‘zlar: *sleng, avtomatlashtirish, teglash, sun’iy intellekt, til tahlili, NLP.*

Global raqamli makonda foydalanuvchilar sonining ortishi matnli ma’lumotlarning mislsiz darajada ko‘payishiga olib keldi. Xususan, ijtimoiy tarmoqlar, bloglar va messenjerlar orqali har kuni milliardlab matnlar yaratilmoqda. Ushbu matnlarning muhim xususiyatlaridan biri – norasmiylik va slenglarning keng qo‘llanilishi. Sleng tili tez o‘zgaruvchan, kontekstga bog‘liq va ko‘pincha rasmiy lug‘atlardan chetda qoladi. Shu bois, slengni avtomatik tarzda aniqlash va teglash masalasi sun’iy intellekt tizimlari uchun dolzarb muammolardan biri hisoblanadi.

Sleng – bu ma’lum ijtimoiy guruh yoki jamiyat qatlamida qo‘llaniladigan, norasmiy, obrazli va emotsional boyitilgan so‘z va iboralar tizimi. Sleng odatda kundalik muloqotda, yoshlar tilida, professional jargonda yoki internet muhitida keng qo‘llaniladi[1]. Sleng birliklarining asosiy xususiyatlari: normativ bo‘lmaganlik, tez yangilanish, kontekstga bog‘liqlik, emotsional yuklama va hokazo.

Sleng birliklari ko‘pincha qisqartirishlar, metaforalar, leksik yangi hosilalar, fonetik o‘zgartirishlar orqali yaratiladi. Bu esa ularni aniqlashni murakkablashtiradi. Matnlarda slengni aniqlash bir nechta nazariy yondashuvlar orqali amalga oshiriladi. Ulardan eng birinchisi, lug‘atga asoslangan yondashuv bo‘lib, bu usulda oldindan tuzilgan sleng lug‘atlaridan foydalaniladi. Lug‘atdagi birliklar matnda qidirilib, aniqlansa, teg qo‘yiladi. Kamchiligi – lug‘at doimiy yangilanib turishi kerak va yangi paydo bo‘lgan slenglarni qamrab olmaydi. Ikkinchi nazariy yondashuv qoidaga asoslangan yondashuvdir. Bu yondashuv sleng birliklarini fonetik yoki morfologik xususiyatlariga ko‘ra aniqlashga qaratiladi. Misol uchun, qisqartmalar yoki ba’zi o‘zgacha yozuv shakllari orqali. Shu bilan birga, slenglarni aniqlashda statistik modellardan foydalaniladi. Masalan, n-gram modellari yoki TF-IDF usuli orqali matndagi odatiy bo‘lmagan birliklar aniqlanadi [1]. Bundan tashqari, mashinaviy o‘rganish algoritmlari – logistika regressiyasi, Naive Bayes, SVM yoki chuqur o‘rganish (Deep Learning) modellari qo‘llaniladi.

So‘nggi yillarda slengni aniqlashda transformer arxitekturasi asosidagi modellar samaradorlikni oshirdi. Bu modellar kontekstni chuqur tushunishga asoslanganligi sababli, yangi va kutilmagan sleng shakllarini ham muvaffaqiyatli aniqlash imkonini beradi.

Matnda slengni aniqlagach, uni teg (ya’ni belgi) bilan belgilash jarayoni muhim bosqich hisoblanadi. Teglash orqali matndagi sleng birliklari yanada aniqlik bilan tasniflanadi. Bu esa keyingi til tahlilida (masalan, sentiment tahlili yoki matnni normallashtirishda) qo‘l keladi.

Slengni teglashda quyidagi muammolar kuzatiladi:

1. Ko‘pma’nolilik (polisemiya). Slenglarning ba’zilari tildagi boshqa leksik hodisalar: metafora, o‘xshatish, sinekdoxa kabi hodisalarga o‘xshab ketadi. Ammo

shunday bo‘lsa ham mazkur hodisalar bir-biridan farqli holatda qarashni talab etadigan ayrim xususiyatlarga ega. So‘zlarning ma’no ko‘chish hodisalari, umuman ko‘chma ma’noda qo‘llangan birliklar har bir so‘zlovchi tomonidan, nutq vaziyatidan kelib chiqqan holda individual hosil qilinishi mumkin. Ammo slenglarda bunday xususiyat yo‘q, ular an’anaviy foydalanishda bo‘ladi.

2. Kontekstga bog‘liqlik. Kontekstual semantika muhim hisoblanadi. Slengni aniqlash uchun faqat so‘zning o‘zi emas, uning atrofidagi kontekst ham tahlil qilinishi kerak. Har qanday slengning mazmun-mohiyati, u muayyan so‘zlar bilan birgalikda kelganda, kontekstda anglashiladi. Nutqdan ajratilgan holatda ko‘pchilik o‘zbekcha slenglar slenglik ma’nosini yo‘qotadi.

3. Tez yangilanish va eskirish. Tilda mavjud slenglar tez-tez yangilanib turadi. Bunga chet tillardan kirib keladigam slenglarni misol qilib keltirish mumkin. Masalan, “lol” so‘zi kontekstdan kelib chiqib kulgu ifodasi yoki boshqa ma’no berishi mumkin. Teglash modellarining ustunligi shundaki, ular kontekstni inobatga olib, to‘g‘ri teglash imkonini yaratadi.

4. Slengning aniqligi va chegaralari. Aksariyat slenglar rasmiy lug‘atlarga kirmagan bo‘lishi mumkin va u kontekstdan kontekstga o‘zgaradi.

5. Slenglar diaxron (vaqt o‘tishi bilan o‘zgaruvchan) va diatopik (mintaqaga bog‘liq) bo‘ladi. Masalan: “otvet qaytarish” – oddiy javob berish emas, balki ko‘proq “keskin munosabat bildirish” ma’nosini anglatadi.

6. Korpus muammosi (ma’lumot yetishmovchiligi). Ko‘pchilik NLP (Natural Language Processing) modellari rasmiy, toza matnlar (masalan, yangiliklar, ilmiy maqolalar) asosida o‘rgatilgan. Ma’lumotlar yetarli bo‘lmagani uchun “low-resource language” muammosi yuzaga kelishi mumkin. Modellar (masalan, BERT, RoBERTa) kontekstni yaxshi tushunsa ham, slengda bu har doim yetarli bo‘lmaydi.

7. Yangi slenglar va dinamik rivojlanish. Sleng tezda yangilanadi. Yangi so‘zlar ijtimoiy tarmoqlar, trendlar orqali tez tarqaladi.

Slenglarni avtommatik teglash (ijtimoiy tarmoq, chat, forumlar asosida), kontekstga moslashtirilgan modellarini o‘rganish, slenglar bazasini kengaytirish (xalqdan to‘plangan misollar bilan) orqali yanada samarali natijalar berish mumkin. Slenglarni teglash (ya’ni, ularni belgilash, ajratish yoki aniqlash) tilshunoslikda, tarjima, NLP (natural language processing), hamda madaniyatshunoslikda muhim ahamiyatga ega. Slenglar – bu tilning norasmiy, ko‘proq og‘zaki shaklida ishlatiladigan elementlari. Ularni teglash orqal sleng va adabiy til o‘rtasidagi farq aniqlanadi, til o‘zgarishi va rivojlanishini kuzatish osonlashadi, muayyan yosh guruhlari yoki submadaniyatlar tilini o‘rganish imkoniyati yaratiladi. Shuningdek, slenglar ko‘p hollarda kontekstga bog‘liq bo‘ladi va ularni noto‘g‘ri tarjima qilish

mazmunni buzib qo‘yadi. Slengni aniqlash to‘g‘ri ekvivalent topishga yordam beradi. So‘zning ohangini va uslubini saqlab qolishga xizmat qiladi. Masalan, chatbotlar, avtomatik tarjimonlar, yoki ijtimoiy tarmoqdagi kontentni tahlil qiluvchi tizimlar uchun to‘g‘ri ma’nosini tushunish hamda kontekstga mos javob qaytarish orqali slengni aniqlay olish muhim. Qaysi sleng qaysi mintaqada, yosh guruhida yoki submadaniyatda keng ishlatilishini aniqlash mumkin. Qolaversa, ijtimoiy tarmoqlardagi til o‘zgarishlarini tahlil qilish qulaylashadi.

Yuqorida keltirganimiz singari slenglarni aniqlashda statistik modellardan foydalanish keng tarqalgan. Masalan, n-gram modellari yoki TF-IDF (Term Frequency-Inverse Document Frequency) usuli orqali matndagi odatiy bo‘lmagan birliklar ajratib olinadi. n-gram modellar – bu matndagi so‘zlar ketma-ketligini o‘rganishga asoslangan statistik model. Ular matndagi so‘zlarni qismlarga ajratib, o‘xshashliklarni aniqlashda yordam beradi. n – bu ketma-ketlikdagi so‘zlar sonini bildiradi [1], masalan:

1) 1-gram (unigram): har bir alohida so‘z.
2) 2-gram (bigram): ikki so‘zli ketma-ketlik (masalan, “yaxshi kitob”).
3) 3-gram (trigram): uch so‘zli ketma-ketlik (masalan, “men kitob o‘qiyman”).

n-gram modellarining afzalligi shundaki, ular kontekstga asoslanib, o‘xshash so‘z birikmalarini aniqlashda yaxshi natijalar beradi. n-gram modelida slengni aniqlashda, odatiy bo‘lmagan so‘zlar yoki qisqartmalarning yuqori chastotaga ega bo‘lgan kombinatsiyalarini topish mumkin. Shuningdek, slenglarning tez-tez ishlatiladigan shakllari yoki yangi paydo bo‘lgan shakllari n-gram modellarini qo‘llash orqali aniqlanadi.

Kelgusida slenglarni aniqlash va teglash sohasida quyidagi yo‘nalishlarda ilmiy izlanishlar olib borish bilan bog‘liq vazifalar mavjud:

T/R	Yo‘nalish	Mexanizm
1	Til korpusini kengaytirish	Sleng birliklarini doimiy ravishda yangilab borish orqali
2	Ko‘p tillilikni qo‘llab-quvvatlash hamda tildagi rang-baranglikni saqlash	Sleng tilni jonlantiradi. Teglash orqali bu boylikni yo‘qotmasdan ko‘rsatish mumkin. Ko‘p tillilikda bu juda ham muhim.
3	Kontekstga asoslangan dinamik teglash tizimlarini yaratish	Tarjima yoki tilni avtomatlashtirish tizimi kontekstni to‘g‘ri tushunadi
4	Real vaqt rejimida ishlaydigan slengni aniqlash botlarini ishlab chiqish	Til o‘rganish jarayonida yoki avtomatik tarjima qilinayotgan tizimlar slenglarni ajrata oladigan bo‘lsa, ular noto‘g‘ri tushunish xavfidan saqlanadi.
5	Ko‘p modalli yondashuv	Sleng ko‘pincha matn bilan cheklanmaydi – videolar, audio formatlarda ham keng qo‘llaniladi.

		Slengni aniqlashni faqat matnga emas, balki boshqa formatlarga ham kengaytirish.
6	Slengni aniqlash jarayonini tezlashtirish va zamonaviy o‘zgarishlarga moslashuvchan qilish	Slenglarni tezkor aniqlash uchun umumiy kontekstni tushunish modellarini qo‘llash orqali.

Xulosa qilib aytganda, matnlarda slenglarni avtomatik aniqlash va teglash zamonaviy tilshunoslik va sun’iy intellekt kesishmasida dolzarb tadqiqot yo‘nalishidir. Bundan tashqari, ijtimoiy tarmoqlar monitoringi, brend reputatsiyasini boshqarish, ijtimoiy tahlil va psixolingvistik tadqiqotlar uchun ham slengni avtomatik aniqlash muhim ahamiyat kasb etadi. Amaliy natijalar shuni ko‘rsatadiki, chuqur o‘rganishga asoslangan modellar yuqori aniqlikni ta’minlaydi. Kelgusida algoritmlarni takomillashtirish va ma’lumotlar bazasini kengaytirish orqali ushbu tizimlarni yanada samarali qilish mumkin.

Foydalanilgan adabiyotlar:

1. Devlin J., Chang M. W., Lee K., & Toutanova K. (2019). “BERT: Pre-training of deep bidirectional transformers for language understanding.” NAACL.
2. Eisenstein, J. (2013). “What to do about bad language on the internet.” Proceedings of NAACL.
3. Zhang Y., and Yang Q. (2018). “An overview of multi-task learning in deep neural networks.” IEEE Transactions.
4. Joachims T. – Text categorization with support vector machines: Learning with many relevant features. 1998, European Conference on Machine Learning, 137-142 p.
5. Матюшенко Е. Современный молодежный сленг: формирование и функционирование. автореф. дисс. кандидата филологических наук – Волгоград, 2007. – 4 б.
6. Muhamedova S. Kompyuter lingvistikasi(metodik qo‘llanma). – Toshkent, 2007.; Po‘latov A., Muhamedova S. Kompyuter lingvistikasi (o‘quv qo‘llanma). – Toshkent, 2008 – 98-b.

UDK: 811.512.133`42

O‘ZBEK TILIDAGI SODDA GAPLARNING IFODA MAQSADIGA KO‘RA TURLARINI MODELLASHTIRISH MASALASI

Abdullayeva Oqila Xolmo‘minovna,
doktoranti DSc
abdullayevaoqila@gmail.com
ToshDO‘TAU

Bozorqulova O‘g‘iloy Erkin qizi
Kompyuter lingvistikasi mutaxassisligi magistranti
ogiloybozorqulova62@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur tadqiqotda o‘zbek tilidagi sodda gaplarning ifoda maqsadiga ko‘ra turlari (darak, so‘roq, buyruq, undov gaplar) bo‘yicha mashinali o‘qitishga mo‘ljallangan modellar tuzilgan va har bir tur uchun namunalar keltirilgan. Ushbu tadqiqot uchun besh yuzta gap saralab olingan hamda ushbu gaplar namuna sifatida berilgan. Maqolada tadqiqot ishining tilshunoslik va kompyuter lingvistikasidagi yangi natijalar berishi ilmiy asoslangan.

Abstract. In this study, machine learning models were developed for the types of simple sentences in the Uzbek language according to their expressive purpose (declarative, interrogative, imperative, exclamatory) and examples were provided for each type. A total of five hundred sentences were selected for the study and are presented as examples. The study provides a scientific foundation for future research in the fields of linguistics and computational linguistics.

Аннотация. В данном исследовании были разработаны модели машинного обучения для типов простых предложений в узбекском языке в соответствии с их экспрессивной целью (декларативное, вопросительное, повелительное, восклицательное) и для каждого типа были приведены примеры. Всего для исследования было отобрано пятьсот предложений, которые представлены в качестве примеров. Исследование дает научную основу для будущих исследований в области лингвистики и компьютерной лингвистики.

Kalit so‘zlar: *sodda gaplar, darak gap, so‘roq gap, buyruq gap, undov gap, model, modellashtirish, sintaktik tahlil.*

Zamonaviy axborot - texnologiyalari rivoji boshqa sohalar singari tilshunoslikka ham o‘z ta’sirini o‘tkazmasdan qolmadi. Tabiiy tillarni mashina uchun tushunarli bo‘lgan sun’iy tilga o‘tkazish masalasi tilshunoslik oldidagi dolzarb masalaga aylandi, shu boisdan, XX asrning 40-yillarining boshidan rivojlana

boshlagan ushbu ta’limotlar natijasida kompyuter lingvistikasi, korpus lingvistikasi, mashina tarjimasi kabi bir qancha yo‘nalishlar vujudga keldi [1]. Jahon tilshunosligida ushbu sohada diqqatga sazovor amaliy ishlar bajarilgan bo‘lib, Chomskiyning Transformational grammar (Transformatsion grammatika), Dependency grammar (bog‘lanishli grammatika), Lexical functional grammar (leksik-funksional grammatika), Cognitive grammar (Kognitiv grammatika), Probabilistic parsing (Ehtimolli bog‘lanish), Combinatory Categorial grammar (Kombinatorlik kategorial grammatika) kabi nazariyalari amaliy tilshunoslik uchun xizmat qilib kelmoqda. Bundan tashqari, gap bo‘laklarini modellashtirish va analiz qilish uchun kuchli NLP algoritmlari mashina xotirasi bilan ishlaydigan DeepL va Google Translate kabi tarjima dasturlarida ishlab chiqilgan. Tilni qayta ishlash jarayonida gap bo‘laklarini modellashtirish muhim ahamiyat kasb etuvchi omil hisoblanadi. Modellashtirish jarayonida tadqiqotchi obyektni emas, uning modelini o‘rganadi [1]. Til modellarini ishlab chiqish tabiiy tilni qayta ishlashda (NLP – natural language processing) asosiy masalalardan biridir. Shuningdek, til modellarini ishlab chiqish tilshunoslik, kompyuter lingvistikasi va boshqa sohalarda amalga oshadigan avtomatlashtirish jarayoniga ijobiy ta’sir ko‘rsatadi. Ushbu tadqiqotda o‘zbek tilidagi sodda gaplarning ifoda maqsadiga ko‘ra turlari bo‘yicha modellar tuziladi va namunalar bo‘yicha tahlil qilinadi.

Sodda gap – bir hodisa yoki fikrni ifodalovchi, odatda bir predikativ markazga ega bo‘lgan gapdir [2]. Sodda gaplar o‘z ichida turli kategoriyalar bo‘yicha guruhlarga bo‘linadi. Ular orasida sodda gaplarning ifoda maqsadiga ko‘ra turlari: darak, so‘roq, buyruq va undov kabilarga bo‘linadi. So‘zlovchi yoki yozuvchi obektiv borliqdagi narsa, hodisalarni gap orqali aks ettiradi. Bunda o‘z oldiga har xil maqsad qo‘yishi mumkin. Masalan u gap orqali biror voqea yoki hodisa haqida xabar beradi yoki o‘ziga aniq bo‘lmagan narsa haqida savol beradi va savolga javob olishni istaydi yoki o‘z xohishini (buyruq, iltimos kabi) ifodalashni ko‘zda tutadi. Ana shu ifoda maqsadiga nuqtayi nazaridan gaplar darak, so‘roq, buyruq kabi turlarga bo‘linadi [3]. Gaplarni modellashtirish masalasida bajariladigan asosiy vazifalardan biri bu – gapni sintaktik bo‘laklarga ajratish hisoblanadi. Sodda gaplar bir necha ega va ikkinchi darajali bo‘laklardan va bitta kesimlik belgisidan iborat bo‘ladi. (Shu sababli keyingi o‘rinlarda ega – E, to‘ldiruvchi – T, hol – H, aniqlovchi – A, undov gapni hosil qiluvchi birlik – U.b, so‘roq gapni hosil qiluvchi birlik – S.b belgilari bilan belgilanadi). Shu jihatdan sodda gaplardan darak gapning gap bo‘laklari bilan kelishiga ko‘ra quyidagicha modellashtirish mumkin.

So‘zlovchining biror voqea-hodisa haqida xabar berishi yoki ma’lum predmetda muayyan belgining mavjudligini tasdiq yoki inkor qilishini ifodalaydigan gap darak gap deyiladi [4]. O‘zbek tili lug‘at tarkibi boy va grammatik xilma-xillikka ega til bo‘lgani uchun ham til modellarini tuzish nisbatan murakkablik tug‘dirishi aniq. Shunday bo‘lsada, sodda gaplarning turlari bo‘yicha modellar

yaratish sintaktik analizatorning rivojiga ham hissa qo‘shmay qolmaydi. O‘zbek tilida darak shakldagi sodda gaplarning modellari quyidagi jadvalda namunalar orqali keltirilgan:

1-jadval. Darak shakldagi sodda gaplarning modellari

E + A + T + H + K	U qadrli mehmonlarni ochiq chehra bilan qarshiladi.
A + E + T + H + K	Butun qishloq eshonga yetti bukilib ta’zim qilardi.
A1 + A2 + E + T + H + K	Bizning butun qadr-u qimmatimiz tafakkurda birday jamlangan.
E + T + H + K	Nasimiy ham unga ipsiz bog‘langan.
E + K	Ukam keldi.
E + H + T + K	Ochil buva bugun chinor bilan xayrlashmoqda.
E + T + K	Taq-tuq hammaga eshitildi.

Ushbu modellardan ko‘rinib turibdiki, o‘zbek tilida eganing o‘rni har doim ham gapning boshida kelavermaydi. Ammo aksariyat hollarda ushbu qoidani istisno qiluvchi holatlar ham mavjud. Ega, asosan, aniqlovchiga bog‘langanligi uchun ham yo aniqlovchidan oldin, yo aniqlovchidan keyin ishlatiladi. Yuqoridagi misollardan ham buni ko‘rish mumkin. Shuningdek ushbu modellar tabiiy tilni qayta ishlash jarayonini nisbatan osonlashtiradi. Ammo bu modellar orasida **inversiya hodisasiga uchragan sodda darak gaplarning modellari** keltirilmagan.

So‘zlovchining o‘ziga noma’lum bo‘lgan biror voqeani suhbatdoshdan bilib olish istagini, zaruratini ifodalovchi gap so‘roq gap hisoblanadi [4]. Bilamizki so‘roq gaplar uch xil vosita yordamida hosil qilinadi. Bular: so‘roq olmoshlari, so‘roq yuklamalari, so‘roq ohangi kabilar. Quyida so‘roq shaklidagi sodda gaplarning modellarini berib o‘tamiz (2-jadval).

2-jadval. So‘roq shaklidagi sodda gaplarning modellari

T + T + K (S.b)	Sizni otga ham ishingiz tushib turadimi?
H (S.b) + K	Qachon ketgan?
A + E + K (S.b)	Sizning safaringiz uzoqmi?
E + H (S.b) + T + K	Sen nega ishingni bajarmading?
E + K (S.b)	Men qarab turaymi?
E + H + H (S.b) +K	Biz bugun qayerga boramiz?

So‘roq shakldagi sodda gaplarning modellari orasida so‘roq olmoshi bilan kelgan sodda gaplar ham mavjud bo‘lib, olmosh mustaqil so‘roqqa javob bo‘lib kela olishi va gapda sintaktik bo‘lak vazifasini bajara olishiga ko‘ra hol yoki to‘ldiruvchi vazifasida kela oladi. Shu sababli ham so‘roq gapni hosil qiluvchi birligimizni (S.b) yuqoridagi shaklda qavs ichida berib o‘tdik. Ammo so‘roq shakldagi sodda gaplarning modellari orasida ritorik so‘roq gaplarning modellari mavjud emas.

Buyurish, iltimos, taklif, xohish, o‘git-istak, xitob, ogohlantirish kabi har xil ma’nolarni ifodalovchi gaplar buyruq gaplar hisoblanadi [4]. Buyruq gaplar odatda ikkinchi shaxsga qaratilgan bo‘lib, shu sababli ularning kesimlari ikkinchi shaxs

buyruq maylida ifodalanadi. Shuningdek birinchi va uchinchi shaxs buyruq formalari ham buyruq gaplarning kesimi bo‘lib kela oladi. Quyida buyruq shaklidagi sodda gaplarning modellari berilgan bo‘lib, modellar namunadagi misollar bilan keltirilgan (3-jadval).

3-jadval. Buyruq shakldagi sodda gaplarning modellari

H + K (U.b)	Shoshilmasdan gapir!
K + T (U.b)	Torting, qo‘lingizni!
A + T + H + K	Bugungi ishni ertaga qo‘yma.
T + K	Kitobni o‘qiy qoling
K (U.b)	Chekilmasin!
E + T + H + K	Sizlar darslarni qunt bilan o‘rganinglar.
E + A + T + K	Davronov g‘alaba pasportiga imzo cheksin.

Tilimizda buyruq gapni faqat buyruq shaklidagi sodda gaplarni buyruq fe’li bilan hosil mumkin degan qat’iy qoida mavjud emas. Shu sababli ushbu qoidani istisno qiluvchi buyruq gaplar ham mavjud: *Shu ondayoq minglab ayollarning chinqirishi eshitiladi: - Non, Non, Non.* Tildagi grammatik turfa xillik til modellarini yaratishga ta’sir ko‘rsatadi. Bu yuqoridagi gapdan sezish mumkin.

Darak, so‘roq, buyruq gaplar kuchli his-hayajon bilan talaffuz qilinishi va bu so‘zlovchining ichki his-hayajoni va ruhiy kechinmalarida o‘z ifodasini topishi mumkin. Emotsional qimmat kasb etuvchi bunday gaplarga undov gaplar deyiladi. Mazmuniga ko‘ra undov gaplar darak-undov gap, so‘roq-undov gap, buyruq-undov gaplarga bo‘linadi. Undov gaplar asosiy gap turlarining maxsus undov intonatsiyasi bilan aytilishi asosida vujudga keladi [4].

Undov gaplarning quyidagicha modellarini tuzib chiqdik va ularni misollar yordamida ifodaladik. Quyidan o‘rin olgan modellar o‘zbek tilida eng ko‘p uchraydigan undov gaplarning namunalaridan tuzilgan. Maxsus saralab olingan 500 ta gaplar orasidan undov gaplarning quyidagi modellarini tuzishga muvaffaq bo‘ldik:

4-jadval. Undov gaplarning modellari

U.b + T + K + E	Eh, bog‘larga burkansin vatanim!
T + K + H	Nurga to‘lsin uyingiz!
K + H	Chiqmaysizmi dalaga?!
U.b + A + T + S.b + K	Eh, bahor sayliga nima yetsin!
K	Yurasizmi?!

Undov gaplar yuqorida ta’kidlanganidek, asosiy gaplarga intonatsion ma’no qo‘shilishi natijasida hosil bo‘ladi va bu namunalarda keltirilgan gaplarda ham aks etgan.

Zamonaviy lingvistika sohasida tabiiy tillarni sun’iy intellekt tizimlariga integratsiyalash masalasi kun tartibida muhim mavzulardan biriga aylanmoqda. Shu nuqtayi nazardan, ushbu tadqiqot doirasida o‘zbek tilidagi sodda gaplarning ifoda maqsadiga ko‘ra modellashtirilishi masalasi dolzarb ilmiy-amaliy ahamiyatga ega bo‘lgan muhim yondashuv hisoblanadi. Tadqiqot davomida sodda gaplarning to‘rtta asosiy turiga – darak, so‘roq, buyruq va undov gaplarga oid sintaktik modellar ishlab chiqilib, tahlil qilindi. Har bir tur uchun amaliy qo‘llanmalar va misollar asosida aniq tavsiyalar taqdim etildi.

Avvalo, darak gaplar modellarining tahlilidan ko‘rinib turibdiki, o‘zbek tilidagi sodda gaplarda ega odatda gap boshida joylashadi. Biroq istisno holatlarda, aniqlovchilar bilan bog‘langan holda, boshqa pozitsiyalarda ham uchrashi mumkin. Ushbu xususiyat tilimizdagi sintaktik erkinlikni va eganing pozitsiyasini aniqlashdagi murakkablikni hisobga olgan holda, tabiiy tilni qayta ishlash (NLP) tizimlarida zarur bo‘lgan aniqlik va fleksibillikni ta’minlaydi.

So‘roq gaplarning modellarida esa, so‘roq birligi (so‘roq olmoshi yoki yuklamasi) asosiy identifikator sifatida ajralib turadi. Ayniqsa, so‘roq gaplardagi ritorik shakllarning mavjud modellashtirilmagan holati kelgusidagi tadqiqotlar uchun yangi imkoniyatlar ochadi. Tadqiqotning muhim jihatlaridan biri – 500 ta gap misolida statistik tahlil o‘tkazilganidir. Ushbu amaliy yondashuv modellarni mashinali o‘rganish (machine learning) va tabiiy tilni qayta ishlash (NLP) tizimlarida qo‘llash uchun mustahkam asos yaratadi. Shunday qilib, mazkur modellar kelajakdagi kompyuter lingvistikasi loyihalarida, jumladan avtomatik tarjima, matnning sintaktik va semantik tahlili hamda gap tuzilishini aniqlash kabi amaliy masalalarda asosiy ko‘rsatkich sifatida xizmat qilishi mumkin. Shu boisdan, ushbu tadqiqot o‘zbek tili sintaksisini zamonaviy texnologik yechimlar bilan integratsiya qilish uchun fundamental poydevor sifatida xizmat qilishi mumkin. Kelgusida mazkur modellar asosida murakkab gap turlarini, ritorik konstruksiyalarni va kontekstual pragmatika masalalarini tahlil qilish imkoniyati yaratiladi.

Foydalanilgan adabiyotlar:

1. Po‘latov A. Kompyuter lingvistikasi. - Toshkent, 2009.
2. G‘ulomov A. Sodda gaplar. - Toshkent, 1955.
3. G‘ulomov A., Asqarova M. Hozirgi o‘zbek adabiy tili. Sintaksis. - Toshkent, 1965. -B.52-53.
4. Abdurahmonov G‘., Sulaymonov A., Xoliyorov X., Omonturdiyev J. Hozirgi o‘zbek adabiy tili. Sintaksis. - Toshkent, 1979.

RAVISH SO‘Z TURKUMINI GRAMMATIK POS TEGLASH ALGORITMLARI: NAZARIY YONDASHUV, AMALIY TATBIQ VA KELGUSIDAGI YO‘NALISHLAR

Xudayberganov Nizomaddin Uktambay o‘g‘li,
o‘qituvchi
nizomaddin@navoiy-uni.uz
ToshDO‘TAU

Karimova Zilola Dilmurod qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
karimovazilola791@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqola o‘zbek tilidagi ravish so‘z turkumini aniqlash va grammatik POS(Part-of-Speech) teglash algoritmlarining nazariy asoslari, amaliy tatbiq usullari va eksperimental natijalarini chuqur tahlil qiladi. Tadqiqotda statistik modellar (HMM,CRF), qoidaviy yondashuv va morfologik tahlil usullari birgalikda o‘rganiladi. Jadval va blok-sxemalar yordamida algoritm strukturalari va samaradorlik ko‘rsatkichlari vizual tarzda ifodalanadi. Asosiy maqsad avtomatlashtirilgan tilni qayta ishlash tizimlarida ravish so‘zlarini, ularning ma’noviy va tuzilishiga ko‘ra turlarini aniqlashning aniqligi va samaradorligini oshirishga qaratilgan yangi metodologiyalarni ishlab chiqishdir.

Abstract. This article provides an in-depth analysis of the theoretical foundations, practical implementations, and algorithms for identifying adverbial word classes models(HMM, CRF), rule-based approaches, and morphological analysis techniques. Through diagrams, tables, and performance metrics are visually presented. The objective is to develop novel methodologies to enhance the accuracy and efficiency of automated natural language processing systems in adverb identification.

Аннотация. В данной статье представлен детальный анализ теоретических основ, практической реализации и алгоритмов для идентификации наречных классов слов с использованием моделей (HMM, CRF), подходов на основе правил и методов морфологического анализа. Диаграммы, таблицы и показатели производительности визуально представлены. Цель исследования разработка новых методологий для повышения точности и эффективности автоматизированных систем обработки естественного языка при идентификации наречий.

Kalit so‘zlar: *ravish so‘z turkumi, POS teglash, POS algoritmlari, morfologik tahlil, statistik modellar, qoidaviy yondashuv*

Kirish. So‘z turkumlarini (POS) belgilash – tabiiy tilni qayta ishlash (NLP) sohasida eng muhim mavzulardan biri bo‘lib, asosiy qurilish bloklaridan va amaliyotlaridan biridir. Shu sababli, so‘z turkumlarini belgilash NLPda keng qo‘llaniladigan mavzu bo‘lib, matndagi har bir so‘zga kontekstga muvofiq tegishli sintaktik yorliqni berishni maqsad qiladi. So‘z turkumlarini avtomatik ravishda aniqlash POS tagging yoki grammatik belgilash deb ataladi. Bu jarayon fe’l, sifat, ravish, ot va boshqa so‘z turkumlariga tegishli bo‘lgan so‘zlarni tasniflashni o‘z ichiga oladi. Ushbu texnologiya mashinada tarjima qilish, so‘z ma’nolarini ajratish, savollarga javob berish, sintaktik tahlil qilish va boshqa ko‘plab NLP ilovalarida qo‘llaniladi. So‘z turkumlarini aniqlash oddiy so‘zlarni ma’lum bir turkumga moslashtirishdan ancha murakkabroq. Chunki bir so‘z turli kontekstlarda har xil so‘z turkumiga ega bo‘lishi mumkin. Shu sababli, POS belgilash uchun yagona universal qoidalar mavjud emas. Matn korpusidagi so‘z turkumlarini qo‘lda aniqlash juda qiyin. Doimiy ravishda yangi so‘zlar va kontekstlar paydo bo‘lib turadi, shuning uchun POS belgilashni qo‘lda bajarish samarali emas. Shu sababli, biz POS belgilash jarayonida mashinani o‘rganishga asoslangan usullarga tayanamiz [1].

Jahon tilshunosligida ham bir qancha olimlar POS teglash bo‘yicha samarali izlanishlar va tadqiqotlar olib borishgan. Ulardan S.K. Cameron POSning asosiy yetakchilaridan biridir va u ijobiy tashkilotlarni yaratish va rivojlantirish haqida ko‘plab ilmiy maqolalar yozgan. Uning asarlari tashkilotlarda yuksak samaradorlik, jamoaviy ishlash va ijobiy psixologiyaning o‘rni haqida keng tahlillarni o‘z ichiga oladi. R.E. Quinn POSni rivojlantirishda muhim rol o‘ynagan olimlardandir. U o‘z tadqiqotlarida organizatsion o‘zgarishlar va ijobiy liderlikni qanday amalga oshirish kerakligi haqida asosiy nazariyalarni ishlab chiqqan. Quinn o‘z tadqiqotlarida tashkilotlardagi ijobiy o‘zgarishlar va innovatsiyalarni qanday boshqarish mumkinligi haqida chuqur tahlil qilgan. U liderlik, ijobiy tuzilmalar va o‘zgarishlarni amalga oshirishga oid ko‘plab ishlanmalarga ega. Jane Dutton POSning tadqiqotida katta hissa qo‘shgan olimlardan biridir. Uning asarlari, ayniqsa, tashkilotlarda rahbarlik va jamoaviy munosabatlar asosida o‘rganilgan. Dutton ijobiy psixologiyaning tashkilotlarda qanday ishlashini va rahbarlarning ishlash samaradorligini qanday oshirishi mumkinligini o‘rganadi.

O‘zbek tilshunosligida kompyuter lingvistika sohasining o‘rganilishi XXI asrdan boshlandi. Bugunga qadar sohaga oid ko‘plab tadqiqotlar olib borilmoqda. O‘zbek tilida POS teglash o‘ziga xos xususiyat bilan o‘zbek tilining boyligini chuqur va keng ekanini yaqqol ifoda eta oladi. Shuning uchun ham POS teglash masalasini maxsus o‘rganish va dasturiy ta’minotini ishlab chiqish katta ahamiyat kasb etadi. O‘zbek tilining milliy korpusi ishlab chiqildi. Shunga qaramay, teglash jarajonida muammolar vujudga kelayotgani, bu jarayondagi kamchiliklar mavjudligi bu sohaning dolzarb ekanligini belgilaydi.

Leksik birliklarni grammatik POS teglash (Part-of-Speech tagging) – bu tilshunoslikda so‘zlarning grammatika jihatdan qanday rol o‘ynashini aniqlash, ya’ni so‘z turkumini belgilash bilan bog‘liq muhim tadqiqot sohasidir[2]. Bu sohada O‘zbekistonlik olimlar ham tadqiqotlar olib borishgan, lekin bu yo‘nalishning o‘zbek tilshunosligida nisbatan yangi va rivojlanayotgan sohaga kirishi mumkin. Leksik birliklar va grammatik POS teglashni o‘rganish asosan kompyuter lingvistikasi, sun’iy intellekt va tabiiy tilga ishlov berish (NLP) texnologiyalariga asoslanadi.

Adabiyotlar tahlili va metodologiya

O‘zbek tilshunosligida teglash masalasi bo‘yicha amaliy va nazariy ishlar olib borilmoqda. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi kafedrasi o‘qituvchilari tomonidan POS teglash boyicha bir qancha ishlar olib borilmoqda. Xususan, B.Elov, Sh.Xamroyeva, N.Xudoyberganov, O.Abdullayeva va boshqalarning ishlari diqqatga sazovordir. POS teglash – bu berilgan gapdagi har bir so‘z shaklga uning turkum (ot, fe’l, sifat, son, ravish yoki olmosh)ga mansubligini belgilash (teglash) vazifasidir. POS teglash tabiiy tilga ishlov berish (Natural Language Processing, NLP) sohasining asosiyvazifalaridan biri bo‘lib, pipeline konveyerining muhimbosqichi hisoblanadi. POS teglash korpusdagi har bir so‘zga grammatik kategoriya belgilashni o‘z ichiga oladi. Ya’ni so‘zning morfologik xuxusiyatlarini aniqlashda ushbu teglashdan foydalanish samarali hisoblanadi. Ushbu teglashda otlar, fe’llar, sifatlar, ergash gaplar, olmoshlar, bog‘lovchilar, yuklamalar va boshqalarni teglashni kiritish mumkin. POS belgilari jumla tuzilmalarini tahlil qilish va lingvistik tadqiqotlar o‘tkazishda yordam beradi.O‘zbek maxsus lingvistik korpuslarini tuzish va ularni teglash masalasi bo‘yicha Sh.Hamroyeva, N.Abdurahmonova M.Abjalova A.Eshmo‘minov Sh.Gulyamova, O.Abdullayeva, D.Elova kabilarning tadqiqotlari alohida ahamiyatga molik.

Zamonaviy texnologiyalar, jumladan, kompyuter lingvistikasi sohasida ravish so‘z turkumini avtomatik aniqlash dolzarb masalalardan biri hisoblanadi. Bu jarayonda Part-of-Speech (POS) teglash algoritmlaridan foydalaniladi. POS teglash jarayonida tilning o‘ziga xos morfologik va sintaktik xususiyatlari inobatga olinadi, bu esa lingvistik modellarni samarali qo‘llashni talab qiladi[3].

So‘z turkumlari tilshunoslikda tilning grammatik tizimini tushunishga yordam beradigan asosiy konseptlardan biridir. Ravish (pronominal) so‘z turkumi, o‘zining grammatik funksiyasi va strukturasi jihatidan boshqa so‘z turkumlaridan farq qiladi[4]. Ravish so‘zlarning grammatik teglashini avtomatik ravishda amalga oshirish esa tilni tahlil qilishda muhim ahamiyatga ega, chunki bu jarayon tilning sintaktikva semantik tuzilishini anglash uchun zarur bo‘lgan asosiy bosqichlardan biridir[5]. Maqolada ravish so‘z turkumini grammatik POS teglash algoritmlari, lingvistik modellari, ularning turlari va metodologiyalari haqida so‘z boradi.

Asosiy qism. O‘zbek tilining murakkab morfologik tuzilishini hisobga olish muhimdir. Ushbu jarayonda quyidagi bosqichlar amalga oshiriladi:

• Asos va qo‘shimchalarni ajratish: So‘zning asosiy qismlari tahlil qilinadi.

• Blok-sxema: Quyidagi diagramma algoritmning asosiy bosqichlarini ifodalaydi:

Quyidagi jadvalda amaliy misollar va ularning teglash natijalari keltirilgan:

1-jadval. Ravishlarni aniqlash va POS teglash bo‘yicha amaliy natijalar

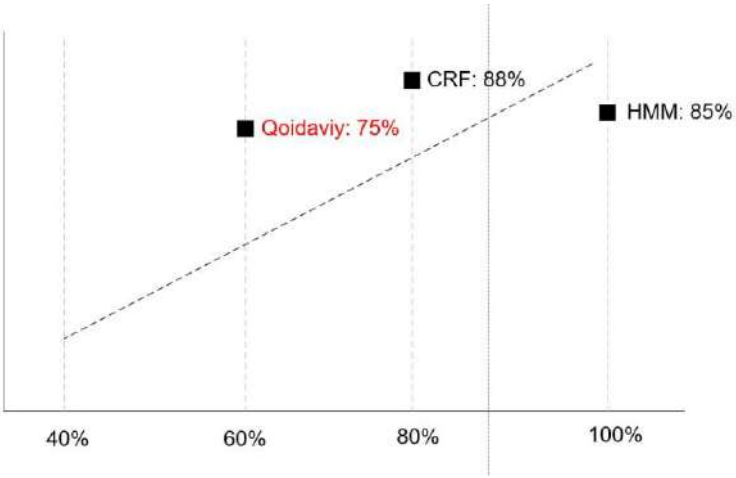
Matn namunasi	Aniqlangan ravish so‘zlari	Algoritm natijasi (POS teglash)
Sardor juda sekin yugurdi	juda, sekin	Ravish: (juda, sekin) – matn kontekstiga mos ravishda
Ra’no tomonidan o‘qilgan she’r o‘zgacha tarzda eshitildi.	o‘zgacha tarzda	Ravish: [o‘zgacha, tarzda] – semantik izoh bilan
U o‘g‘rilikni juda tez payqagan edi.	Juda tez	Ravish: [juda] – qo‘shimcha ibora sifatida
U nihoyatda puxta va tezjavob berdi	nihoyatda, puxta, tez	Ravish: [nihoyatda, puxta, tez] – umumiy tahlil

Bu misollar yordamida statistik va qoidaviyyondashuvlarning samaradorligi, aniqligi va kamchiliklari tajriba asosida tahlil qilinadi.

Tadqiqot jarayonida quyidagi bosqichlar amalga oshirildi:

1. Matn yig‘ish: O‘zbek tilidagi akademik maqolalar, yangiliklar va adabiyotlardan iborat korpus shakllantirildi.
2. Avtomatik teglash: Dastlabki avtomatlashtirilgan teglash algoritmlari yordamida natijalar olinib, keyin qo‘lda tahrir qilindi.
3. Model sinovlari: HMM, CRF va qoidaviy yondashuvlar yordamida modellar o‘rgatilib, ularning aniqlikko‘rsatkichlari taqqoslandi.

Quyidagi diagramma eksperimental natijalarning aniqlik ko‘rsatkichlarini aks ettiradi:



2-rasm. Algoritm aniqlik ko‘rsatkichlari diagrammasi

Natijalar shuni ko‘rsatadiki, CRF modeli kontekstual bog‘lanishlarni hisobga olib eng yuqori aniqlikka erishgan. Shuningdek, qoidaviy yondashuv barqaror natija bergan, ammo moslashuvchanlik jihatidan statistic modellarga nisbatan pastroq ekanligi aniqlangan.

Eksperimental jarayonda quyidagi xatoliklar va cheklovlar aniqlangan:

• *Morfologik noaniqliklar:* Ba’zi hollarda o‘zak va qo‘shimchalarni to‘g‘ri ajratib bo‘lmaslik.

• *Kontekstual cheklovlar:* Ravish so‘zlarining kontekstini to‘liq aks ettirishda qiyinchiliklar[6].

• *Qoidaviy yondashuvning moslashuvchanligi:* Qo‘lda yangilanish zarurati va keng korpusga moslashtirishda qiyinchiliklar[6].

Kelgusida quyidagi takomillashtirishlar rejalashtirilmoqda:

• Korpus hajmini va xilma-xilligini oshirish;

• Statistik va qoidaviy yondashuvlarni gibridlashtirish;

• Chuqur o‘rganish algoritmlari (LSTM, Transformer) integratsiyasini amalga oshirish.

Ushbu tadqiqotda quyidagi metodologik bosqichlar qo‘llanildi:

• *Ma’lumot yig‘ish:* O‘zbekiston va xalqaro manbalardan olingan korpuslar asosida ma’lumotlar to‘planildi.

• *Avtomatik teglash:* Dastlabki avtomatlashtirilgan algoritmlar yordamida tezkor natija olish, so‘ngra inson nazorati ostida qayta tahrir qilish.

• *Sinov va taqqoslash:* Turli modellarning aniqlik va samaradorligini eksperimental tarzda sinab ko‘rish[7].

Amaliy jarayonda quyidagilar bajarildi:

• *Modellarning o‘rgatilishi:* HMM, CRF va qoidaviy yondashuvlar alohida sinovdan o‘tkazildi.

• *Kombinatsion yondashuvlar:* Ensemble metodlari yordamida yakuniy natija sifatini oshirishga intilish.

• *Vizualizatsiya:* Diagrammalar va jadvallar yordamida natijalar va xatoliklar vizual tarzda aks ettirildi.

• *Neyraviy tarmoqlar:* LSTM va Transformer kabi chuqur o‘rganish metodlari orqali POS teglashsamaradorligini oshirish.

• *Semantik vektor modellar:* Word2Vec, GloVe va BERT kabi metodlardan foydalanish orqali so‘zlar orasidagi semantik bog‘lanishni aniqlash.

• *Ensemble metodlar:* Turli modellarning natijalarini birlashtirish orqali umumiy aniqlikni oshirish[8].

Kelgusida quyidagi yo‘nalishlar kengaytirilishi rejalashtirilmoqda:

• *Hibrid tizimlar:*Statistik, qoidaviy va chuqur o‘rganish yondashuvlarini integratsiya qilish.

• *Korpusni kengaytirish:* Dialektlar, stilistik xususiyatlar va turli mavzularni inobatga olgan korpuslarni yaratish.

• *Real vaqt rejimi:* Interaktiv va mobil platformalarda POS teglash tizimlarini tatbiq etish.

• *Keng qamrovli sinovlar:* Aniqlik, tezlik va resurs samaradorligini o‘lchash bo‘yicha eksperimental sinovlarni kengaytirish.

Taqdim etilgan natijalar asosida quyidagi xulosalar chiqarildi:

• Statistik modellarning ustunligi: CRF va HMM modellari katta hajmdagi korpuslarda samaradorligini isbotladi, ayniqsa CRF modeli kontekstual bog‘lanishlarga asoslanganligi sababli yuqori aniqlikka erishdi.

• Qoidaviy yondashuvning barqarorligi: Qoidaviy usullar barqaror va tushunarli natijalarni ta’minlasa-da, moslashuvchanlik jihatidan cheklanganligi aniqlangan.

• Hibrid metodlar istiqbollari: Statistik va qoidaviy yondashuvlarni chuqur o‘rganish metodlari bilan birlashtirish orqali aniqlik va samaradorlikni yanada oshirish mumkinligi aniqlangan.

Xulosa.

Ushbu maqolada o‘zbek tilidagi ravish so‘z turkumini grammatik POS teglash algoritmlarining nazariy asoslari, metodologiyasi, amaliy tatbiq jarayoni va eksperimental natijalari chuqur tahlil qilindi. Natijalar shuni ko‘rsatadiki, CRF modeli eng yuqori aniqlikni ta’minlasa-da, qoidaviy yondashuvlarning barqarorligi va tushunarli natijalari ham e’tiborga loyiqdir. Kelgusida hibrid yondashuvlar va chuqur o‘rganish algoritmlarini integratsiya qilish orqali real vaqt rejimida ishlovchi, yuqori aniqlikka ega tizimlar ishlab chiqish istiqbollari mavjud.

Foydalanilgan adabiyotlar:

1. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition (3rd ed.). Prentice Hall. P. 95.

2. Manning, C. D., & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press. P. 133.

3. Brants, T. (2000). TnT – A Statistical Part-of-Speech Tagger. In Proceedings of the Sixth Conference on Applied Natural Language Processing. P. 224-231.

4. O‘zbek Tili Morfologiyasi. (2018). Tashkent: Fan va Texnika. P. 23-24

5. Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In Proceedings of HLT-NAACL 2003. P. 173-180.

6. Habash, N. (2010). Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers. P. 8-9.

7. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. In Advances in Neural Information Processing Systems P. 3111-3119.

8. Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP). P. 1532-1543.

MATNNI TASNIFLASH VA ULARNING SOHALAR KESIMIDA QO‘LLANILISH MASALASI

Xusainova Zilola Yuldashevna,
Filologiya fanlari bo‘yicha falsafa doktori
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Shirinboyeva Marjona,
Filologiya yo‘nalishi talabasi
shirinboyevamarjona36@gmail.com
ToshDO‘TAU

Annotatsiya. Matnni tasniflash – bu matnlarni mazmun, uslub, maqsad yoki tarkibiy jihatdan guruhlash jarayoni. Zamonaviy axborot texnologiyalari rivojlanishi natijasida matnni tasniflash usullari ko‘plab sohalarda qo‘llanilmoqda. Xususan, hujjatlarni avtomatik ajratish, elektron pochta spamlarini aniqlash va ijtimoiy tarmoqlardagi kontent monitoringi kabi vazifalarda keng foydalaniladi. Tibbiyotda, yuridik sohada matn tasnifi muhim ahamiyat kasb etadi. Matnni tasniflashda asosan sun’iy intellekt va mashinali o‘qitish texnologiyalari muhim yo‘nalishlaridan biri hisoblanadi. Bu jarayon matnlarni avtomatik ravishda oldindan belgilangan toifalarga ajratishni anglatadi. Masalan, ijtimoiy tarmoqlardagi postlarni sentimentini aniqlashda, ya’ni ijobiy, salbiy yoki neytral deb tasniflashda qo‘llaniladi. Shu bilan birga, NLP metodlari matnlarning ma’nosini chuqur tahlil qilish imkonini beradi. Har bir sohada matn tasniflashning o‘ziga xos xususiyatlari va talablariga to‘g‘ri tanlangan algoritm ishlab chiqiladi.

Abstract. Text classification is the process of grouping texts by content, style, purpose, or structure. As a result of the development of modern information technologies, text classification methods are used in many areas. In particular, they are widely used in tasks such as automatic document sorting, detecting email spam, and monitoring content on social networks. Text classification is of great importance in medicine and the legal field. Text classification is mainly one of the important areas of artificial intelligence and machine learning technologies in text classification. This process involves automatically dividing texts into predefined categories. For example, it is used to determine the sentiment of posts on social networks, that is, classify them as positive, negative, or neutral. At the same time, NLP methods allow for a deep analysis of the meaning of texts. In each area, an algorithm is developed that is correctly selected for the specific characteristics and requirements of text classification.

Аннотация. Классификация текстов — это процесс группировки текстов по содержанию, стилю, назначению или структуре. В результате развития современных информационных технологий методы классификации

текстов нашли применение во многих областях. В частности, он широко используется в таких задачах, как автоматическая сортировка документов, обнаружение спама в электронной почте и мониторинг контента социальных сетей. Классификация текстов важна в медицине и юридической сфере. Классификация текстов — одна из важнейших областей применения технологий искусственного интеллекта и машинного обучения. Этот процесс подразумевает автоматическую классификацию текстов по предопределенным категориям. Например, он используется для определения тональности постов в социальных сетях, то есть для их классификации как положительных, отрицательных или нейтральных. В то же время методы НЛП позволяют проводить глубокий анализ смысла текстов. В каждой области разрабатывается алгоритм, адаптированный под конкретные характеристики и требования классификации текста.

Kalit so‘zlar. *Matnni tasniflash, NLP, ML, axborotni tartiblash, sentiment tahlil, sohaviy tasnif.*

Matnni tasniflash – bu matnlarni ma’lum belgilarga ko‘ra guruhlarga ajratish va ularni avtomatik tarzda toifalash jarayonidir. Bugungi kunda axborot texnologiyalari jadal rivojlanishi natijasida matn ko‘rinishidagi ma’lumotlar hajmi keskin oshmoqda. Ushbu ma’lumotlar oqimini matnlarni tizimli ravishda tasniflash va ularni sohaga oid kategoriyalarga ajratish dolzarb ilmiy-texnik masalaga aylanmoqda. Matnni tasniflash turli mezonlar (mazmun, struktura, til xususiyatlari) asosida matnlarni sistematik ravishda guruhlash jarayonidir. Ushbu jarayon tabiiy tilni qayta ishlash, mashinali o‘qitish va sun’iy intellekt texnologiyalarining rivojlanishi bilan yanada takomillashib bormoqda. Biroq, sifatli ma’lumotlarning yetishmasligi va til xilma-xilligi bu sohada muhim muammolar sifatida qolmoqda. Ushbu maqolada matnni tasniflash jarayonining nazariy asoslarini va ularning turli sohalar kesimidagi amaliy qo‘llanilishini ilmiy asosda tahlil qilishga qaratilgan.

Matnni tasniflash zamonaviy axborot texnologiyalari va sun’iy intellekt sohasining muhim tadqiqot yo‘nalishlaridan biri sifatida shakllangan. Ushbu jarayon matnli ma’lumotlarni oldindan belgilangan toifalar yoki kategoriyalarga avtomatik ravishda ajratishni nazarda tutadi[1]. Matnni tasniflash tabiiy tilni qayta ishlash (Natural Language Processing) va mashinali o‘qitish (Machine Learning) texnologiyalariga tayangan holda amalga oshiriladi. Har bir sohada matnlarning semantik va kontekstual xususiyatlarini hisobga olgan holda moslashtirilgan modellardan foydalaniladi[2]. Shuningdek, matnni tasniflash texnologiyalari axborot oqimini boshqarish, tezkor qaror qabul qilish va umumiy samaradorlikni oshirishda muhim ahamiyat kasb etmoqda.

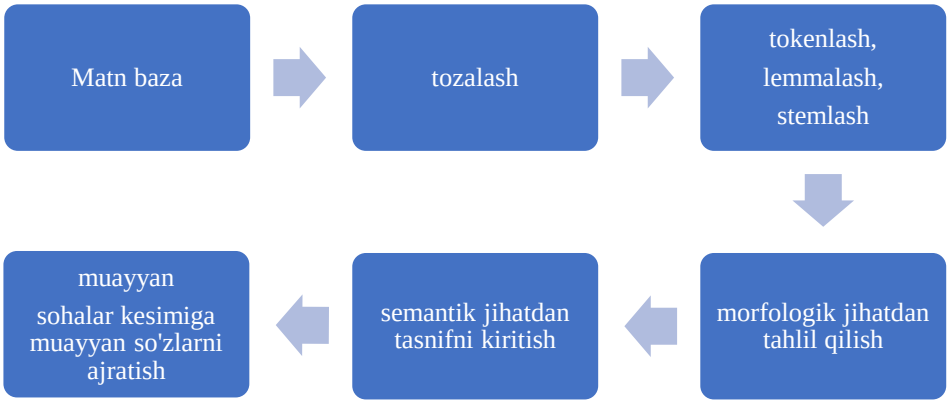
Ayniqsa, matnlarning sohalarda tasniflashda ancha vaqt talab etadi, ammo uning hozirda aniq yechimi mavjud emas. Bo‘lsa ham tahlillarning samaradorlik

ko‘rsatkichi 100% natija bermaydi. Bu kabi muammolar bor ekan ularni tahlil qilish va yechim topishga bo‘lgan qiziqlash va talablar ortib boradi. Matnlarni tasniflash jarayonida ularni avtomatik ravishda toifalarga ajratish uchun ma’lum so‘zlarni matn tarkibidagi kalit so‘zlarga qarab uning sentiment scorini ijobiy, salbiy kabi guruhlarga yoki qaysi sohaga oid ekanligini aniqlab berishda muhim ahamiyatga ega. Masalan: internet tarmog‘ida jurnalistik matnlar, ilmiy yoki badiiy matnlar soniya sayin ko‘payib bormoqda ularni ajratish ko‘proq vaqt talab etadi. Jurnalistik matnlarni ajratishning o‘zi ham bir necha turlarga ajratiladi. Misol tariqasida sport, musiqa, tibbiyot, adabiyot, tabiat haqidagi matnlar. Bular tarkibida ma’lun so‘zlarni ayni shu soha ekanligini o‘rganib ularni ajratadi. Misol uchun, “Bugun yurtimizning janubi-sharqiy qismida yomg‘irli ob-havo bo‘lishi kutilmoqda” gapida ob-havo, yomg‘ir va janubi-sharqiy so‘zlari orqali bu matnning ob-havo haqida ma’lumot berayotganini bilishimiz mumkin.

Matnni tasniflash bo‘yicha olib borilgan dastlabki tadqiqotlar ko‘proq qoidaga asoslangan (rule-based) yondashuvlarga tayangan[3]. Ushbu usullar matndagi muhim kalit so‘zlar va sintaktik tuzilmalarga asoslanib tasniflashni amalga oshirgan. Biroq, qoidaga asoslangan yondashuvlar katta hajmdagi va murakkab ma’lumotlar bilan ishlashda o‘zining cheklovlarini namoyon qilgan. Keyingi bosqichda statistik metodlar va mashinaviy o‘rganish asosidagi yondashuvlar rivojlandi[4]. Xususan, Naive Bayes, Support Vector Machines (SVM) va Decision Trees kabi modellar matnni tasniflash samaradorligini oshirdi. So‘nggi yillarda chuqur o‘rganish (Deep Learning) usullari, xususan, neyron tarmoqlar va transformer modellar (masalan, BERT) matnlarni yanada aniqroq tasniflash imkonini berdi. Masalan, Devlin va boshqalar[5] tomonidan ishlab chiqilgan BERT (Bidirectional Encoder Representations from Transformers) modeli kabi pre-train qilingan til modellarining rivojlanishi matnni kontekstual ravishda chuqur tahlil qilishga imkon yaratdi. Matnni tasniflash bo‘yicha tadqiqotlar tibbiyot, moliya va ijtimoiy tarmoqlar monitoringi kabi ko‘plab sohalarda qo‘llanila boshlandi. Masalan, klinik yozuvlardan kasallikni aniqlash, kreditni baholash kabi amaliy holatlarni kuzatishimiz mumkin. Biroq mavjud metodologiyalarning cheklovlari ham mavjud. Dastlabki usullar kontekst va semantika chuqur tahlil qilinmagani uchun murakkab matn tuzilmalari bilan ishlashda samarasiz bo‘lgan. Shuningdek, kam resursli tillar uchun matnni tasniflash modellari hali ham yetarlicha rivojlanmagan. Mavjud nazariyalar asosan katta hajmdagi strukturalangan ma’lumotlarga tayanadi, bu esa kichik korpuslar uchun qiyinchilik tug‘diradi. Mashinali o‘qitsh modellarining samaradorligi esa ko‘pincha katta hajmdagi, yaxshi belgilangan ma’lumot to‘plamlariga bog‘liqdir. Deep learning asosidagi modellar esa hisoblash resurslariga va ma’lumotga bo‘lgan yuqori talab tufayli cheklovlarga ega. Shuningdek, soha kesimida matnni tasniflashda maxsus domenga moslashtirilgan modellarni ishlab chiqish zarurati tug‘iladi.

Ko‘plab tadqiqotlar umumiy tasniflash modellariga qaratilgan, ma’lum bir soha uchun moslashgan tasniflagichlar hali to‘liq mukammallashmagan. Shuningdek, matnlarning noaniqligi, teglanmagan va imlo xatolariga nisbatan modellar samaradorligini oshirish masalasi yetarlicha yechilmagan.

Sohalar bo‘yicha matn tasniflashning bizga afzal tomonlari shundaki, bu jarayon tez va unumdor bo‘lishi hamda ishchi kuchiga bo‘lgan talabning kamayishidir.



Tadqiqotlardan olingan natijalar matnni tasniflash texnologiyalari turli sohalarda axborot boshqaruvini samarali tashkil etishga xizmat qilishini ko‘rsatadi. Shuningdek, har bir soha uchun moslashtirilgan model va algoritmlarni tanlash samaradorlikni oshiradi. Jurnalistika, onlayn harid, sport, tibiiyot, tilshunoslik va matematika sohalarida aniqlik talablari yuqori bo‘lganligi sababli, chuqur o‘rganishga asoslangan (Deep Learning) yondashuvlar ustunlik qilmoqda.

Online savdo sohasida

Agar biz biror narsani online xarid qilmoqchi bo‘lsak mahsulot bizga qay holatda kelishi, bizga mahsulot qaday ta’sir qilishi, ayniqsa, kosmetika mahsulotlari, kiyim-kechak, qurilish mollari, hozirgi paytda keng ommalashgan audio kitoblar va doimiy reklamalarda aylanayotgan turli xil dori vositalarini sifatini aniqlashdagi mijozlarning fikrlarini bilishda yordam beradi. Bu savdo sohasidagi mijozlardan biri agar “Bu mahsulot menga yoqdi”, “Yaxshi”, “Mahsulot menga manzur keldi”, “Zo‘r”, “Mahsulot sifatli va arzon” kabi fikrlar bildirsa demak bu fikrlarni biz ongli ravishda iliq fikrlar ekanligini anglaymiz. Ammo, buni sun’iy intellektga avtomatik ajratish uchun avvalo, o‘rgatish kerak. “Yoqmadi”, “Menga uncha manzur emas”, yana salbiy baho ifodalovchi bir qancha stikerlardan foydalanish orqali salbiy munosabatlarni bilib olishimiz mumkin. Bu fikrlar munosabatni bildirsa, yana bir jihati borki “buyurtma qilmoqchiman”, “yetkazib berildi”, “mahsulot haqida quyidagi havoladan bilib olishingiz mumkin” kabi bir qancha avtomatik javob berish mumkin bo‘ladi. Keling ijtimoiy tarmoqlardagi ba’zi savdo kanallaridagi commentlardan misol tariqasida ko‘rib chiqamiz.

1. Kompyuter sotadigan Instagram commentida “**qmatku**” tarzidagi fikr qoldirilgan ushbu fikrda biz imloviy xatoni ko‘rishimiz mumkin ammo bu kabi holatlar yana topiladi va buni bazaga kiritish orqali salbiy ma’no berishini bilishimiz mumkin. Yana bir qancha misollar keltiramiz:

“**Qimmat** va **bir tiyingayam qimmat** bu keca 3 millionga coria 5 abrativka 8 xotira 512 360 graduagaca iciladi va sensor 12 avlod.”, — salbiy.

“Men ishlataman **Juda yaxshi** laptop”, — ijobiy.

2. Kiyim do‘kon kommentidan:

🔥🔥 — ijobiy

😔😔 — salbiy

“**juda yaxshi** narx ham sifati ham yaxshi”, ijobiy.

“Yetkazib berish **uncha yaxshi emas** ekan 1 hafta deganda **illa** oldim”, ushbu fikrda salbiy baho bilan birgalikda sheva so‘z **“illa”** ham ishtirok etgan.

Jurnalistika sohasida

Matnlarni mavzu ko‘lami va ma’no jihatdan salbiy yoki ijobiy guruhlarga tasniflashda kerak bo‘ladi. Masalan; “Keldiyorovaga bu mukofot Parij-2024 Olimpiadasida yaponiyalik taniqli dzyudochi Uta Abe ustidan qozonilgan g‘alaba uchun taqdim etildi”, kabi mavzuga doir maqolani biz **“dyuzdochi”, “Uta Abe”** kabi teglar orqali sport mavzuda ekanligini, “Domla sharhlarni o‘qibdi-da, osmonga qarab charaqlagan quyoshni Alisherga, qora bulutni esa Husayn Mirzoga ko‘rsatib, Alisherga “Ofarin” debdi”, jumlasidan adabiyotga oid bo‘lgan “Alisher”, “Husayn Mirzo” teglari orqali bilishimiz, “O‘zbekistonda hafta davomida iliq va asosan quruq ob-havo bo‘lishi kutilmoqda.” jumlasidagi “iliq”, “ob-havo” va “quruq” teglari orqali ob-havo mavzusida ekanligini va albatta, bu kabi keltirilgan fikrlar ijobiy ekanligini aniqlashga yordam bersa, “2025-yil 24-fevral. Bugun Rossiyaning Ukraina zaminiga boshlagan keng ko‘lamli urushiga 3 yil to‘ldi. Bu voqelik nafaqat ikki malakatda balki, butun dunyo bo‘ylab sado berdi: minglab insonlar halok bo‘ldi, ko‘z o‘ngimizda shahar vayronaga aylandi, boshqa davlatlar o‘z tomonlarini hamon aniqlamoqda.”, jumlasidan esa biz bu mavzudagi **“urush”** tokeni orqali urush mavzusida ekanligini va “halok bo‘ldi”, “vayronaga aylandi” kabi so‘zlar orqali esa salbiy ma’no ifodalayotganini bilishimiz mumkin.

Tibbiyot sohasida

Kimdir sog‘lig‘idan shikoyat qilib,

– haroratning ko‘tarilishi, bosh og‘rig‘i, isitma, yo‘tal, tomoq og‘rig‘i kabi belgilar borligini aytsa, bu orqali bemorda grippga chalinganligini osongina aniqlash va unga kerakli dori hamda muolajalar haqida ma’lumot berish mumkin bo‘ladi.

– holsizlik, tez-tez kasal bo‘lish, tezda vazn yo‘qotish, sariqlik, yo‘tal va ovozning bo‘g‘ilishi kabi belgilar esa saratondan darak berishi hamda zudlik bilan davolash ishlarini boshlash haqida ma’lumot va tavsiyalar beradi. Bu kabi misollardan ularni “tibbiy terapiya” “nevrologiya” kabi bolimlarga xos ekanligini bilishingiz mumkin.

– yana bir tomondan “qabulga yozildi”, “x- navbatdagi bemor”, “qon tahlili” kabi tokenlar orqali bir qancha zaruriy ma’lumotlarga ega bo‘lish mumkin.

Tilshunoslik sohasida

Lug‘atlar, imlo tekshiruvi, bosh harflar imlosi, chiziqcha va tire bilan bog‘liq bilan muammolar, punktuatsion belgilar, transkripsiya kabi bir qancha masalalar bor.

– Lug‘atchilikning eng ommalashgan turi ingliz-o‘zbek, rus-o‘zbek, turk-o‘zbek kabi tarjima lug‘atlari hisoblanadi. Ammo arab, fors va tojik kabi izofali tillarda bu biroz oqsayotganini ko‘rishimiz mumkin. Izofalar bizni tilimizga xos bo‘lmaganligi sababli tarjima sohasida ularda qator muammolar kelib chiqyapti.

– Eski o‘zbek tilidan yurtimizda ancha yillar foydalanilganini hisobga olsak, bu yozuvdagi asarlarni hozirgi o‘zbek adabiy tiliga o‘girish va bu jarayondagi ishlar hali o‘z yechimiga to‘la ega emasligini ko‘rimiz mumkin. To‘g‘ri eski o‘zbek tili va yozuvi haqida bir qancha qo‘llanma, lug‘atlar ishlab chiqigan bo‘lsa hamki, ushbu masalani avtomatik hal qilish kerak ekanligi yana bir masaladir.

– Punktuatsion belgilarni to‘g‘ri ishlatilishida hamma ham xatolarga yo‘l qo‘yishi mumkin. Ammo tez-tez ishlatilinuvchi nuqta, vergul, ikki nuqtani ishlatishda ularni uyushiq holatda kelganda verguldan foydalanish, kesimlikni shakklantiruvchi qo‘shimchalardan so‘ng nuqtaning ishlatilishi va uyushiq so‘zlardan oldin,ko‘chirma gapdan oldin qo‘llangan muallif gapidan so‘ng ishlatishni hal etish masalasi ham bor.

– Tire va chiziqcha bilan bog‘liq muammo ham muhim masalalardan biridir. Qay holatda tire, qay holatda chiziqchaning ishlatilishiga ko‘pchilik ikkilanib qoladi. Tire punktuatsion, chiziqcha esa orfografik belgidir. Juft va takror so‘zlar orasida chiziqcha ishlatilishi kabi qator muammolar o‘z yechimini kutmoqda.

Algebra-matematika sohasida

Matn tasnifi berilgan misol, masala, tenglama, tengsizlik, funksiya kabi tur va oson, qiyin, o‘rtacha darajalarga ajratishda muhim ahamiyatga egadir.

Dialektologiya sohasida

Sheva matnlaridagi so‘zlarni adabiy tilga o‘girish, unli harflarni old va orqa qatorga, undoshlarni esa qay holatda talaffuz etilgan bo‘lsa uning adabiy tildagi muqobiliga almashtirishda kerak bo‘ladi.

XULOSA

Matn tasniflash turli sohalarda ma’lumotlarni tahlil qilish va qaror qabul qilish jarayonlarini optimallashtirishning muhim vositasi bo‘lib, AI texnologiyalarining rivojlanishi bilan uning ahamiyati tobora ortib bormoqda. Matnlarni tasniflash kelgusida ham sun’iy intellekt yutuqlari bilan uyg‘un holda rivojlanib, axborot jamiyatining samaradorligini oshirishda muhim rol o‘ynashi kutilmoqda.

Foydalanilgan adabiyotlar:

1. Aggarwal, C. C. (2012). Mining Text Data. Springer.

2. Camacho-Collados, J., & Pilehvar, M. T. (2018). From Word to Sense Embeddings: A Survey on Vector Representations of Meaning. Journal of Artificial Intelligence Research.

3. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys.

4. Joachims, T. (1998). Text categorization with Support Vector Machines: Learning with many relevant features. ECML.

5. Devlin, J., et al. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL.

6. B. Elov, Sh. Hamroyeva, R. Alayev, Z. Xusainova, U. Yodgorov – O‘zbek tili korpusi matnlarini qayta ishlash usullari – Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali, October 2023 – Volume 1, Issue 3.

7. Xusainova, Z., & Yangibayeva, S. (2024). Til korpusi turlari. Uzbekistan: Language and Culture, 2(2).

8. Elov, B. B., Hamroyeva, S., & Xusainova, Z. (2022). NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar. TerDU, filologik tadqiqotlar: til, adabiyot, ta’lim, 5-6.

SLENGLARNI AVTOMATIK TEGLASHDA LUG‘ATGA VA MASHINALI O‘QITISHGA ASOSLANGAN YONDASHUVLAR

Bekmuradova Iroda Zokir qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
irodabekmuradova9@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada slenglarni avtomatik teglashning lug‘atga va mashinali o‘qitishga asoslangan usullari xususida so‘z yuritiladi. Bunda lug‘atga asoslangan usulning bosqichlari ketma-ketlikda ko‘rsatib, mazkur usulning afzalliklari va kamchiliklari keltiriladi. Shuningdek, maqolada slenglarni avtomatik teglash uchun Bayes va SVM (support vector machine) modellari ham taklif etilib, bu modellarga xos dastlabki umumiy tavsiflar ham berilib, ilmiy tahlil qilinadi.

Abstract. This article discusses dictionary-based and machine learning-based methods for automatic tagging of slang. The steps of the dictionary-based method are shown sequentially, and the advantages and disadvantages of this method are presented. The article also proposes Bayesian and SVM (support vector machine) models for automatic tagging of slang, provides initial general descriptions of these models, and conducts scientific analysis.

Аннотация. В этой статье рассматриваются методы автоматической разметки сленга на основе словаря и машинного обучения. Здесь будут последовательно показаны этапы метода, основанного на словаре, а также представлены преимущества и недостатки этого метода. В статье также предлагаются байесовские модели и модели SVM (машина опорных векторов) для автоматической разметки сленга, даются начальные общие описания этих моделей и проводится научный анализ.

Kalit so‘zlar: *sleng, teg, teglash, lug‘at, Bayes algoritmi, Support vector machine.*

Ma’lumki, til korpusida sleng birliklarni ajratib tahlil qilish, ularga avtomatik teglash orqali bunday birliklar qatnashgan matnning u yoki bu tahlili amalga oshirilganda slenglarni oddiy leksik birliklar kabi tahlil qilib ketish, avtomatik tarjima jarayonida so‘zni to‘g‘ridan-to‘g‘ri belgilangan tilga tarjima qilib qo‘yish kabi xatoliklarni to‘xtatishga erishish mumkin.

Til korpuslari o‘z maqsadi va ichki tuzilishiga ko‘ra turlicha bo‘lishiga qaramay, ularni umumlashtiruvchi asosiy xususiyat shundan iboratki, ular til birliklarini avtomatik teglash hamda ularni ma’lum bir mezonlarga ko‘ra tahlil qilish

yoki shunga o‘xshash lingvistik tahlillarni amalga oshirish imkonini beradi. Korpusda tahlil uchun kiritilgan matnlardagi birliklarni teglashga oid turli yondashuvlar va bir qator matematik usullar mavjud. Shulardan asosiylari: qoidaga asoslangan yondashuv, lug‘atga asoslangan yondashuv, mashinali o‘qitishga asoslangan yondashuv, statistik va transformatsion yondashuv va hokazolar [2].

Har qanday til korpusi tilning leksik va grammatik xususiyatlarini chuqur o‘rganish imkonini beradi, til tizimini hamda uning o‘zgaruvchanligini kuzatishga xizmat qiladi. Milliy til korpusi esa, tilning barcha xususiyatlarini aks ettiruvchi og‘zaki va yozma matnlarni o‘z ichiga olgani bois, tilning dinamikasini, so‘zlar va grammatik shakllarning kelib chiqishi hamda qo‘llanishini tahlil qilishda, matnlarning lingvistik xususiyatlarini o‘rganishda muhim ahamiyat kasb etadi.

Matndagi so‘z va iboralarni teglash tabiiy tilni qayta ishlash jarayonida muhim vazifalardan biri hisoblanadi. Teglash orqali so‘zlarning tuzilishi, ma’nosi va ularning kontekstdagi o‘rni haqida batafsil ma’lumot beriladi. Bu jarayonda matn tarkibidagi so‘z yoki iboralarga maxsus izohlar qo‘shiladi. Izohlar tadqiqot maqsadiga qarab, grammatik yoki leksik ma’no, shuningdek, boshqa muhim xususiyatlarni ko‘rsatishi mumkin.

Teglar (inglizchasiga “tags”) – bu matn haqidagi qo‘shimcha ma’lumotni o‘zida mujassam etuvchi belgilardir. Matnli ma’lumotlarni (korpuslarni) teglash uchun bir necha universitetlar matnlarning qaysi parametrlarini teglash kerakligini tavsiflovchi tizimni maxsus ishlab chiqdi. Ushbu tizim XMLdan foydalanadi va Text Encoding Initiative Guidelines (TEI Guidelines) deb nomlanadi. Bu kodlash, teglash va indeksirlash mumkin bo‘lgan matnlarning turli xil xususiyatlarining ro‘yxati hisoblanadi. Masalan, tizim matndagi turli xil tuzatishlar, iqtiboslar, qisqrtmalar, atoqli otlar, initsial, akronimlar, chet el so‘zlari va boshqalarni sanab o‘tadi.[1]

So‘zlarni teglashning lug‘atga asoslangan yondashuvi quyidagi bosqichlarni o‘z ichiga oladi:

1-rasm. Slenglarni avtomatik teglashning lug‘atga asoslangan usuli bosqichlari

Lug‘atga asoslangan teglash usuli oldindan tuzilgan lug‘atlardan foydalangan holda matn birliklariga mos teg yoki toifani belgilashga asoslanadi. Masalan, sleng birliklarini avtomatik teglashda dastlab tildagi sleng birliklari aniqlanib, maxsus lug‘at shakllantiriladi. So‘ngra matn kiritilganida, u lug‘atdagi birliklarga mosligi bo‘yicha tekshiriladi va mos keladigan birliklarga teg biriktiriladi. Ushbu yondashuv nisbatan soddaligi bilan ajralib turadi hamda muayyan matn turlari uchun, xususan, aniq kalit so‘zlar mavjud bo‘lsa, yuqori aniqlikdagi natijalarni ta’minlashi mumkin.

Shu bilan birga, lug‘atga asoslangan yondashuvning ayrim kamchiliklari mavjud. Xususan, teglash natijalarining sifat darajasi lug‘atning to‘liqligi va aniqligiga bevosita bog‘liq. Agar lug‘at barcha zaruriy birliklarni o‘z ichiga olmasa, natijaviy tahlil sifati pasayadi. Shuningdek, kontekstga bog‘liq yoki noaniq birliklarni teglashda muayyan cheklovlarga duch kelinadi. Ushbu usulning samaradorligini oshirish uchun kontekstual lug‘atlardan foydalanish, mashinani o‘qitish modellarini joriy etish yoki bir nechta lug‘atlarni birlashtirish orqali natijalarni yaxshilash mumkin.

Tabiiy tilni qayta ishlash (NLP) sohasida slenglarning avtomatik aniqlanishi va teglanishi muhim vazifalardan biri hisoblanadi. Slenglar norasmiy nutq qatlamiga tegishli bo‘lib, doimiy ravishda o‘zgarib boradi va odatiy lug‘aviy tizimlarga sig‘maydi. Shuning uchun, slenglarni samarali tarzda tasniflash va ularni to‘g‘ri teglash uchun kuchli va moslashuvchan algoritmlar talab etiladi. Slenglarni teglashda ML (mashinali o‘qitish)ga asoslangan yondashuv bu jarayonga kompyuter omilini kiritish, ya’ni mos algoritmlarni qo‘llagan holda, matn tarkibida kelgan slengni o‘ziga xos ehtimoliylik darajasiga nisbatan teglashni o‘z ichiga oladi. Bunday toifalar ierarxik tuzilishga ega bo‘lib, ML ga asoslangan usullar kompyuter lingvistikasida faol qo‘llaniladi. Tabiiy tildagi matnlarni ikki bosqichli tahlil qilishning birinchi bosqichida, lug‘atda mavjud bo‘lgan so‘z tahlil qilinayotgan matnning har bir so‘zi bilan taqqoslanadi. Agar so‘z lug‘atda mavjud bo‘lmasa u holda keying bosqichga o‘tiladi. Ikkinchi bosqichda, jarayonda shakllangan qoidalarga muvofiq mavjud bo‘lgan teglash variantlaridan yagona to‘g‘ri namunasi tanlanadi. Ushbu yondashuvning afzalligi aniqlikning ustuvor bo‘lishidir. Ammo bu jarayonda qoidalar sonining ko‘payishi aniqlikning kamayishiga sabab bo‘lishi mumkin. Avtomatik teglashning mashinali o‘qitish usullari slenglar va shunga o‘xshash noaniq so‘zlarni izohlash muammosini hal qilishda samaralidir. Bugungi kunda jahon kompyuter lingvistikasida ushbu sohada faol izlanishlar olib borilmoqda. Matnni tasniflash, undagi birliklarni teglashga doir bir qator ishlarni kuzatib, shunga amin bo‘ldikki, vaznni multiplikativ tuzatish algoritmi, yashirin semantik tahlil, transformatsiyani o‘rgatishga asoslangan algoritm, differentsial grammatika, qatorlar ro‘yxati usuli, Bayes algoritmi, perseptron usullar kabilar

noaniqlikni bartaraf etish muammosini hal eta oladi [2]. Shulardan Bayes algoritmi va uning matndagi birliklarni teglashga doir xususiyatlarini ko‘rib chiqamiz.

Bayes algoritmi obyektning ehtimolini, uning xususiyatlarini va tegishli bo‘lgan guruhini aniqlashni o‘rganuvchi algoritmdir. XVIII asr yashagan matematik olim Tomas Bayes nomi bilan ataladigan ushbu algoritm bir qator sohalarda keng tarqalgan. Mazkur algoritmning markazida Bayes xulosasi yotadi, bu usul bizga yangi dalillar yoki ma’lumotlar paydo bo‘lganda gipoteza haqidagi farazni yangilash imkonini beradi [7]. U ehtimollik tasniflagichi sifatida ham tanilgan. Masalan, so‘zni muayyan xususiyatlari, unga qo‘shilgan qo‘shimchalar va bog‘lanib kelgan so‘zlarga qarab aniq tahlil qilish qiyin, chunki xuddi shu jihatiga ko‘ra o‘xshash xususiyatlarga ega ko‘plab so‘zlar mavjud. Ammo, bu narsa haqida ehtimolli bashorat qilish mumkin va bu yerda Naive Bayes algoritmi qo‘l keladi. Bayes algoritmi, ayniqsa Naive Bayes usuli tabiiy tilni qayta ishlashda muhim rol o‘ynaydi. NLP da u orqali matnni tasniflash, sentiment tahlil, spamni aniqlash, hujjatlarni turkumlashtirish, matn tilini aniqlash, nomlangan obyektni tanish (NER), matnlarni klasterlash va modellashtirish kabi masalalarni hal qilish mumkin. Bayes nazariyasi berilgan dalillar to‘plamidan (B) muayyan gipotezaga (A) kelish ustida ishlaydi.

Slenglarni avtomatik teglashda qo‘llash mumkin bo‘lgan yana bir model Support Vector Machine (SVM) – nazorat ostidagi o‘rganish (supervised learning) asosida ishlovchi klassifikatsiya modeli bo‘lib, u matnlarni tasniflashda keng qo‘llanilishi bilan mashhur [3]. SVM modelining asosiy maqsadi – ma’lumotlar to‘plamini maksimal gipertekislik (optimal ajratuvchi chiziq) orqali ajratishdir.

SVM modeli kichik va murakkab ma’lumot to‘plamlarida yaxshi natija ko‘rsatgani, chiziqli va chiziqli bo‘lmagan tasniflash vazifalariga moslashuvchanligi [4], so‘zlarning semantik va sintaktik o‘ziga xosliklarini hisobga olgan holda slenglarni teg guruhlariga ajrata olishi kabi xususiyatlari bilan slenglarni teglashda samarali hisoblanadi:

Slenglarni teglash uchun SVM modeli quyidagi bosqichlarda ishlatiladi:

1. Ma’lumotlarni to‘plash va tayyorlash

– Sleng so‘zlar bazasini yaratish (ijtimoiy tarmoqlar, forumlar, chat loglaridan).

– Slenglar teglarini aniqlash va ularni toifalarga ajratish (masalan, yoshlar slengi, texnologik sleng, professional jargon va h.k.).

2. Xususiyatlarni (Features) aniqlash

– TF-IDF (Term Frequency-Inverse Document Frequency) yoki so‘z embedding usullari orqali har bir so‘z uchun vektorli taqdimot yaratish [6].

– Morfologik va semantik belgilarni ajratib olish.

3. SVM modelini o‘qitish va tekshirish

– Olingan xususiyatlar asosida SVM modeli o‘qitiladi.

– Model samaradorligini baholash uchun ma’lumotlar to‘plami trening va test to‘plamlariga bo‘linadi [5].

– F1-score, Precision, Recall kabi metrikalar asosida natijalarni tahlil qilinadi.

4. Natijalarni tahlil qilish va optimallashtirish

– Model natijalarini yaxshilash uchun Hyperparameter tuning (C, kernel turi, gamma) amalga oshiriladi [6].

Agar zarur bo‘lsa, boshqa NLP usullari bilan birlashtirish (masalan, LSTM yoki Word2Vec bilan kombinatsiya qilish) ham mumkin.

Xulosa qilib aytganda, til korpusida sleng birliklarini aniqlash va avtomatik teglash jarayonini to‘g‘ri yo‘lga qo‘yish – zamonaviy til texnologiyalarining eng dolzarb vazifalaridan biridir. Bu nafaqat avtomatik tarjima jarayonidagi xatoliklarni kamaytiradi, balki slenglarni oddiy leksik birliklardan farqlash orqali matnni yanada aniqroq tahlil qilish imkonini beradi. Ayniqsa, lug‘atga asoslangan yondashuv, o‘zining soddaligi va yuqori aniqligi bilan ajralib turadi. Bu yondashuv, to‘liq va dolzarb lug‘at bazasi shakllantirilgan taqdirda, tilshunoslik, kompyuter lingvistikasi hamda matnlarni avtomatik qayta ishlash sohalarida katta imkoniyatlarni ochadi.

Shu bilan birga, texnologiya rivojlanishi bilan bir qatorda, slenglarning o‘zi ham doimiy o‘zgarishda ekani, ularni muntazam yangilab borishni taqozo qiladi. Shuningdek, faqat lug‘atga tayanish yetarli emasligini anglash kerak – kombinatsiyalashgan yondashuvlardan foydalanish, ya’ni statistik, qoidaga asoslangan va mashinali o‘qitishga asoslangan usullarni ham birlashtirish orqali ancha mukammal tizimni yaratish mumkin. Bu esa, kelajakda nafaqat sleng birliklarini, balki boshqa murakkab til hodisalarini ham yanada chuqurroq o‘rganishga zamin yaratadi. Shunday ekan, slenglarni avtomatik teglash bo‘yicha olib borilayotgan izlanishlar nafaqat lingvistik nazariya uchun, balki amaliy sohalar – masalan, sun’iy intellekt yordamida matn tahlili, ma’lumotlarni avtomatik indekslash va hatto ijtimoiy tarmoqlarda til monitoringi uchun ham muhim ahamiyat kasb etadi. Mashinali o‘qitish metodlari yordamida slenglar va ularning kontekstlarini avtomatik tarzda tahlil qilish imkoniyatlarini ko‘rib chiqish mumkin. Bunda mashinali o‘qitish algoritmlari yoki neyron tarmoqlar, deep learning metodlari yordamida sleng so‘zlarini avtomatik tarzda tanib olish tizimini ishlab chiqish muhim ahamiyat kasb etadi. Shuningdek, slenglarning til tizimida qanday o‘zgarishlarni keltirib chiqarganini tahlil qilishda mashinali o‘qitish modellari katta yordam beradi. Bu boradagi tadqiqotlar esa slenglarning tilshunoslikdagi o‘rni va ijtimoiy kontekstga bog‘liq ravishda qanday rivojlanishini yanada chuqurroq o‘rganishga yo‘l ochadi.

Foydalanilgan adabiyotlar:

1. Abjalova M. (2023). Korpus birliklarini teglash va annotatsiyalash masalasi. “Zamonaviy ilm-fan va ta’lim istiqbollari” respublika konferensiyasi materiallari, (pp. 345-351). Toshkent.
2. Демская-Кульчицкая О.М., Семеренко В.Р., Ющенко Р.А. 2001: 72
3. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. European Conference on Machine Learning, 137-142.
4. Cristianini, N., & Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press.
5. Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. JMLR.
6. Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification.
7. An introduction to Naive Bayes Algorithm for beginners. (n.d.). Retrieved from Turing: http://www.turing.com
8. http://www.turing.com

ZAMON QO‘SHIMCHALARINI AVTOMATIK ANIQLASH MODELI

Erkinjonova Go‘zal Baxtiyor qizi,
Komputer lingvistikasi mutaxassisligi magistranti
uzbekistan-24@mail.ru
ToshDO‘TAU

Annotatsiya. Tilshunoslikda fe’l so‘z turkumiga mansub so‘zlarning zamonini aniqlash muhim ahamiyatga ega. Fe’llarning zamonlarini aniqlash orqali matn mazmunini to‘g‘ri tushunish, tarjima qilish va tabiiy tilni qayta ishlash (NLP) sohasida ko‘plab muammolarni hal qilish mumkin. Zamon qo‘shimchalarini aniqlashda lingvistik jihatdan o‘zbek tilidagi fe’l zamon kategoriyasining morfologik xususiyatlari tahlil qilinadi. Amaliy jihatdan esa qoidalarga asoslangan metod yordamida modelni qurish jarayoni tavsiflanadi.

Abstract. In linguistics, it is important to determine the tense of words belonging to the verb group. Determining verb tenses can solve many problems in text comprehension, translation, and natural language processing (NLP). Morphological characteristics of the verb tense category in the Uzbek language are analyzed from a linguistic point of view when determining the tenses. In practical terms, the process of building a model using the rule-based method is described.

Аннотация. В лингвистике важно определять время слов, принадлежащих к группе глаголов. Определение времен глаголов может решить многие проблемы понимания текста, перевода и обработки естественного языка (NLP). Морфологические характеристики категории времени глагола в узбекском языке анализируются с лингвистической точки зрения при определении времен. В практическом плане описывается процесс построения модели с использованием метода, основанного на правилах.

Kalit so‘zlar: *Zamon qo‘shimchalari, qoidalarga asoslangan metod, NLP, Python.*

Kirish

Dunyo tilshunosligida XX asr oxirlarida til korpuslarini tuzish ularni lingvistik annotatsiyalash dolzarblasha boshladi. XXI asrda ko‘plab so‘zlarni o‘z ichiga olgan korpuslar paydo bo‘la boshladi. Natijada fe’l zamonlarining avtomatik tahlil qilish muhim masalaga aylana boshladi. Fe’l so‘z turkumida zamon qo‘shimchalarini aniqlash biroz murakkab jarayon hisoblanadi, sababi fe’l zamonlarining ko‘chishi, konteksda boshqa bir zamonga ishora qilishi avtomatik tahlilda qiyinchiliklar keltiradi.

Metodologiya

Jahon korpus linvistikasida zamon shakllarini ko‘plab olimlar tadqiq etishgan sirasiga M.Verhagen, G.Marsik, N. UzZaman, L.Bierkandt, V. Gast kabi olimlar faqat muayyan grammatik shakllar orqali ifodalangan zamon ma’nosi emas, balki turli vaqt ma’noli lug‘aviy birliklar, shuningdek, birikmalar va gaplarning o‘zaro munosabatidan kelib chiqadigan zamon ma’nosini modellashtirish masalalari ham o‘rganiladi. O‘z tadqiqotlarida korpus va uning tuzilishi, razmetka standartlari, matnning avtomatik morfologik va sintaktik tahlilini amalga oshirish kabi masalalarga e’tibor bergan olimlar V.P.Zaxarov, S.Y.Bogdanova, A.B.Kutuzov, Y.V.Nedoshivina, D.Y.Gruzdev, I.O.Kuznetsov, I.M.Nojov kabilardir.

O‘zbek tilshunosligida kompyuter lingvistikasi, matnga leksikografik ishlov berish va lingvostatistik tahlil etish borasida ko‘plab tadqiqotlar olib borilgan. Jumladan, O‘zbek korpus lingvistikasida ham zamon shakllarini annotatsiyalash yuzasidan fikrlar morfologik teglash, morfoanalizator ishlab chiqish masalalari doirasidagi tadqiqotlarda bildirilgan. Sh.Hamroyeva “O‘zbek tili morfologik analizatorining lingvistik ta’minoti” mavzusidagi dissertatsiyasida o‘zbek tili morfoanalizatori uchun grammatik kategoriyalarni o‘zbek tili qonuniyatlari asosida qat’iy belgilash lozimligini ta’kidlab, o‘zbek tilidagi 11 ta grammatik kategoriyani, jumladan, zamon kategoriyasini ham ko‘rsatib o‘tadi [1]. M.Abjalova o‘zbek tili matnlarining avtomatik tahlili va tahriri bo‘yicha amalga oshirgan tadqiqotida tahlilning morfologik bosqichi doirasida zamon shakllariga ham to‘xtaladi [2]. Sh. Hamroyeva, O‘.Xoliyorov, G.Abdualimovalar tomonidan tuzilgan “Fe’l shakllari lug‘ati”da zamonning o‘zida 49ta, variantlari bilan esa 66ta shakl mavjudligi qayd etiladi [3]. Zamon kategoriyasi shakllari to‘liqsiz fe’llar bilan hosil qilingan analitik shakllar hisobiga shunday miqdorni ko‘rsatishi mumkin. A.Q.Po‘latov, T.P.Mo‘minova, I.O.Po‘latovalarning “Dunyoviy o‘zbek tili” kitobida ingliz tilida bir fe’l taxminan 100ta shaklga ega bo‘lgani holda o‘zbek tilida uning shakllari 100000dan oshishi haqida ma’lumot beriladi [4].

Asosiy qism

O‘zbek tilida zamon fe’lning muhim grammatik kategoriyalaridan biri bo‘lib, uch asosiy shaklga ega: Hozirgi zamon, o‘tgan zamon, kelasi zamon.

№	Zamon shakllari	Qo‘shimchalari
1.	Hozirgi zamon	-yap,-yotir/ ayotir, -yotib/ ayotib, -moqda

2.	O‘tgan zamon	-di, -gan/ kan/ qan, -b/ ib, -r edi/ ar edi, //rdi/ ardi, -yotgan edi/ ayotgan edi, -yotib edi/ ayotib edi, -guvchi edi, -ydigan edi/ adigan edi, -moqda edi, -moqchi edi.
3.	Kelasi zamon	- y/ a, - ar, -r, -moqchi, -y/ ay - y/ a, - ar, -r, -moqchi, -y/ ay

Zamonni ifodalovchi grammatik shakllarning annotatsiyasi morfologik tahlil uchun muhim ahamiyat kasb etadi. Zamon shakllarining morfologik tahlilda ya’ni POS taggingda fe’l so‘z turkumiga oid so‘zshaklga grammatik xarasteristika berishda muhim hisoblanadi. Zamon qo‘shimchalarini avtomatik tahlil qilish uchun formal grammatikadan foydalaniladi. Sababi formal grammatika asosi o‘zak va affiks qoidalardan iborat.

Qoidalarga asoslangan model (Rule-based) orqali zamon kategoriyasini avtomatik aniqlash mumkin. Ushbu usul lingvistik qoidalarga asoslangan holda ishlaydi. Qoidalarga asoslangan yondashuv aniq va tushunarlilik darajasi yuqori bo‘lib, aynan morfologik tahlil talab qilinuvchi tizimlar uchun samarali yondashuv hisoblanadi.

Qoidalarga asoslangan model quyidagicha ishlaydi:

1. O‘zbek tilidagi zamonni ifodalovchi grammatik shakllarning tizimli ro‘yxatini shakllantirish.

2. Berilgan matndan fe’llar ajratib olinadi

3. Matn tarkibidan zamonni ifodalovchi grammatik shakllarni topishga xizmat qiluvchi lingvistik modellar ishlab chiqilib va ular asosida zamon shakllarini aniqlanadi.

4. Qoidalarga asoslanib qaysi zamon ekanligi aniqlanadi.

Zamon qo‘shimchalarini avtomatik aniqlashda python dasturlash tilidan foydalanildi. Bunda asosiy vazifa shaxs-son qo‘shimchalarini olib tashlash, fe’lning oxiri o‘tgan, hozirgi yoki kelasi zamon qo‘shimchalari bilan tugashini tekshirish va aniqlangan fe’lni tegishli zamon ro‘yxatiga joylashtirish bilan yakunlandi. Zamon qo‘shimchalarini avtomatik aniqlashda quyidagi metodlardan foydalanildi.

Funksiyalar	**Vazifasi**
.split()	Matnni so‘zlarga ajratish uchun ishlatilgan.

.endswith(qoshimcha)	So‘z oxiridagi qo‘shimchalarni tekshirish uchun ishlatilgan
[:-len(qoshimcha)]	So‘zdan qo‘shimchani olib tashlash uchun ishlatilgan
append()	So‘zlarni yangi ro‘yxatlarga qo‘shish uchun ishlatilgan.
for ... in ...:	Ro‘yxatlar ustida aylanish uchun ishlatilgan.
break	So‘zdan qo‘shimcha topilganda, keraksiz tekshiruvlarni to‘xtatish uchun ishlatilgan.

Foydalanilgan metodlar orqali ushbu natijaga erishildi.

Misol (*bordim, bormoqchisan, qilmoqdamiz*)

O‘tgan zamon so‘zlari:

Asosiy: *bor*, Qo‘shimcha: -di

Hozirgi zamon so‘zlari:

Asosiy: *qil*, Qo‘shimcha: -moqda

Kelasi zamon so‘zlari:

Asosiy: *bor*, Qo‘shimcha: -moqchi

Xulosa

O‘zbek tilshunosligida zamonni ifodalovchi grammatik shakllar va ularning tasnifi munozarali masalalardan hisoblanadi. Korpus tilshunosligida zamon kategoriyasini aniqlash uchun uning grammatik darajasi, semantik darajasi va kontekstual darajasini qamrab olish zarur. Natijada zamon qo‘shimchalarini matnda aniqlash birmuncha osonlashadi.

Foydalanilgan adabiyotlar:

1. Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таьминоти: Филол. фан. док. ...дис. – Тошкент, 2021. – Б. 150.
2. Абжалова М. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун): Филол. фан. бўйича фалсафа доктори (PhD) …дисс. – Фарғона, 2019. – Б.125–126.
3. Hamroyeva Sh., Xoliyoyorov O‘., Abdualimova G. O‘zbek tilining fe’l shakllari lug‘ati. (“kiymoq” fe’li misolida) I jild. – Termiz, 2021. – B. 19.
4. Пўлатов А.Қ., Мўминова Т.П., Пўлатова И.О. Дунёвий ўзбек тили. Ўзбек тилида феъл шакллари ва уларнинг рус, инглиз тилидаги кўринишлари. – Тошкент: Universitet, 2003. – Б. 3.

5. Elova D. Fe’l grammatik shakllarining uslubiy xoslanishiga ko‘ra lingvistik annotatsiyalash masalasi xususida.
6. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Бухоро, 2018. – Б. 3.
7. Elov B., Hamroyeva Sh., Abdullayeva O., Uzakova M. O‘zbek tilida POS tegging masalasi: muammo va takliflar / Uzbekistan: Language and Culture. Applied philology. 2022/2(5). – 51-68-b.
8. Ғуломов А. Феъл. – Тошкент: Фан, 1954. – Б. 19–38
9. Сулаймонов А., Ҳожиев А., Жўраева Ж. Феъл замонлари. – Тошкент, 1962. – 168 б.
10. https://ucrel.lancs.ac.uk/papers/ClawsWordTaggingSystemRG87.pdf
11. https://uznatcorpara

UDK: 811.512.133`42

TURKIY TILLARDAGI TREEBANKLAR KLASSIFIKATSIYASI

Bozorqulova O‘g‘iloy Erkin qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
ogiloybozorqulova62@gmail.com
ToshDO‘TAU

Annotatsiya. Mazkur tadqiqotda turkiy tillar uchun yaratilgan treebanklarning tasnifi o‘rganilib, ularning lingvistik xususiyatlari hamda qo‘llanilish sohalari atroflicha tahlil qilinadi. Turkiy tillarning grammatik tuzilishi va so‘z yasash xususiyatlari va ushbu tushunchalar treebanklarga qay darajada ta’sir etishiga ham alohida to‘xtalib o‘tiladi.

Abstract. This study analyses the classification of treebanks created for Turkic languages, their linguistic features, and their scope of use. A particular focus is placed on the extent to which the grammatical structure and word formation features of Turkic languages affect treebanks.

Аннотация. В данном исследовании анализируется классификация древесных банков, созданных для тюркских языков, их лингвистические особенности и сфера использования. Особое внимание уделяется тому, в какой степени грамматическая структура и особенности словообразования тюркских языков влияют на банки деревьев.

Kalit so‘zlar: *Treebank, turkiy tillar, sintaktik tahlil, morfologik teglash, tabiiy tilni qayta ishlash, avtomatik analiz.*

Zamonaviy kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasida treebanklar (sintaktik korpuslar) muhim resurs sifatida ajralib turadi. Treebank – bu tabiiy tilni qayta ishlash va kompyuter lingvistikasi uchun maxsus ishlab chiqilgan korpus bo‘lib, undagi har bir gap sintaktik tahlil qilinib, daraxt strukturasida (tree structure) annotatsiya qilingan. Boshqacha qilib aytganda, treebank faqat matnlarni saqlab qolishdan tashqari, ularning grammatik tuzilishi, jumladan, so‘z turkumlari, fraza strukturalari va sintaktik bog‘lanishlari haqidagi ma’lumotlarni ham taqdim etadi. Bu resurslar zamonaviy lingvistik tahlil va NLP modellari uchun asosiy vosita sifatida qo‘llaniladi. Treebankning asosiy xususiyatlari quyidagilardan iborat: sintaktik annotatsiyalash jarayonida har bir so‘zga uning grammatik roli (masalan, ot, fe’l, sifat va boshqalar) va sintaktik bog‘lanishlar (qaysi so‘z qaysi so‘z bilan bog‘langanligi) aniqlanadi. Misol uchun, “Kitobni stolga qo‘ydim” gapini tahlil qilsak, “kitobni” - to‘ldiruvchi, “stolga” - hol,

“qo‘ydim” esa kesim sifatida belgilangan. Treebank ushbu grammatik va sintaktik ma’lumotlarni tizimli ravishda tashkil etib, lingvistik tahlillar hamda NLP tadqiqotlari uchun muhim manba vazifasini bajaradi.

Daraxt shaklidagi strukturalash jarayonida gap tarkibidagi so‘zlar o‘rtasidagi sintaktik munosabatlar daraxt shaklida ifodalanadi. Ushbu yondashuv sintaktik tahlilning aniq va tizimli tasvirini taqdim etib, har bir so‘zning grammatik roli hamda boshqa so‘zlar bilan bog‘lanish xususiyatlarini vizual ravishda aks ettiradi. Daraxt strukturalash tabiiy tilni qayta ishlash va lingvistik tahlil uchun asosiy vosita sifatida xizmat qiladi. Treebanklar constituency (bo‘laklar) yoki dependency (bog‘lanishlar) asosida qurilishi mumkin. Bunday strukturalar har bir gapning sintaktik munosabatlarini aniq va tizimli shaklda tasvirlash imkonini beradi. Treebanklar tabiiy tilni qayta ishlash sohasida juda muhim manba hisoblanib, NLP modellarini (masalan, sintaktik tahlil dasturlari va mashina tarjimasi tizimlari) o‘qitishda qo‘llaniladi. Shuningdek, ular til qonuniyatlarini o‘rganish, grammatik xatolarni aniqlash va tuzatish kabi vazifalarni amalga oshirishda keng foydalaniladi. Ushbu resurslarning amaliy ahamiyati lingvistika va kompyuter fanlari kesishmasida katta o‘rin egallaydi. Treebanklar matnlarning grammatik tuzilishini daraxt shaklida (masalan, dependency parsing yoki constituency parsing) aniqlash orqali mashinalarga tilni tushunish va generatsiya qilish imkoniyatini taqdim etadi. Shunga qaramay, dunyoda 180 milliondan ortiq so‘zlashuvchiga ega bo‘lgan turkiy tillar (jumladan, turk, o‘zbek, qozoq, uyg‘ur, qirg‘iz va boshqalar) uchun treebank resurslari nisbatan kam bo‘lib, ular tizimli klassifikatsiyaga ega emas. Ushbu tadqiqotning maqsadi turkiy tillar uchun yaratilgan treebanklarni lingvistik, texnologik va annotatsiya jihatlari asosida tasniflash, mavjud resurslarning afzalliklari va kamchiliklarini aniqlash hamda kelajakdagi loyihalar uchun tavsiyalar ishlab chiqishdan iborat.

Turkiy tillar o‘ziga xos morfologik va sintaktik xususiyatlari bilan ajralib turadi. Ushbu xususiyatlar treebank yaratishda Yevropa tillari (masalan, ingliz tili) uchun ishlab chiqilgan an’anaviy annotatsiya sxemalarini moslashtirish jarayonida qiyinchiliklar keltirib chiqaradi. Masalan, Universal Dependencies (UD) loyihasini turkiy tillarga moslashtirish bir qancha muammolarni yuzaga keltiradi, chunki bu tillarning morfosintaktik tuzilishi UD talablariga to‘liq mos kelmaydi. Natijada, turkiy tillar uchun treebanklarni alohida tasniflash va ularga moslashtirilgan annotatsiya qoidalarini ishlab chiqish zarurati paydo bo‘ladi. Shunga qaramay, turkiy tillar uchun treebanklarning yetishmasligi mintaqada NLP texnologiyalarining sekin rivojlanishiga sabab bo‘lmoqda. Misol uchun, o‘zbek tilida ishlaydigan ChatGPTga o‘xshash modellar hali cheklangan darajada faoliyat olib bormoqda, chunki ular o‘zbek tilining murakkab sintaktik tuzilishini to‘liq o‘zlashtira olmagan. Ushbu holat turkiy tillar uchun chuqur lingvistik tadqiqotlar va zamonaviy NLP resurslarini yaratish zarurligini yaqqol ko‘rsatadi. Hozirgi vaqtda

turkiy tillar ichida faqat turk tili uchun keng qamrovli treebank resurslari mavjud. Ulardan biri – Turkish Web Treebank [2], bu 1 million so‘zdan iborat korpus bo‘lib, Universal Dependencies (UD) sxemasi asosida annotatsiya qilingan. Shuningdek, METU-Sabancı Treebank [3] turk tilining morfologik tahliliga qaratilgan. Boshqa turkiy tillar, jumladan uyg‘ur va qozoq tillari uchun treebanklar faqat akademik loyihalar doirasida yaratilgan. Masalan, Kazakh Dependency Treebank [4] 30 000 so‘zdan iborat bo‘lib, asosan gazeta matnlarini o‘z ichiga oladi.

Jahonning ko‘zga ko‘ringan tillari, jumladan ingliz, nemis va fransuz tillarida treebank dasturlarini yaratish dastlab bir qancha muammolarni hal qilishni talab etgan. Turkiy tillarda treebank tizimini yaratishda esa tilning o‘ziga xos imkoniyatlari va lingvistik xususiyatlarini inobatga olgan holda ish olib borish zarur. Xususan, morfologik murakkablik va sintaktik moslikni Universal Dependencies (UD) annotatsiyasiga moslashtirish masalasi dolzarb ahamiyat kasb etadi. Shu bilan birga, turkiy tillarning morfosintaktik tuzilmalari UD talablariga to‘liq mos kelmasligi sababli bir qancha nomuvofiqliklar yuzaga keladi. Bu nomuvofiqliklar sintaktik tartib va morfologik shakllar asosida namoyon bo‘lib, treebank tizimini ishlab chiqish jarayonida maxsus yondashuvlarni talab qiladi. Chunki turkiy tillarda so‘zlar affikslar orqali ko‘plab grammatik ma’nolarni ifodalaydi. Bundan tashqari turkiy tillarda sintaktik tartib quyidagi ko‘rinishda belgilanadi: SOV – “Men [subyekt] kitob [obyekt] o‘qiyman [fe’l]. Bu kabi xususiyatlar esa treebank annotatsiyasini murakkablashtiradi. Ammo bu kabi muammolarga qaramasdan turkiy tillarga mansub ba’zi tillarda ushbu tizim yaratilgan bo‘lib, bugungi kunda tizimni rivojlantirish uchun say-harakatlar amalga oshirilmoqda. Quyida turkiy tillarda mavjud treebanklar haqida asosiy ma’lumotlar bilan to‘ldirilgan jadval keltirilgan.

1-jadval. Turkiy tillarda mavjud treebanklar tasnifi

Til	Treebank nomi	Korpus hajmi	Annotatsiya sxemasi	Mavjudligi	Xususiyatlar/ kamchiliklari
Turk tili	Turkish Web Treebank (TWT)	1 million so‘z	Universal Dependencies (UD)	Ochiq manba	Murakkab morfologiya, ko‘p qatlamli annotatsiya
Turk tili	METU – Sabanci Treebank	500 000 so‘z	LFG (Lexical-Functional Grammar)	Akademik loyiha	Morfologik tahlilga qaratilgan
Qozoq tili	Kazakh UD Treebank	30 000 so‘z	Universal Dependencies	Ochiq manba	Asosan gazeta matnlari qamrab olingan

Uyg‘ur tili	Uyghur UD Treebank	20 000 so‘z	UD + alohida morfologik belgilar	Tajribaviy	NLP vositalari to‘liq qo‘llanilmagan
Qirg‘iz tili	-	-	-	Yo‘q	Hech qanday treebanklar mavjud emas
Turkman tili	-	-	-	Yo‘q	Faqat lug‘atlar va grammatik qo‘llanmalar
Tatar tili	Tatar Corpus (Arenas et al.)	10 000 so‘z	UD (qisman)	Cheklangan	Noto‘liq annotatsiya, asosan folklor matnlari

Yuqoridagi treebanklar orasida faqat turk va qozoq tillari uchun nisbatan tizimli treebanklar mavjud. Shunga qaramay, mavjud treebanklar bir qator kamchiliklarga ega, jumladan, matn turlarining bir xilligi (asosan gazeta va adabiy matnlar bilan cheklangan) va dialektlar hamda shevalarni qamrab olmasligi. Lotin-kirill alifbosidagi tillar, xususan o‘zbek tili uchun esa Universal Dependencies (UD) annotatsiyasini moslashtirish nisbatan osonroq kechadi. UD qoidalariga ko‘ra, standart POS teglar mavjud bo‘lib, bu teglar so‘z turkumlarining nomlari orqali (masalan, noun, verb, adjective) annotatsiya qilinadi. Shuningdek, UDda so‘z tartibi SVO (subject-verb-object) formatida belgilanadi. Masalan, “I read a book” gapida I – subject (ega), read – verb (kesim), a book – object (to‘ldiruvchi) sifatida aniqlanadi. Ushbu belgilash tizimi dunyo tillarining ko‘pchiligi uchun mos keladi. Biroq, turkiy tillarning o‘ziga xos xususiyatlari bor. Turkiy tillar agglyutinativ tillar guruhiga mansub bo‘lib, bu tillar murakkab morfologik belgilarni o‘z ichiga oladi. Bundan tashqari, turkiy tillarning so‘z tartibi SOV (subject-object-verb) shaklida bo‘lib, bu dunyo tillarining ko‘pchiligidagi SVO tartibidan farqlanadi. Ushbu morfosintaktik xususiyatlar turkiy tillar uchun UD annotatsiyasini kengaytirish zaruratini ko‘rsatadi. Global tajribani turkiy tillarga tatbiq qilish orqali ham muammoni hal qilish mumkin. Masalan, boshqa agglyutinativ tillar, jumladan fin va venger tillari uchun Fin UD treebankida morfologik belgilar XPOS va UPOS annotatsiyasi orqali qo‘shilgan[1]. Japanese BCCWJ [5] korpusiga esa matn turlari (rasmiy va norasmiy) bo‘yicha segmentatsiya kiritilib, tizimdagi muammolar bartaraf etilgan. Ushbu tajribalar turkiy tillar uchun ham Universal Dependencies (UD) annotatsiyasini yaxshilashda qo‘llanilishi mumkin.

2-jadval. UD va Turkiy tillar annotatsiyalarining xususiyatlari taqqoslangan.

Universal Dependencies	**Turkiy tillar annotatsiyasi**
Standart POS teglar	Murakkab morfologik belgilar
SVO sintaktik tartibi	SOV sintaktik tartibi

Global bog‘lanishlar | Tilga xos bog‘lanishlar

Treebanklarni yaratishga qaratilgan ilmiy ishlar va maqolalarda turkiy tillar va UD annotatsiyalari orqasidagi farqlar va o‘zaro moslashtirish masalasi ilgari suriladi. Quyidagi jadvalda Universal Dependencies va Turkiy tillar annotatsiyasi o‘rtasidagi farqlar keltirib o‘tilgan.

3- jadval. Universal Dependencies va Turkiy tillar annotatsiyasi o‘rtasidagi asosiy farqlar

Xususiyat	**UD standarti**	**Turkiy tillar (Moslashtirilgan)**	**Misollar**
Morfologik annotatsiya	Oddiy Pos teglar (Noun, Verb kabi)	Qo‘shimcha morfologik belgilar (Case, Voice).	*Uyg‘ur: Noun+Case=Genitive*
Sintaktik bog‘lanishlar	Asosiy bog‘lanishlar (nsubj, obj, obl).	Tilga xos bog‘lanishlar (compound:lvc,aux:q).	*O‘zbek: compound: lvc (yordamchi fe’l).*
So‘z tartibi	SVO (Subject-Verb_Object) ga mos	SOV (Subject-Object-Verb) bilan moslashtirilgan.	*“Men kitob o‘qidim” nsubj (o‘qidim, Men).*
Affikslar	So‘z sifatida ajratilmaydi	Morfemalarga bo‘linishi mumkin.	*Turk: “okudugum” “oku+dug+um”.*
Case marking	Case xususiyati oddiy (Nom, Acc, Gen).	Batafsil case belgilari (Abl, Loc, Equ).	*Qozoq: Case = Equ*
Qo‘llanilish darajasi	100+ tilda standart	Faqat 5-6 turkiy tilda (turk, qozoq).	*Uyg‘ur UD beta versiyasida*

Yuqoridagi jadvaldan kelib chiqadiki, Universal Dependencies (UD) loyihasi global miqyosdagi tillarning turli xususiyatlarini qamrab oladi, bu esa UD tizimining odatiy morfologik belgilar to‘plamidan ham ko‘rinadi. Turkiy tillarga moslashtirilgan UD sxemasiga qo‘shimcha morfologik belgilar kiritilishi tilning o‘ziga xos xususiyatlarini hisobga olish zaruratini ifodalaydi. Masalan, **case** turkiy tillarda so‘z turkumining holatini belgilash vazifasini bajaradi, **voice** esa fe’llarning modallik belgilarini aniqlash uchun ishlatiladi. Bu kategoriyalar o‘z ichida ichki guruhlarga bo‘linishi mumkin, bu esa tilni yanada chuqurroq tahlil qilish imkonini beradi. Sintaktik tartib masalasi ham bu jarayonda muhim o‘rin egallaydi. UD standartlariga ko‘ra so‘zlar morfemalarga ajratilmaydi, biroq turkiy tillar o‘zining grammatik o‘ziga xosliklari sababli bu talabni qo‘llab-quvvatlamaydi. Agglyutinativ xususiyatga ega bo‘lgan turkiy tillarda murakkab so‘zlarni morfemalarga ajratish zarur. Masalan, qozoq tilidagi “барыпжатырмын” (borayotirman) so‘zi quyidagicha morfemalarga bo‘linadi: **бар** – fe’l (bor), - **ып** – zamon qo‘shimchasi,

- **жатыр** – hozirgi zamon qo‘shimchasi, - **мын** 1-shaxs shaxs-son qo‘shimchasi (-man). Ushbu xususiyatlar turkiy tillarning grammatik tizimini aniq va to‘liq tasvirlash uchun morfemalashni talab qiladi. Bundan tashqari, UD tizimi dunyo bo‘ylab 100 dan ortiq tillarga moslashtirilgan va faol qo‘llanilmoqda. Ammo turkiy tillar oilasidagi tillar faqat cheklangan darajada moslashtirilgan variantlaridan foydalanmoqda. Bu holat turkiy tillarning lingvistik o‘ziga xosliklarini inobatga olgan holda, UD sxemasini yanada kengaytirish zarurligini ko‘rsatadi. Bu o‘z navbatida turkiy tillar uchun mustahkam va moslashtirilgan treebank resurslarini yaratishda muhim ahamiyat kasb etadi.

Xulosa qilib aytganda, turkiy tillar uchun yaratilgan treebanklar lingvistik va kompyuter lingvistikasi nuqtayi nazaridan alohida ahamiyatga ega. Ushbu resurslar sintaktik tahlil jarayonini avtomatlashtirish, tabiiy tilni qayta ishlash tizimlarini rivojlantirish va turkiy tillarning grammatik xususiyatlarini chuqur tadqiq qilishda muhim manba bo‘lib xizmat qiladi. Turkiy tillar agglutinativ xususiyatlari bilan ajralib turgani sababli, ularga mos treebanklar so‘z shakllari, morfologik tahlil va sintaktik tuzilish nuqtayi nazaridan murakkablik kasb etadi. Mazkur tasnif turkiy tillar uchun mavjud treebanklarning tuzilishini, ularning lingvistik modelga asoslangan yondashuvlarini hamda annotatsiya tizimlarining farqlarini aniq ko‘rsatadi. Shu bilan birga, treebanklar o‘rtasidagi o‘xshashlik va farqlar tahlil qilinib, ularning tabiiy tilni qayta ishlashdagi samaradorligi baholangan. O‘rganilgan treebanklar turli standartlarga asoslangan bo‘lib, ba’zilari Universal Dependencies (UD) formatida ishlab chiqilgan, boshqalari esa milliy lingvistik an’analarga tayangan. Natijada, turkiy tillar uchun treebank yaratish jarayonida umumiy qoidalar va standartlar katta ahamiyat kasb etishi ta’kidlangan. Shu bilan birga, har bir tilning o‘ziga xos grammatik va morfologik xususiyatlarini hisobga olish zarurligi ham alohida qayd etilgan. Bunday yondashuv turkiy tillar uchun mos treebank resurslarini yaratishga imkoniyat beradi.

Foydalanilgan adabiyotlar:

1. Pyysalo S., Kanerva J., Missila A., Laipala V., Jinter F. The Finnish universal dependencies Treebank/ Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015). – Finlandiya, 2015. 164-165-p.
2. Pamay T., Sulubacak U., Torunog‘lu-Selamet D., Eryigit G. The annotation process of the ITU WEB Treebank / Proceedings of LAW IX - The 9th Linguistic Annotation Workshop. – Colorado, 2015. 95-p.
3. Sulubacak U., Eryigit G., Pamay T. A revisted dependency Turkish Treebank/ Proceedings of TurCLing 2016, the 1st International Conference on Turkic Computational Linguistics. – Turkey, 2016. 4-p.

4. Tyers F., Washington J. Towards a free/open-source universal-dependency treebank for Kazakh / 3rd International Conference on Computer Processing in Turkic Languages (TURKLANG 2015). – Kazan, 2015. 3-p.

5. Omura M., Asahara M. UD-Japanese BCCWJ: Universal Dependencies Annotation for the Balanced Corpus of Contemporary Written Japanese / Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). – Brussel, 2018. 117-p.

O‘ZBEK TILIDAGI SO‘Z BIRLIKLARINI OLMOSH SO‘Z TURKUMI BO‘YICHA POS TEGLASHDAGI MUAMMOLAR VA YECHIMLAR

Karimova Zilola Dilmurod qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
karimovazilola791@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tilidagi so‘z birliklarini Olmosh so‘z turkumi doirasida avtomatik teglash (POS tagging) orqali aniqlashdagi asosiy muammolar va ularning yechimlari tahlil qilinadi. Xususan, olmoshlarning morfologik shakllari, kontekstual ko‘rinishlari va leksik-semantik xususiyatlari tufayli yuzaga keladigan xatolar ko‘rib chiqiladi. Muammolarning texnik va lingvistik jihatlari o‘rganilib, tegishli qismlarni takomillashtirish bo‘yicha takliflar bildiriladi.

Abstract. This paper analyzes the core challenges of automatic part-of-speech (POS) tagging of pronouns in the Uzbek language. It focuses on morphological forms, contextual usage, and lexical-semantic characteristics of pronouns that often cause tagging errors. Statistical data and diagrams based on annotated Uzbek corpora are presented, and several practical solutions are proposed to enhance model accuracy.

Аннотация. В данной статье рассматриваются основные проблемы автоматической части речи (POS tagging) для местоимений в узбекском языке. Особое внимание уделяется морфологическим формам местоимений, их контекстуальным проявлениям и лексико-семантическим особенностям, вызывающим ошибки в системах автоматической обработки. На основе аннотированных корпусов узбекского языка проведён статистический анализ и предложены возможные решения для повышения точности.

Kalit so‘zlar: *Olmosh so‘z turkumi, statistik tahlil, POS tagging, morfologik analiz, avtomatik belgilash, korpus tilshunosligi*

KIRISH

Zamonaviy kompyuter lingvistikasi sohasida tabiiy tilni avtomatik qayta ishlash masalalari tobora dolzarb bo‘lib bormoqda. Xususan, so‘z turkumlarini avtomatik aniqlash – Part-of-Speech (POS) tagging texnologiyasi tilshunoslik tadqiqotlarida asosiy vosita sifatida xizmat qilmoqda. Ko‘plab rivojlangan tillarda (ingliz, rus, fransuz va boshqalar) bu yo‘nalishda mustahkam korpuslar, model va

teglovchilar mavjud bo‘lsa-da, o‘zbek tilida POS tagging texnologiyasining to‘liq shakllanmaganligi ilmiy-amaliy izlanishlar zaruratini yuzaga keltirmoqda[1].

Hozirgi kunda tabiiy tilni qayta ishlash (NLP) sohasida matndagi so‘zlarning kategoriyalarini aniqlash muammosi dolzarb hisoblanadi. So‘zlarning ot, olmosh, fe’l, ravish kabi kategoriyalarga mansubligini aniqlash POS (Parts of Speech) tagging deb ataladi. Bu jarayon matnni tahlil qilishda va mashinali o‘qitishda muhim ahamiyatga ega[2].

O‘zbek tilidagi olmoshlar semantik va sintaktik jihatdan murakkab birliklar hisoblanadi. Ularda kontekstga bog‘liq holda o‘zgaruvchanlik mavjudligi tufayli, avtomatik tizimlar tomonidan noto‘g‘ri belgilanish hollari ko‘p uchraydi. Xususan, shaxs olmoshlari (*“men”, “sen”, “u”*), ko‘rsatish olmoshlari (*“bu”, “shu”, “o‘sha”*) orasidagi formal yaqinlik, yoki so‘roq (*“kim”, “nima”*) va gumon olmoshlari (*“kimdir”, “qandaydir”*) orasidagi semantik noaniqlik natijasida POS tagging tizimlarida xatoliklar kuzatiladi.

Olmoshlarning turlari – kishilik, ko‘rsatish, so‘roq, o‘zlik, belgilash, bo‘lishsizlik va gumon olmoshlaridir. Ularning har biri morfologik va semantik jihatdan farqlanadi. Ular o‘z funksional yuklamasiga ko‘ra gapda turli sintaktik rolni bajaradi. Masalan, kishilik olmoshlari ko‘pincha ega vazifasida ishlatilsa, ko‘rsatish olmoshlari aniqlik kirituvchi yoki ko‘rsatkich bo‘lib xizmat qiladi.

Bundan tashqari, olmoshlar tuzilishiga ko‘ra quyidagi ko‘rinishlarda uchraydi:

1. *Sodda olmoshlar:* sen, u, biz, bu, shu
2. *Qo‘shma olmoshlar:* har kim, har nima, hech kim
3. *Juft olmoshlar:* kim-kim, nima-nima
4. *Takroriy olmoshlar:* o‘sha-o‘sha, shu-shu

1-jadval. Olmoshlarning gapdagi grammatik roli

№	Olmosh turi	Namuna	Grammatik rol
	Kishilik	men, sen, u	Subyekt, obyekt
	Ko‘rsatish	bu, shu, o‘sha	Aniqlik kirituvchi
	So‘roq	kim, nima, qaysi	Savol bildiruvchi
	Gumon	kimdir, nimadir	Noaniqlik bildiruvchi
	O‘zlik	o‘zim, o‘zing	Refleksiv holat
	Belgilash	har kim, har nima	Universal
	Bo‘lishsizlik	hech narsa, hech kim	Inkor holat

Bu jadvaldagi olmoshlar POS teglovchi modellarni tuzishda har bir olmosh turini alohida yondashuv bilan ko‘rib chiqishni talab etadi. Shuningdek, olmoshlar sintaktik roliga qarab turlicha turlarga kiritilishi, ular asosida avtomatik teg qo‘yuvchi algoritmlarni sinovdan o‘tkazish zaruratini vujudga keltiradi.

Asosiy qism. Nutqning bir qismi (POS) – so‘zning grammatik funksiyasini tavsiflash usuli. POS teglari NLP ning yuqori darajadagi funksiyalarida qo‘llanilsa-da, ularni o‘z-o‘zidan tushunish muhim va ularni matn tahlilida foydali maqsadlarda ishlatish mumkin[3].

Olmosh so‘z turkumi ikki xil ko‘rinishga ega, ya’ni *sof olmosh tabiatli so‘zlar* hamda *olmosh vazifasida qo‘llaniluvchi so‘zlar*. Sof olmosh so‘zlar o‘zida olmoshning barcha jihatlarini aks ettiradi, uning tub mohiyati ham olmosh ma’nosini ifodalaydi, ularga misol sifatida men, sen, ular, kimdir, nimadir va hokazolarni keltirish mumkin. Olmosh vazifasida qo‘llaniluvchi istilohlar esa tabiatan olmosh xarakteriga ega bo‘lsa-da, biroq ularning kelib chiqishi aslan boshqa so‘z turkumlariga ya’ni ot, sifat so‘z turkumlariga oid bo‘ladi. Nomlari qayd etilgan har ikki toifa bir-biridan ba’zi jihatlariga ko‘ra farqlanadi hamda ularning ayrim o‘xshash jihatlari ham mavjud[4]. Uni quyidagi 2-jadval asosida ko‘rib chiqamiz.

2-jadval.Olmosh so‘z turkumining qo‘llanish o‘rinlari

№	Sof olmosh tabiatli so‘zlar	Olmosh vazifasida keluvchi so‘zlar
	Olmoshning barcha xususiyatlarini o‘zida aks ettiradi	Olmoshning barcha xususiyatlarini o‘zida aks ettiradi
	Kelib chiqishiga ko‘ra ham olmosh turkumiga mansub bo‘ladi	Kelib chiqishigako‘ra boshqa so‘z turkumlariga oid bo‘lishi mumkin
		Ayrim hollarda asl so‘z turkumi bajargan vazifasini bajaradi (masalan, sifat so‘z turkumiga oid bo‘lsa, sifat so‘z turkumining xususiyatini aks ettirishi mumkin)

O‘zbek tilida olmoshlarning POS teglashdagi asosiy muammolaridan biri ularning ko‘p funksiyali va kontekstsiz ajratishning murakkabligidir. Masalan, “u” so‘zi kishilik olmoshi sifatida ham (*“U maktabga bordi”*), ko‘rsatish olmoshi sifatida ham (*“U bola sinfda qolibdi”*) ishlatilishi mumkin. Bu holat avtomatik teglashda ikki xil tegni talqin qilishga olib keladi. Shuningdek, so‘roq va gumon olmoshlari ham ko‘p hollarda bir-biriga yaqin strukturada bo‘lib, ayniqsa matn korpuslarida notekis ishlatilgan. Masalan, *“kim”* (so‘roq) va *“kimdir”* (gumon) olmoshlari korpuslarda ko‘pincha yagona teg bilan belgilangan, bu esa statistik modellar uchun chalg‘ituvchi hisoblanadi[5].

Olmoshlar tuzilmasiga ko‘ra, ularni sodda, qo‘shma, juft va takroriy turlarga ajratish mumkin. Sodda olmoshlar bir morfemali bo‘lsa, qo‘shma olmoshlar ikki yoki undan ortiq so‘zning birikmasidan tashkil topadi *(“har kim”, “biror narsa”)*. Juft olmoshlar esa ikki marta takrorlangan yoki o‘zaro juftlashgan olmoshlardir (*“kim-kim”, “nima-nima”*). Takroriy olmoshlar esa ayni bir olmoshning ikki marta takrorlanishi orqali hosil bo‘ladi (*“shu-shu”, “o‘sha-o‘sha”*). Bu strukturalarning

ko‘pligi va kontekstdagi rolini hisobga olgan holda, POS teglashda morfologik analizatorlarning aniqligini oshirish uchun chuqur semantik va sintaktik tahlil zarur. Masalan, mashinali o‘qitish asosidagi modellar uchun olmoshlarni kontekstual

Misollar bilan boyitish, yoki BERT modeli asosida maxsus mexanizmni joriy qilish taklif etiladi. Yuqoridagi misollar shuni ko‘rsatadiki, POS teglashda nafaqat sirtqi ko‘rinish, balki olmoshlarning gapdagi vazifasi, atrofdagi birliklar bilan munosabati ham muhimdir. Bu esa olmoshlarning alohida korpuslarda semantik belgilanishi, qo‘shimcha atributlar bilan boyitilishi, teglashda rule-based va statistik yondashuvlarning kombinatsiyasini talab qiladi[6].

Olmoshlar bilan bog‘liq POS teglashdagi muammolarni quyidagicha tahlil qilishimiz mumkin:

- Olmoshlarning mazmuni va ularning grammatik funksiyalari kontekstga bog‘liq bo‘ladi. Masalan, “u” so‘zi “u” olmoshi bo‘lishi mumkin, ammo u shaxsni bildirgan holda ishlatilganida “u”ning o‘zi haqida gapirsa, boshqacha xususiyatga ega bo‘lishi mumkin. Bu kabi muammolarni to‘g‘ri tahlil qilishda POS tagging tizimlari uchun qiyinchilik tug‘diradi.
- Kishilik va o‘zlik olmoshlarini ajratish: Ba’zi olmoshlar bir nechta shartli vaziyatlarga ega bo‘lishi mumkin. Misol uchun, “men” olmoshi shaxsni bildiradi, ammo “o‘zim” olmoshi o‘zlikni bildiradi. Bu ikki turdagi olmoshlarni to‘g‘ri ajratib olish tizim uchun qiyin bo‘lishi mumkin.
- Ko‘rsatish olmoshlarining aniq tahlili: “Bu” va “shu” kabi ko‘rsatish olmoshlari, odatda, predmetni yoki shaxsni aniq belgilash uchun ishlatiladi, ammo kontekstga qarab ular farqlanishi mumkin. Bu holatni aniqlash uchun kontekstni tahlil qilish talab etiladi.
- Bo‘lishsizlik olmoshlari: Bo‘lishsizlik olmoshlari (“hech kim”, “hech narsa”) juda ko‘p holatlarda, ayniqsa so‘roq gaplarida yoki inkor javoblarda ishlatiladi. Ularni to‘g‘ri tegishli kategoriyaga o‘tkazish uchun muayyan algoritmlar ishlab chiqilishi zarur.

3-jadval Olmoshlar bilan bog‘liq POS teglashdagi muammolar

№	Muammo turi	Tavsif	Misollar
	Morfologik	Olmoshlarning boshqa so‘z turkumlari bilan shaklan o‘xshashligi, qo‘shimchalarga ega bo‘lishi	“Uning kitobi”, “O‘zini yo‘qotdi”
	Kontekstual	Gapdagi o‘rni va vazifasiga qarab nolmoshlarning noto‘g‘ri turkumlanishi.	“Bu yaxshi fikr”, “Bu esa boshqa masala”
	Leksik-semantik	Ma’no va referentga bog‘liq holda olmoshlarning chalkash teglanishi	“Hech kim kelmadi”, “Kim bo‘ldi u?”

POS tagging tizimlarida olmoshlarni tahlil qilishda quyidagicha yechimlarni berishimiz mumkin:

1. *Statistik va ma’lumotlarga asoslangan modellar:* HMM (Hidden Markov Model) va boshqa statistik modellar, masalan, Viterbi algoritmi yordamida olmoshlarni tahlil qilishda kontekstga asoslanib to‘g‘ri POS tegini aniqlash mumkin. Bu modellarda har bir so‘zning ehtimoliy pozitsiyasini aniqlash orqali olmoshlarning to‘g‘ri kategoriyasini belgilash mumkin[7].

2. *Deep Learning metodlari:* O‘qitilgan neyron tarmoqlari yordamida olmoshlar turini aniqlashda, masalan, LSTM (Long Short-Term Memory) yoki BERT (Bidirectional Encoder Representations from Transformers) modellarida foydalanish mumkin. Ushbu metodlar kontekstni chuqur o‘rganib, olmoshlarning to‘g‘ri tahlilini amalga oshiradi[8].

3. *Qayta ishlash va morfologik tahlil:* Olmoshlarni to‘g‘ri ajratish uchun morfologik tahlil qilish va qayta ishlash texnikalari ishlatiladi. Bu jarayonlarda so‘zning tarkibiy qismlari, affikslar va sintaktik o‘zgarishlar tahlil qilinadi.

XULOSA.

Olmosh so‘z turkumi o‘zbek tilining murakkab va ko‘p qatlamli leksik-morfologik qatlamlaridan biri bo‘lib, ularni avtomatlashtirilgan tizimlar orqali aniqlash va teg belgilash (POS tagging) amaliyotida bir qator muammolar yuzaga keladi. Ushbu maqolada biz aynan morfologik, kontekstual va leksik-semantik muammolarni tahlil qildik va ularning har biri POS tagging tizimlariga qanday ta’sir ko‘rsatishini ko‘rsatib berdik. Morfologik muammolar asosan olmoshlarning boshqa so‘z turkumlari bilan shaklan o‘xshash bo‘lishi va morfemik jihatdan bir nechta variantlarga ega bo‘lishi natijasida yuzaga keladi. Kontekstual muammolar esa, ayniqsa ko‘rsatish, so‘roq va belgilash olmoshlariga xos bo‘lib, bir xil ko‘rinishdagi birliklarning har xil kontekstlarda turli sintaktik vazifalarni bajarishi bilan bog‘liq. Leksik-semantik muammolar esa referentga bog‘liq holda olmoshning asl ma’nosini to‘g‘ri aniqlashdagi murakkabliklar bilan bog‘liq bo‘lib, ular faqat sintaktik va semantik bog‘liqlik orqali hal qilinishi mumkin. Zamonaviy NLP (Natural Language Processing) texnologiyalarining rivojlanishi ushbu muammolarni hal qilish imkonini bermoqda. Xususan, gibrid yondashuvlar – ya’ni morfologik tahlil, kontekstual model (masalan, HMM, CRF, BERT asosidagi modellari) va semantik tahlil birgalikda qo‘llanilganda olmoshlarni to‘g‘ri teglash darajasi sezilarli darajada oshadi. Bundan tashqari, o‘zbek tiliga moslashtirilgan yorliqlash tizimlari, annotatsiyalangan korpuslar, va qo‘llanmalarning mavjudligi ilmiy-amaliy izlanishlar uchun mustahkam zamin yaratmoqda. Shu bois, olmoshlarning POS teglashdagi aniqligi tilshunoslik va kompyuter lingvistikasi uchun nihoyatda muhim bo‘lib, bu boradagi tadqiqotlar yanada chuqurlashtirilmog‘i zarur. O‘zbek tilining turkiy tillar orasida o‘ziga xos morfologik xususiyatlarga ega ekanini inobatga olgan holda, mahalliy tilshunos va dasturchilar hamkorligida ilg‘or va aniq ishlovchi tizimlar ishlab chiqilishi dolzarb masala sanaladi.

Foydalanilgan adabiyotlar:

1. Marie Novotná, Markéta Masopustová. Using Syntax Analyser SET as a GrammarChecker for Czech.2018.9.
2. Elov B., Hamroyeva Sh., Xusainova Z., Xudayberganov N., Yodgorov U., Yuldashev A. POS tagging of uzbek texts using Hidden Markov Models (HMM) and Viterbi algorithm.2023.105-bet.
3. Toshpo‘latov Z. Hozirgi o’zbek adabiy tili. –T.: Fan (2017.45-bet)
4. Brayan Roepke. A Quick Guide to Part of Speech Tagging. 2021.14.
5. Ne’matov H., Rasulov R..O‘zbek tili system leksikologiyasi asoslari.-T.:Fan, 1975.36-bet
6. Mirtojiyev M. O‘zbek tilida polisemiya.-T.:Fan,1975.48.
7. Jurafsky and Martin. Speech and Language Processing, 3rd edition, Draft chapters (Stanford NLP)2018.56.
8. Karel Pala. Comments onCzechMorphological Tagset (2018.67-69)
9. https://part of speech and taggings, natural language processing
10. Rada Mihalcea, Raymond Mooney, Dan Jurafsky. Word Classes and Part of Speech Tagging. (2018.10-12)

IV SHO‘BA. MASHINA TARJIMASI

GRAMMATIK TEGLASHDA MASHINALI O‘QITISH YONDASHUVI

Hamroyeva Shahlo Mirdjanovna,
Filologiya fanlari doktori, professor v.b.
shaxlo.xamrayeva@navoiy-uni.uz
ToshDO‘TAU

Erkinjonova Go‘zal Baxtiyor qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
uzbekistan-24@mail.ru
ToshDO‘TAU

Annotatsiya. Grammatik teglash tabiiy tilni qayta ishlashning (NLP) muhim sohasi hisoblanadi. So‘zlarning gapdagi sintaktik va morfologik rolini aniqlash uchun ishlatiladi. Mashinali o‘qitish yondashuvlari grammatik teglashda an’anaviy qoidalarga asoslangan usullarga nisbatan sezilarli yaxshilanishlarni ta’minlamoqda. Ushbu maqolada mashinali o‘qitish asosida grammatik teglash usullari, ularning ishlash tamoyillari va o‘zbek tilida qo‘llanilishi tahlil qilinadi.

Annotation: Grammar tagging is an important field in natural language processing (NLP). It is used to determine the syntactic and morphological roles of words in a sentence. Machine learning approaches provide significant improvements over traditional rule-based methods in grammar tagging. This article analyzes machine learning-based grammar tagging methods, their working principles, and their application in the Uzbek language.

Аннотация: Грамматическая разметка является важной областью обработки естественного языка (NLP). Она используется для определения синтаксических и морфологических ролей слов в предложении. Подходы машинного обучения обеспечивают значительные улучшения по сравнению с традиционными методами, основанными на правилах, в грамматической разметке. В этой статье анализируются методы грамматической разметки на основе машинного обучения, их принципы работы и применение в узбекском языке.

Kalit so‘zlar: *Mashinali o‘qitish, chuqur o‘rganish, grammatik teglash, qoidalarga asoslangan yondashuv, NLP, o‘zbek tili, POS tagging.*

KIRISH

Mashinali o‘qitish (Machine Learning, ML) – bu sun’iy intellektning (SI) bir bo‘limi hisoblanadi. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) va til modellari uchun mashinali o‘qitish juda muhim, chunki bu texnologiya tilni tushunish, tahlil qilish va avtomatik tarzda ishlov berish uchun juda zarur bo‘ladi. Mashinali o‘qitish tilni avtomatik tarzda tushunish va qayta ishlashni

ta’minlaydigan turli usullarni rivojlantirishga yordam beradi: Mashinali o‘qitishning teglashdagi o‘rni shundaki, model o‘rgatish jarayonida ma’lum bir til va uning tarkibiy qismlarini taniy oladi. Bu jarayon odatda katta hajmdagi annotatsiyalangan ma’lumotlar bilan amalga oshiriladi. Mashinali o‘qitish metodlari, xususan, supervized o‘qitish (supervised learning), bu turdagi tahlillarni avtomatlashtirish uchun ishlatiladi.

Jahon kompyuter lingvistikasida grammatik teglash bo‘yicha ko‘plab tadqiqotlar olib borilgan. Masalan, Penn Treebank korpusi asosida yaratilgan ingliz tili uchun taggerlar, hamda Universal Dependencies loyihasi doirasida turli tillar bo‘yicha teglash tizimlari ishlab chiqilgan. So‘nggi yillarda BERT, GPT va XLM-R kabi transformer modellaridan foydalangan holda yuqori aniqlikdagi grammatik teglash tizimlari ishlab chiqilmoqda, Bojanowski va boshqalar (2017)[1] tomonidan ishlab chiqilgan fastText modeli turli tillardagi grammatik teglashga qo‘llanilishi mumkinligini ko‘rsatgan.

O‘zbek tilida grammatik teglash bo‘yicha tadqiqotlar so‘nggi yillarda boshlangan bo ‘lishiga qaramay, izlanishlar olib borilmoqda. O‘zbek tilining aglutinativ tabiati tufayli, klassik HMM va qoidalarga asoslangan yondashuvlar samaradorligi cheklangan. Shuningdek, chuqur o‘rganish modellaridan foydalanish bo‘yicha ham ayrim eksperimental ishlanmalar mavjud. Xususan, morfologik tahlil va sintaktik teglash bo‘yicha ilg‘or tadqiqotlar olib borilmoqda. Xususan Sh. Hamrayeva “O‘zbek tili morfologik analizatorining lingvistik ta’minoti”[2] dissertatsiyasida grammatik teglash haqida batafsil ma’lumot bergan.

Mashinali o‘qitish usullari grammatik teglashni avtomatlashtirishda muhim hisoblanadi. Turli korpuslar matnni tahlil qilish, so‘z turkumlarini teglash va sintaktik tahlilnii aniqlash imkonini beradi. Ayniqsa, nazoratli o‘rganish va chuqur o‘rganish usullari bu sohada samarali natijalar bermoqda. Grammatik teglashda statistik modellar, neyron tarmoqlar va qoidalarga asoslangan tizimlar ishlatiladi.

Nazoratli o‘rganish (supervised learning) modellari tegli korpuslardan foydalanadi. Ular oldindan belgilangan so‘z turkumlari asosida o‘qitiladi va yangi matnlarda teg qo‘yish uchun ishlatiladi. Bunday modellar orasida Naive Bayes, SVM, CRF va LSTM asosida qurilgan modellar ajralib turadi.

Nazoratsiz o‘rganish(unsupervised learning) modellari esa avvaldan teglanmagan matnlar asosida o‘z-o‘zidan tuzilmalarni aniqlashga harakat qiladi. Bu yondashuvning asosiy kamchiligi u nazoratli modellar darajasida yuqori aniqlikni ta’minlay olmaydi. Ammo, bu modellar yangi til resurslari uchun moslashuvchan yechim bo‘lishi mumkin. Chuqur o‘rganish(supervised learning) grammatik teglashda katta yutuqlarga erishgan usullardan biridir. Ayniqsa, RNN, LSTM va Transformer modellarining ishlatilishi natijasida yuqori aniqlikga erishish mumkin bo‘ldi. Ushbu modellar kontekstni hisobga olish orqali murakkab tuzilmalarda ham

yaxshi natijalar ko‘rsatadi. Chuqur o‘rganish modellarining samaradorligi katta hajmdagi ma’lumotlar to‘plamidan foydalanganda oshadi.

Qoidalarga asoslangan yondashuv sintaktik va morfologik qoidalarga asoslanadi. Ushbu tizimlar lingvist mutaxassislar tomonidan ishlab chiqilgan qoidalarga tayangan holda teglash jarayonini amalga oshiradi. Grammatik qoidalarga asoslangan teglash tizimlari an’anaviy yondashuv bo‘lib, turli tillarda ishlatiladi. Ular tilda mavjud grammatik xususiyatlarga asoslanib, so‘zlarning rolini aniqlaydi. Masalan, o‘zbek tilida ot, fe’l, sifat singari so‘z turkumlarini aniq teglash uchun morfologik xususiyatlarga tayanish talab etiladi.

Neyron tarmoqlar sathma-sath o‘rganish printsipiga asoslanadi. Har bir qatlam oldingi qatlamdan olingan ma’lumotlarni qayta ishlaydi va chuqurroq xususiyatlarni ajratib oladi. Ayniqsa, konvolyutsion neyron tarmoqlar (CNN) va rekurrent neyron tarmoqlar (RNN) tabiiy tilni qayta ishlash jarayonlarida samarali natijalar beradi. Transformer arxitekturasi esa uzoq masofali bog‘liqliklarni o‘rganish va tahlil qilishda yuqori samaradorlikni ko‘rsatadi.

O‘zbek tilidagi grammatik teglash jarayonida Deep Learning modellarining natijalari yuqori bo‘lishi mumkin. Bu tilning morfologik xususiyatlari, jumladan, affiksatsiya tizimi va sintaktik strukturalari tahlil qilinib, teg qo‘yish jarayoni amalga oshiriladi.

Grammatik teglashning keyingi rivojlanish bosqichlari, ayniqsa, Deep Learning texnologiyalaridan foydalanish bilan bog‘liq bo‘ladi. Bugungi kunda mashinali o‘qitish modellarining samaradorligi faqatgina turli tillarga tegishli murakkab korpuslar orqali yuqori aniqlikni ta’minlash bilan cheklanmaydi, balki o‘zbek tilidagi kabi morfologik jihatdan boy tillarda ham yuqori natijalarga erishishga imkon yaratadi. Shu bilan birga, yangi modellar va usullar, masalan, *Transformer* va *BERT* kabi avvalgi metodlardan yanada samaraliroq natijalarga olib kelmoqda.

Transformer arxitekturasi grammatik teglashda yuqori samaradorlikni ko‘rsatgan eng ilg‘or texnologiyalardan biridir. Ushbu modelda so‘zlarning ma’nosi va ularning o‘zaro bog‘lanishlari davomiy va murakkab kontekstda aniqlanadi. Transformer modelining asosiy afzalliklaridan biri bu uning *uzoq masofali bog‘liqliklarni* o‘rganish qobiliyatidir. Natijada, grammatik strukturalarni tahlil qilishda samaradorlik oshadi. BERT (Bidirectional Encoder Representations from Transformers) modelining alohida ahamiyati shundaki, u so‘zlarni ikki yo‘nalishda, ya’ni oldingi va keyingi kontekst asosida o‘rganadi. Bu orqali so‘zlar orasidagi bog‘lanishni yanada aniqroq tahlil qilish imkonini beradi. BERT asosidagi modellar o‘zbek tilida grammatik teglashni yanada aniqlik bilan amalga oshirishi mumkin. BERT ing pre-trening qobiliyatlari yordamida model tilning barcha sintaktik va morfologik xususiyatlarini o‘rganadi va yangi matnlar uchun moslashadi. Ayniqsa,

o‘zbek tilining aglutinativ tabiati tufayli, so‘zlarning turli affiksatsiya shakllari va ularning o‘zgarishlarini tahlil qilishda BERT modellaridan foydalanish samarali bo‘ladi.

Deep Learning ,xususan, *Recurrent Neural Networks (RNN), Long Short-Term Memory (LSTM)* va *Convolutional Neural Networks (CNN),* grammatik teglashda yana bir qadam oldinga siljish imkonini beradi. Bu modellar, ayniqsa, tilning sintaktik va morfologik strukturalarini chuqurroq o‘rganish imkonini beradi, chunki ular uzun muddatli bog‘liqliklarni aniqlash va yaxshiroq kontekstni tushunish uchun yaratilgan. *LSTM* va *RNN* kabi modellar so‘zlar o‘rtasidagi uzluksiz bog‘liqliklarni ta’minlashda juda foydali, chunki ular vaqt ketma-ketligini tahlil qilishda kuchli hisoblanadi. Bu modellar, ayniqsa, sintaktik bog‘lanishlarni aniqlashda samarali ishlaydi. O‘zbek tilidagi so‘zlar o‘rtasidagi morfologik o‘zgarishlarni, affikslar va ularning o‘zaro bog‘lanishini aniq tasavvur qilishga imkon yaratadi. *CNN* modellarining esa so‘zlarning lokal xususiyatlarini aniqlashda samarali ekanligi ma’lum. Bu modellarning yordamida so‘zlarning strukturaviy jihatlarini va ularning turkumlarini tez va samarali tarzda aniqlash mumkin.

Grammatik teglash uchun mashinali o‘qitish metodlari korpuslarning sifatiga va hajmiga katta bog‘liq. *Korpuslar* – bu tilni tahlil qilishda asosiy resursdir, chunki ularning yordamida modellarga tilning sintaktik va morfologik strukturalari o‘rgatiladi. O‘zbek tilini grammatik teglashini avtomatlashtirishda, tilning barcha affiksatsiya tizimlarini, so‘z turkumlarini va ularning o‘zaro bog‘lanishlarini aniqlash uchun yirik va diversifikatsiyalangan annotatsiyalangan korpuslar kerak bo‘ladi.

Fe’l so‘z turkumini zamon va shaxs-son qo‘shimchalarini aniqlash uchun 15 mingta matnni fe’l so‘z turkumi qo‘lda grammatik teglandi, bu esa zamon va shaxs-son kategoriyalarning avtomatik aniqlash uchun asos yaratildi. Olingan natijalar mashina tarjimasi, matnni avtomatik tahlil qilish va tabiiy tilni qayta ishlash (NLP) tizimlarida qo‘llanilishi mumkin. Ayniqsa, fe’l shakllarining zamon, shaxs-son va nominal shakllardagi taqsimoti avtomatik tarjima algoritmlarida kontekstga mos tarjima variantlarini tanlashda muhim rol o‘ynaydi. Keyingi tadqiqotlarda ushbu teglangan matnlar asosida matnlarda ularning chastotasi, bu natijalardan qanday model tuzish mumkinligi ko‘rib chiqiladi.

Xulosa

Grammatik teglash tabiiy tilni qayta ishlash (NLP) sohasida muhim yo‘nalish bo‘lib, mashinali o‘qitish usullari bu jarayonni samarali va avtomatik tarzda amalga oshirishga yordam bermoqda. Deep learning va supervised learning usullari grammatik teglashda katta yutuqlarga erishish imkonini yaratdi. Transformer modellarining va boshqa chuqur o‘rganish texnologiyalarining qo‘llanilishi

grammatik teglashning samaradorligini oshirishda muhim rol o‘ynaydi. O‘zbek tilida grammatik teglashni yaxshilash uchun zamonaviy modellarni qo‘llash, tilning morfologik xususiyatlarini chuqur tahlil qilish va yirik annotatsiyalangan korpuslardan foydalanish zarur. Kelajakda, bu sohadagi tadqiqotlar davom etib, yanada aniqroq va samarali grammatik teglash tizimlarini ishlab chiqish imkonini beradi.

Foydalanilgan adabiyotlar:

1. **Bojanowski, P., Grave, E., Mikolov, T., & Joulin, A.** (2017). Enriching Word Vectors with Subword Information. arXiv preprint arXiv:1607.04606.
2. Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таьминоти: Филол. фан. док. ...дис. – Тошкент, 2021. – Б. 150.
3. **Collobert, R., & Weston, J.** (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning *(ICML '08). ACM.*
4. **Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. A., Kaiser, Ł., & Polosukhin, I.** (2017). Attention is all you need. In Advances in neural information processing systems (NeurIPS 2017).
5. **Devlin, J., Chang, M. W., Lee, K., & Toutanova, K.** (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
6. Elova D. Fe’l grammatik shakllarining uslubiy xoslanishiga ko‘ra lingvistik annotatsiyalash masalasi xususida.
7. Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Бухоро, 2018. – Б. 3.
8. **Abdukarimov, S., & Tashkent, A.** (2017). O‘zbek tilida morfologik tahlil va sintaktik teglash. O‘zbekiston tilshunoslik va informatika jurnali, 5(1), 120-134.
9. **Shodmonov, M.** (2020). O‘zbek tili va uning grammatik tizimi: Mashinali o‘qitish va tillarni qayta ishlash. Toshkent, O‘zbekiston.
10. **O‘rinov, S., & Saidov, T.** *(2021).* Chuqur o‘rganish va grammatik teglash: O‘zbek tilidagi yondashuvlar. *Informatika va texnologiyalar jurnali, 7(3), 202-217.*
11. https://www.researchgate.net/publication/380823282_O'ZBEK_TILI_KORPUSINI_SINTAKTIK_TEGLASH_MASALASI
12. https://www.researchgate.net/publication/375912394_Korpus_birliklarini_teglash_va_annotatsiyalash

PARALLEL KORPUSLARNI TARJIMA XOTIRASIGA INTEGRATSIYA QILISH TAMOYILLARI

Shamsiyeva Gulshoda Asliddin qizi,
O‘zMU tadqiqotchisi
shamsiyeva_g@nuu.uz
O‘zMU

Annotatsiya. Mashina tarjimasi sohasida parallel korpuslardan foydalanish an’anaviy tarjima jarayonini sezilarli darajada osonlashtiradi. Tarjima xotirasi (Translation Memory, TM) tizimlari tarjima jarayonini avtomatlashtirish va samaradorligini oshirish uchun muhim vosita hisoblanadi. Ushbu maqolada parallel korpuslarni tarjima xotirasiga integratsiya qilish tamoyillari, metodologiyasi va samaradorligi tahlil qilinadi.

Abstract. The utilization of parallel corpora in the field of machine translation significantly streamlines the traditional translation process. Translation Memory (TM) systems are considered crucial tools for automating and enhancing the efficiency of the translation process. This article analyzes the principles, methodology, and effectiveness of integrating parallel corpora into translation memory.

Аннотация. Использование параллельных корпусов в области машинного перевода значительно облегчает традиционный процесс перевода. Системы памяти переводов (Translation Memory, TM) являются важным инструментом для автоматизации и повышения эффективности процесса перевода. В данной статье анализируются принципы, методология и эффективность интеграции параллельных корпусов в память переводов.

Kalit so‘zlar: *tarjima xotirasi, parallel korpus, filtrlash, moslashtirish, indeksatsiya qilish.*

Tarjima xotirasi – bu tarjima jarayonida avvalgi tarjimalarni saqlab qoladigan va takroriy matnlarni avtomatik tarjima qilish imkonini beradigan dasturiy ta’minotdir [1]. Parallel korpus esa turli tillardagi tarjima qilingan matnlarning muvofiqlashtirilgan bazasi bo‘lib, mashina tarjimasi uchun asosiy resurslardan biridir [2]. Zamonaviy tarjima jarayonida ushbu ikki element birgalikda ishlatilishi natijasida tarjima sifati va tezligi sezilarli darajada oshadi.

Bir qancha olimlar, xususan, H.L.Somers [3], A.Way [4], P.Koehn [5], B.Maegaard [6], D.Marcu, W.Wong [7], M.Federico [8], A.Toral [9] v.b. avtomatik tarzda parallel korpuslardan tarjima xotirasini yaratish haqida izlanishlar olib borgan

bo‘lib, ularning asarlarida tarjima xotiralari qanday ishlashi va parallel korpuslardan qanday foydalanilish haqida chuqur tahlillar berilgan, shuningdek, ularda parallel korpus va tarjima xotirasining moslashuvi, parallel korpuslardan foydalanib tarjima modellari yaratish hamda tarjima xotirasini statistika asosidagi mashina tarjimasi bilan integratsiya qilish usullari haqida so‘z boradi.

Tarjima xotiralari zamonaviy tarjimonlar va tarjima agentliklari uchun katta ahamiyatga ega bo‘lib, ular takroriy iboralarni aniqlash, aniq va izchil tarjimalarni ta’minlash uchun foydalaniladi. Shu bilan birga, parallel korpuslar sun’iy intellektga asoslangan tarjima tizimlari uchun asosiy ma’lumot bazasi sifatida xizmat qiladi.

Parallel korpuslarni tarjima xotirasiga integratsiya qilish quyidagi bosqichlarda amalga oshirildi:

1. Korpus yig‘ish va tozalash (filtrlash) – turli manbalardan tarjima matnlarini yig‘ish va strukturaga keltirish (maslan, KazParC). Korpus sifati tarjima jarayonining asosiy elementlaridan biri bo‘lib, ularning sifatini nazorat qilish, jumladan, noaniqliklarni kamaytirish, noto‘g‘ri moslangan segmentlarni chiqarib tashlash va umumiy sifatsiz ma’lumotlarni filtrlash uchun maxsus tozalash algoritmlaridan foydalaniladi.

1.1. Lingvistik tozalash (filtrlash) (Linguistic Filtering). Bunday tozalash usulida “Google Compact Language Detector (CLD3)” yoki “langid.py” kabi vositalardan foydalanilib, ularning vazifasi parallel korpus ichida noto‘g‘ri tillarda kiritilgan matnlarni chiqarib tashlash (tilni aniqlash – Language Identification), belgilar ketma-ketligi, masalan, @@@@@ , XXXXX yoki noto‘g‘ri formatlangan matnlarni olib tashlash (shovqinli ma’lumotlarni filtrlash – Noise Removal) hamda bitta til segmentining tarjima qilingan varianti bilan sintaktik va semantik bog‘lanmaganligi aniqlansa, ushbu segmentni chiqarib tashlash (mantiqiy bog‘lanmagan tarjimalarni aniqlash – Mismatched Translations) dan iborat [10].

1.2. Masofaga asoslangan tozalash (Distance-Based Filtering). Ushbu metod yordamida parallel segmentlarning o‘lchami va muvofiqligi tekshiriladi. Bunda asl va tarjima qilingan jumla orasidagi o‘zgarishlar sonini hisoblab, juda katta farqlarni filtrlash uchun “Levenshtein masofasi (Levenshtein Distance)”, tarjima segmentlarining vektorli ifodalariga asoslangan o‘zaro o‘xshashlikni tekshirish uchun “Kosinus o‘xshashligi (Cosine Similarity)” algoritmlaridan foydalaniladi [5,11].

1.3. Statistika asosida tozalash (Statistical Filtering) usulida tarjima sifati statistik usullar orqali baholanadi. Masalan,

• BLEU Score Filtering – parallel segmentlarning BLEU skori hisoblanadi, agar u juda past bo‘lsa (masalan, 0.1 dan past bo‘lsa), segment olib tashlanadi.

• Language Model Perplexity (Til modelining chalkashligi) – agar tarjima korpusining perplexity ko‘rsatkichi juda baland bo‘lsa, ya’ni, tarjima natijalarining kutilmagan va noto‘g‘ri bo‘lish ehtimoli katta bo‘lsa, segment chiqarib tashlanadi.

• TF-IDF Filtering usuli kam ishlatiladigan yoki juda tez-tez uchraydigan so‘zlar bo‘yicha segmentlarni filtrlash uchun ishlatiladi [12].

1.4. Neyron tarmoqlar asosida tozalash (Neural Filtering) da chuqur o‘rganish (deep learning) asosida noto‘g‘ri parallel segmentlarni avtomatik filtrlash amalga oshiriladi [13]. Parallel korpus sifati tarjima jarayonida hal qiluvchi ahamiyatga ega. Korpus sifatini nazorat qilish uchun tilshunoslik, statistika va neyron tarmoqlarga asoslangan tozalash algoritmlari qo‘llaniladi. Bu algoritmlar orqali noto‘g‘ri tarjimalar, shovqinli ma’lumotlar va semantik jihatdan mos kelmaydigan segmentlar filtrlab tashlanadi.

2. Parallel korpuslarni tarjima xotirasiga integratsiya qilishning keyingi bosqichi moslashtirish va indekslashtirishdir. Parallel korpuslarda tarjima sifatini oshirish va tarjima xotiralarining samaradorligini ta’minlash uchun bu jarayonlar muhim ahamiyat kasb etadi. Parallel matn segmentlarini bog‘lash va ularni tezkor qidirish imkoniyatini ta’minlash orqali tarjima tizimlarining natijalari sezilarli darajada yaxshilanadi. Moslashtirishning asosiy maqsadi har bir tarjima segmentini uning asl nusxasi bilan sintaktik, semantik va pragmatik jihatdan bog‘lashdan iborat. Buning uchun quyidagi metodlardan foydalaniladi:

2.1. To‘g‘ridan-to‘g‘ri moslashtirish (Alignment). Moslashtirishning bu turida manba tilidagi jumlalar va ularning tarjimalari bir-biriga mos kelishi lozim. Bunda quyidagi modellardan foydalaniladi:

• Qoidalarga asoslangan yondashuv (Rule-Based) – jumlalar uzunligi, tinish belgilari va so‘zlarning tarqalishiga asoslanadi [14].

• Statistik yondashuv – parallel korpusdagi so‘z va iboralar chastotasiga asoslanib segmentlarni bog‘laydi [15].

• Mashina o‘rganish metodlari – neyron tarmoqlar va transformer modellar yordamida yanada aniqroq moslashtirishni amalga oshiradi [16].

2.2. So‘z va iboralar darajasida moslashtirish. Ba’zan jumlalar to‘g‘ridan-to‘g‘ri mos kelmaydi, chunki tarjimada iboralar tuzilishi o‘zgarishi mumkin. Bu holatda dinamik dasturlash algoritmlari, xususan, Levenshtein masofasi va DICE koeffitsienti asosida so‘z moslashuvi tekshiriladi [17].

2.3. Indeksatsiya qilish. Parallel matnlarni samarali qidirish va tarjima xotirasidan tezkor foydalanish uchun indeksatsiya qilish (indexing) zarur. Indeksatsiya natijasida katta hajmdagi matnlar tez va samarali qidiriladi, bu esa tarjima tizimining tezligini oshiradi. Indeksatsiyaning eng samarali usullaridan biri – *teskari indeks* (Inverted Index) texnikasidir. Bu usulda har bir so‘zning qaysi hujjatlarda va qaysi segmentlarda uchrashi to‘g‘risidagi *qidiruv jadvali* yaratiladi [18]. Ushbu metod Google va boshqa qidiruv tizimlarida ham keng qo‘llaniladi.

2.4. TF-IDF va vektorli modellar. Indeksatsiyani yaxshilash uchun *TF-IDF (Term Frequency - Inverse Document Frequency)* va *vektorli ifodalash* metodlaridan foydalaniladi. TF-IDF – muhim terminlarni ajratish va parallel matndagi asosiy tarjima birliklarini belgilashga yordam beradi [19]. *Word2Vec*, *FastText* va *BERT* kabi *neyron modellar* qidiruv natijalarini yanada moslashuvchan qilishga yordam beradi [20].

3. Parallel korpuslarni tarjima xotirasiga integratsiya qilishning keyingi bosqichida tarjima jarayonida tarjima xotirasidan foydalaniladi. Tarjima jarayoni murakkab lingvistik va texnologik bosqichlarni o‘z ichiga oladi. Tarjima xotirasi va avtomatik tarjima tizimlari tarjima jarayonini avtomatlashtirish va samaradorligini oshirishga xizmat qiladi. Ushbu texnologiyalar tarjimonlarga vaqtni tejash, natijalarning barqarorligini oshirish va inson xatolarini kamaytirish imkonini beradi [21].

Tarjima xotirasining ishlash tamoyili quyidagi bosqichlar ketma-ketligida amalga oshiriladi:

⇒ Segmentatsiya – matn tarjima birliklariga bo‘linadi (masalan, jumlalar yoki iboralar).

⇒ Mos kelish tekshiruvi – yangi segmentlar oldin tarjima qilingan ma’lumotlar bilan taqqoslanadi.

⇒ Qayta ishlatish va takroriy tarjima – tizim mavjud tarjimalarni taklif qiladi.

⇒ Qo‘lda tahrirlash – tarjimon TM tomonidan taqdim etilgan natijani tekshiradi va tahrirlaydi.

L. Bowkerning ta’kidlashicha [1], TM tizimlari ayniqsa bir xil formatdagi va takroriy jumlalardan iborat hujjatlarda yuqori samaradorlikni ko‘rsatadi.
Zamonaviy avtomatik tarjima tizimlari TM bilan integratsiyalashgan holda ishlaydi. Ushbu tizimlarning evolyutsiyasi uch asosiy modelni o‘z ichiga oladi:

• Qoidalarga asoslangan tarjima (RBMT – Rule-Based Machine Translation) – lingvistik qoidalar va lug‘atlarga asoslangan tarjima texnologiyasi [22].
• Statistik tarjima (SMT – Statistical Machine Translation) – katta hajmdagi parallel korpuslardan foydalangan holda probabilistik usullar yordamida tarjima qiladi [23].
• Neyron tarjima tizimlari (NMT – Neural Machine Translation) – sun’iy neyron tarmoqlari yordamida matnning kontekstini chuqur o‘rganadi va yanada tabiiy tarjima hosil qiladi [24].

Neyron tarjima tizimlari (NMT) transformer modellarini qo‘llagan holda so‘zlar orasidagi bog‘liqlikni tushunadi va kontekstni hisobga olgan holda tarjima qiladi [16].

Tarjima tizimlarining samaradorligini oshirish uchun TM va MT integratsiyasi amalga oshirilmoqda. Bunda:

⇒ Tarjima xotirasidan mavjud tarjimalar ajratib olinadi.

⇒ NMT faqat yangi segmentlarni tarjima qiladi.

⇒ Natijalar TM bazasiga qayta yuklanadi va kelgusida foydalanish uchun saqlanadi [25].

Tadqiqotchilarning ta’kidlashicha, bu yondashuv tarjima jarayonining barqarorligini oshirish va inson aralashuvini minimallashtirish imkonini beradi [8]. Tarjima xotirasi tizimlari hozirda turli sohalarda, jumladan, davlat idoralari, xalqaro tashkilotlar va IT kompaniyalarda keng qo‘llanilmoqda [26].

Kelajakda sun’iy intellekt va chuqur o‘rganish texnologiyalarining rivojlanishi tarjima xotiralarining yanada takomillashishiga olib keladi [21].

Kelajak istiqbollari quyidagi yo‘nalishlarda rivojlanishi kutilmoqda:

⇒ Yuqori sifatli avtomatik tarjima tizimlari – chuqur o‘rganish va neyron tarmoqlari asosida ishlaydigan ilg‘or tizimlar tarjimalarni yanada tabiiy qilishga yordam beradi.

⇒ Ko‘p tilli tarjima xotiralari – mavjud tizimlar faqat ikki tilli parallel korpuslarga asoslangan bo‘lsa, kelajakda ko‘p tilli parallel korpuslar va tarjima xotiralari ommalashishi mumkin.

⇒ Interaktiv va adaptiv tarjima tizimlari – tarjimonlarning ishlash uslubiga moslashadigan va ularning tarjima uslubini o‘rganadigan tizimlar rivojlanishi kutilmoqda.

Tadqiqotlar shuni ko‘rsatadiki, tarjima xotirasi tizimlaridan foydalanish tarjimonlarning mahsuldorligini sezilarli darajada oshiradi. Misol uchun, 371 902 ta parallel jumlalarni o‘z ichiga oluvchi KazParC: Qozoq parallel korpusidan foydalangan holda davlat idoralari uchun maxsus ishlab chiqilgan tarjima tizimlarida tarjima jarayoni 30-40% tezlashgani qayd etilgan. Shuningdek, KazParC loyihasi doirasida Tilmash nomli neyron mashinaviy tarjima modeli ham ishlab chiqilgan bo‘lib, ushbu model BLEU va chrF kabi standart baholash metrikalariga ko‘ra, Google Translate va Yandex Translate kabi yirik kompaniyalar bilan tenglashadi yoki ba’zi hollarda ularni ortda qoldiradi. [https://github.com/IS2AI/KazParC].

Parallel korpusga asoslangan tarjima xotirasi tizimlari zamonaviy tarjima jarayonining muhim elementlaridan biri hisoblanadi. Ular tarjima jarayonini tezlashtirish, sifatini oshirish va iqtisodiy samaradorligini ta’minlash imkonini beradi. Ushbu maqolada ko‘rib chiqilgan metodologiya va yondashuvlar ushbu sohada tadqiqotlarni davom ettirish uchun muhim asos bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Bowker L. Computer-Aided Translation Technology: A Practical Introduction. – Ottawa: University of Ottawa Press, 2002. – 243 p.

2. Dagan I., Church K., Gale W. Robust Bilingual Word Alignment for Machine Aided Translation // ACL Anthology. – 1993.

3. Сомерс Х.Л. Translation Memory Systems // Hutchins J. (ed.) Computers and Translation: A Translator’s Guide. – Amsterdam: Benjamins, 2003. – С. 31–48.
4. Way A. Traditional and Emerging Use-Cases for Machine Translation // Machine Translation. – 2014. – Vol. 28, No. 3-4. – P. 243–254.
5. Koehn P. Europarl: A Parallel Corpus for Statistical Machine Translation // Proceedings of the MT Summit X. – Phuket, Thailand, 2005. – P. 79–86.
6. Maegaard B. Translation Memory: An Efficient Way to Reuse Translations // Proceedings of MT Summit VII. – Singapore, 1999. – P. 331–334.
7. Marcu D., Wong W. A Phrase-Based, Joint Probability Model for Statistical Machine Translation // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2002. – P. 133–139.
8. Federico M., Bertoldi N., Barbaiani M. Integrating Translation Memories into Phrase-Based Statistical Machine Translation // Proc. of ACL Conf. – 2014.
9. Toral A., Muñoz R. A Proposal to Automatically Build and Use Parallel Corpora within the OpenTrad Project // Proceedings of the Workshop on Building and Using Parallel Texts. – New York, USA, 2006. – P. 149–152.
10. Baldwin T., Lui M. Language Identification: The First Step Towards NLP // Computational Linguistics. – 2010.
11. Kocmi T., Bojar O. An Exploration of Sentence Embeddings for Neural Machine Translation. – 2017.
12. Papineni K., Roukos S., Ward T., Zhu W. BLEU: A Method for Automatic Evaluation of Machine Translation // ACL. – 2002.
13. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // NAACL. – 2018.
14. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. – 1993.
15. Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Lafferty J. D., Mercer R. L., Roossin P. S. The Mathematics of Statistical Machine Translation: Parameter Estimation // Computational Linguistics. – 1993. – Vol. 19, No. 2. – P. 263–311.
16. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is All You Need // Proc. of NeurIPS 2017. – Vol. 30. – P. 5998–6008.
17. Melamed I. D. Models of Translational Equivalence among Words // Computational Linguistics. – 2000.
18. Manning C. D., Raghavan P., Schütze H. Introduction to Information Retrieval. – Cambridge: Cambridge University Press, 2008.
19. Sparck Jones K. A Statistical Interpretation of Term Specificity and Its Application in Retrieval // Journal of Documentation. – 1972.
20. Mikolov T., Sutskever I., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // ICLR. – 2013.

21. Somers H. Translation Memory Systems: Enlightenment or Feudalism? // Machine Translation Journal. – 2003.

22. Hutchins W. J., Somers H. L. An Introduction to Machine Translation. – London: Academic Press, 1992. – 362 p.

23. Koehn P. Statistical Machine Translation. – Cambridge: Cambridge University Press, 2010. – 487 p.

24. Bahdanau D., Cho K., Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate // ICLR. – 2015.

25. Koehn P., Knowles R. Six Challenges for Neural Machine Translation // Proc. of the 1st Conf. on Machine Translation (WMT 2017). – P. 28–39.

26. Gough N., Way A. Example-Based Machine Translation Using a Memory-Based Learning Approach // Computational Linguistics. – 2004.

27. Abduraxmonova N. Kompyuter lingvistikasi: darslik. – Toshkent: Nodirabegim, 2021. – 312 b.

28. Abduraxmonova N. Mashina tarjimasining lingvistik ta’minoti: monografiya. – Toshkent, 2018. – 185 b.

29. Abdurakhmonova N., Hayrullayeva G., Urdishev K. Using a Corpus for Teaching Synonymy in Uzbek // Электронная письменность народов Российской Федерации: опыт, проблемы и перспективы: материалы II Междунар. науч. конф. – Уфа, 11–12 декабря 2019 г. – Уфа: Изд-во БашГУ, 2019.

30. Agostini A., Usmanov T., Khamdamov U., Abdurakhmonova N., Mamasaidov M. UZWORDNET: A Lexical-Semantic Database for the Uzbek Language // Proceedings of the 11th Global Wordnet Conference. – University of South Africa (UNISA), 2021. – P. 8–19.

31. Gatiatullin A., Suleymanov D., Prokopyev N., Abdurakhmonova N. Turkic Morpheme: From the Portal to the Linguistic Platform // Aliev R.A. et al. (Eds.) Lecture Notes in Networks and Systems. – Vol. 912. – Springer, Cham, 2024. – https://doi.org/10.1007/978-3-031-53488-1_22

32. KazParC. Kazakh Parallel Corpus for Machine Translation // arXiv preprint. – 2024.

33. Mengliev D. B., Abdurakhmonova N., Hayitbayeva D., Barakhnin V. B. Automating the Transition from Dialectal to Literary Forms in Uzbek Language Texts: An Algorithmic Perspective // 2023 IEEE XVI Int. Sci. and Tech. Conf. APEIE. – Novosibirsk, 2023. – P. 1440–1443. DOI: 10.1109/APEIE59731.2023.10347617.

34. Mengliev D., Barakhnin V., Abdurakhmonova N. Development of Intellectual Web System for Morph Analyzing of Uzbek Words // Applied Sciences. – 2021. – Vol. 11, No. 19. – Art. no. 9117. https://doi.org/10.3390/app11199117

V SHO‘BA. MATNGA QAYTA ISHLOV BERISH TIZIMLARI

JAHON TAJRIBASIDA AVTOMATIK MATN TAHRIRLASH VOSITALARI

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Sobirova Nazira G‘anijon qizi,
tayanch doktoranti
nazirasobirova@gmail.com
ToshDO‘TAU

Annotatsiya. Avtomatik matn tahriri dasturlari sun’iy intellekt (AI) va tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalanib, matnlarni avtomatik tarzda tahrirlash, grammatik xatolarni tuzatish va uslubiy jihatdan yaxshilash imkoniyatini beradi. Ushbu dasturlar nafaqat grammatika va imloni tekshirish, balki kontekstual takliflar berish, uslubiy optimallashtirish va yozish jarayonini avtomatlashtirish uchun ham qo‘llaniladi.Jahon tajribasida avtomatik matn tahriri dasturlari yozish jarayonini sezilarli darajada osonlashtirib, matn sifati va aniqligini oshirishga yordam beradi. Bu dasturlar har xil sohalarda – akademik yozuvlar, biznes hujjatlari, jurnalistika va kundalik yozishmalar uchun qo‘llaniladi. Sun’iy intellekt va tabiiy tilni qayta ishlash texnologiyalari rivojlanishi bilan ushbu dasturlar yanada mukammal va moslashuvchan bo‘lib borayotgani kuzatilmoqda. Mazkur ilmiy maqolada avtomatik matn tahrirlash vositalarining shakllanishi, rivojlanish bosqichlari va amaliyotda qo‘llanilishining jahon tajribasi asosida tizimli tahlili berilgan.

So‘nggi yillarda tabiiy tilni qayta ishlash (Natural Language Processing – NLP) va sun’iy intellekt (Artificial Intelligence – AI) texnologiyalarining jadal sur’atlarda taraqqiy etishi natijasida avtomatik matn tahrirlovchi tizimlar lingvistik xatolarni aniqlash, stilistik tahlil, semantik moslikni ta’minlash kabi murakkab vazifalarni muvaffaqiyatli bajarmoqda. Ushbu vositalar grammatika, imlo, uslub, tinish belgilarini avtomatik ravishda aniqlab, real vaqt rejimida foydalanuvchiga tahririy takliflarni ilgari suradi. Maqolada Grammarly, LanguageTool, ProWritingAid, Microsoft Editor, Hemingway kabi zamonaviy platformalarning texnologik yechimlari, mashina o‘rganish va chuqur o‘rganish modellariga tayanib ishlashi, tilga moslashuvchanlik darajasi va interfeys imkoniyatlari muqobillik asosida tahlil qilinadi. Ushbu vositalarning funksional salohiyati, ko‘p tillilikni qo‘llab-quvvatlash darajasi va ularning algoritmik ishlash mexanizmlari blok-sxema va jadvallar yordamida ifodalanadi. Tadqiqot natijalari shuni ko‘rsatadiki, avtomatik

matn tahrirlovchilar nafaqat matn sifatini oshirish, balki professional yozuv madaniyatini shakllantirishda ham muhim vositaga aylanmoqda.

Annotation. Automatic text editing software uses artificial intelligence (AI) and natural language processing (NLP) technologies to automatically edit texts, correct grammatical errors, and improve stylistic aspects. These programs are used not only for grammar and spell checking, but also for providing contextual suggestions, stylistic optimization, and automating the writing process.

In global practice, automatic text editing tools significantly simplify the writing process and help improve text quality and clarity. These tools are used in various fields – academic writing, business documents, journalism, and everyday communication. With the development of artificial intelligence and natural language processing technologies, these tools are becoming more advanced and adaptive.This scientific article presents a systematic analysis of the formation, stages of development, and practical application of automatic text editing tools based on global experience. In recent years, due to the rapid advancement of NLP (Natural Language Processing) and AI (Artificial Intelligence) technologies, automatic editing systems have been able to successfully perform complex tasks such as identifying linguistic errors, conducting stylistic analysis, and ensuring semantic compatibility.These tools automatically detect grammar, spelling, style, and punctuation errors and offer real-time editorial suggestions to users.

The article analyzes the technological solutions, machine learning and deep learning models, language adaptability, and interface features of modern platforms such as Grammarly, LanguageTool, ProWritingAid, Microsoft Editor, and Hemingway. The functional capacity, multilingual support level, and algorithmic mechanisms of these tools are illustrated using flowcharts and tables. The research findings show that automatic text editors are becoming not only tools for improving text quality but also important instruments in fostering professional writing culture.

Аннотация. Программы автоматического редактирования текста используют технологии искусственного интеллекта (ИИ) и обработки естественного языка (NLP) для автоматического редактирования текстов, исправления грамматических ошибок и улучшения стилистики. Эти программы применяются не только для проверки грамматики и орфографии, но и для предоставления контекстуальных предложений, стилистической оптимизации и автоматизации процесса письма.

Мировой опыт показывает, что автоматические редакторы значительно упрощают процесс написания текста, способствуют повышению его качества и точности. Эти инструменты используются в различных сферах – академическое письмо, деловая документация, журналистика и повседневная переписка. С развитием технологий искусственного интеллекта и обработки

естественного языка данные программы становятся всё более совершенными и адаптивными. В данной научной статье представлен системный анализ формирования, этапов развития и практического применения автоматических средств редактирования текста на основе международного опыта.

В последние годы, благодаря стремительному прогрессу технологий NLP (Natural Language Processing) и AI (Artificial Intelligence), автоматические редакторы успешно выполняют сложные задачи, такие как обнаружение лингвистических ошибок, стилистический анализ и обеспечение семантической согласованности. Эти инструменты автоматически выявляют грамматические, орфографические, стилистические ошибки и пунктуацию, а также предлагают пользователю редакционные предложения в режиме реального времени.

В статье проводится сравнительный анализ технологических решений современных платформ, таких как Grammarly, LanguageTool, ProWritingAid, Microsoft Editor и Hemingway, их моделей машинного и глубокого обучения, уровня адаптации к языкам и интерфейсных возможностей. Функциональные особенности, поддержка многоязычия и алгоритмические механизмы этих инструментов представлены с помощью блок-схем и таблиц. Результаты исследования показывают, что автоматические редакторы становятся не только средством повышения качества текста, но и важным инструментом в формировании профессиональной письменной культуры.

Kalit so‘zlar: *NLP, AI, Grammarly, LanguageTool, ProWritingAid, Microsoft Editor, Hemingway, Machine Learning, Deep Learning.*

KIRISH

Axborot texnologiyalarining rivojlanishi bilan tilshunoslik va sun’iy intellekt sohalari o‘zaro yaqinlashdi. Jumladan, avtomatik matn tahrirlash vositalari inson aralashuvisiz matn sifatini oshirish imkonini bermoqda. Ushbu vositalar grammatik xatolarni aniqlaydi, uslubiy takliflar beradi va ba’zan kontekst asosida butun gapni o‘zgartiradi.Oxirgi yillarda axborot texnologiyalari va sun’iy intellekt sohalarining jadal rivojlanishi natijasida tabiiy tilni qayta ishlash (Natural Language Processing – NLP) yo‘nalishi alohida e’tibor markaziga aylandi. Ayniqsa, avtomatik matn tahrirlash texnologiyalari lingvistik sifatni oshirish, yozma nutqni standartlashtirish va inson xatosini minimallashtirishda muhim vositaga aylanmoqda.

Dastlab imlo va grammatik xatolarni aniqlash bilan cheklangan bu tizimlar bugungi kunda stilistik tahlil, kontekstual tahrir, semantik izchillikni baholash, hatto pragmatik moslikni ta’minlash kabi yuqori darajadagi tilshunoslik vazifalarini ham bajarishga qodir.

Ilmiy tadqiqotlar tahlili

Matnda sodir bo‘ladigan imloviy xatolarni avtomatik aniqlash va tuzatish masalasi jahon miqyosida ko‘plab olimlar tomonidan o‘rganilgan. Jumladan, J.Peterson 1980-yillarda kompyuter yordamida matndagi xatolarni aniqlash va ularni avtomatik tuzatish masalalarini atroflicha tadqiq etib, xatolarni tekshiruvchi va to‘g‘rilovchi dasturlar haqida fikr yuritgan[1].Uning ishlarida kompyuter dasturlarining imlo xatolarini topish va tuzatish algoritmlari, xususan, so‘zlarni lug‘at bilan solishtirish hamda tahrir masofasi (masalan, bir harfni tushirib qoldirish, almashtirish, qo‘shish yoki joyini almashtirish kabi xato turlari) asosida aniqlash usullari ilgari surilgan. Bu usullar keyinchalik ko‘plab imlo tekshirgich dasturlarida qo‘llanilgan bo‘lib, ushbu yo‘nalishdagi dastlabki nazariy asoslarni yaratgan. Avtomatik tahrirlash yo‘nalishida avtomatik tahrir qiluvchi dasturlar matnlarni tahrirlashga yo‘naltirilgan. Hatto tatar, boshqird, qozoq, qirg‘iz, ozarbayjon kabi turkiy tillarga mansub matnlar uchun morfologik va sintaktik analizatorlar yaratilgan, MS WORD sistemasiga avto-tahrir paketlari joylashtirilgan [2].

So‘nggi yillarda xorijiy mamlakatlarda imlo xatolarini avtomatik tuzatish bo‘yicha zamonaviy texnologiyalar jadal rivojlanmoqda. An’anaviy lug‘atga asoslangan tekshirish usullaridan tashqari, statistik va mashinaviy o‘rganish usullari joriy etilmoqda. Misol uchun, R.C. de Amorim va M. Zampieri kabi tadqiqotchilar matndagi xatolarni tuzatishda klasterlash algoritmlaridan foydalanish bo‘yicha tajribalar o‘tkazib, xatoli so‘zlarni to‘g‘ri so‘z shakllariga yaqinlashtirishda fonetik o‘xshashlik va klaster tahlilini birlashtirish usulini taklif qildilar [3].

2013–2014-yillarda Bolgariyada o‘tkazilgan tabiiy tilni qayta ishlash bo‘yicha konferensiyalarda taqdim etilgan ushbu ishlarda turli tillardagi imlo xatolarini tuzatish uchun kesh klasterlash va fonetik klasterlash kabi yondashuvlar sinovdan o‘tkazilgan. Natijada, kompyuter imlo tekshirgichlari aniqlay oladigan xato turlari kengayib, noyob va murakkab xatolar uchun ham tuzatish takliflarini shakllantirish imkoniyati oshdi. Shuningdek, ingliz tilidan tashqari boshqa ko‘plab yevropa tillari, xususan, morfologiyasi murakkab bo‘lgan venger va fin tillari uchun maxsus orfografik tekshiruv tizimlari yaratilib, bunday tizimlarda so‘zning barcha grammatik shakllarini qamraydigan lug‘atlar va qoidalar bazasi ishlab chiqildi. Bularning barchasi xorijiy tajribada to‘plangan bilim va texnologiyalarning o‘zbek tilini avtomatik tekshirish dasturlarini yaratishda qo‘llanilishi muhim ilmiy asos bo‘lib xizmat qiladi.

Jahon miqyosida ingliz, rus, fransuz va boshqa tillar uchun imlo xatolarini aniqlash va tuzatish tizimlari ancha rivojlangan bo‘lsa-da, o‘zbek tili uchun bunday tizimlar endi rivojlana boshladi. O‘zbek tilida matnlarni avtomatik qayta ishlash bilan bog‘liq tadqiqotlar so‘nggi yillarda keng rivojlanib bormoqda. Lekin imloviy xatolarni aniqlash va to‘g‘rilash tizimlari hali to‘liq shakllanmagan. Bunday

tizimlarni ishlab chiqish nafaqat lingvistik asoslarni, balki dasturiy yondashuvlarni ham talab qiladi.

B.Elov [4], Sh.Xamroyeva [5], N.Abdurahmonova [6], M. Abjalova [7] va boshqa olimlar tomonidan izchil ilmiy tadqiqotlar olib borilmoqda. Bugungi kunga kelib olib borilayotgan ilmiy ishlar o‘z natijasini ko‘rsatish darajasiga yetib kelgan.

Jahon tajribasida avtomatik matn tahriri dasturlarining rivojlanish bosqichlari

Boshlang‘ich davr (1950-1970-yillar)

- Noam Chomsky tomonidan ishlab chiqilgan **"Transformational-Generative Grammar"** (1957) nazariyasi tabiiy tilni qayta ishlashga asos bo‘ldi.
- IBM va MIT olimlari lingvistik qoidalar asosida imlo tekshirish modellarini yaratishga harakat qilishdi.
- **SPELL** (1961) – IBM tomonidan ishlab chiqilgan ilk imlo tekshirish tizimlaridan biri.
- **FIND** (1965) – AQSh Mudofaa vazirligi tomonidan tarjima qilish va matnni tahlil qilish uchun yaratilgan dastur.

Qoidaga asoslangan yondashuv (1970-1990-yillar)

- **Grammatik tekshiruvchilar**: Microsoft Word ilk versiyalari (1983) asosiy grammatik xatolarni tekshira boshladi.
- **Hemingway Editor** (1985) – matnni o‘qish qulayligi bo‘yicha tahlil qiluvchi dastur ishlab chiqildi.
- 1976-yilda **Donald Knuth** tomonidan **"TeX"** tizimi ishlab chiqildi, bu avtomatik matn formatlash tizimi bo‘lib, ilmiy maqolalar tayyorlashda keng qo‘llanildi.
- 1983-yilda **Robert Allen** tomonidan ilk grammatik analizatorlar ishlab chiqildi.

Statistik modellar va NLP texnologiyalari (1990-2010-yillar)

- **Grammarly (2009)** – statistik NLP algoritmlaridan foydalangan holda grammatik tekshiruv va uslubni yaxshilashni boshlagan dastur.
- **Google Translate (2006)** – avtomatik tarjima tizimi bo‘lib, statistik mashina tarjimasi (SMT) asosida ishlagan.
- 1999-yilda **Google** qidiruv tizimi yaratildi va bu katta hajmdagi matnlarni avtomatik qayta ishlashga imkon berdi.
- 2000-yillarda **Stanford NLP** va **IBM Watson** tabiiy tilni qayta ishlash uchun kuchli algoritmlar ishlab chiqdi.

Sun’iy intellekt va chuqur o‘rganish asosidagi avtomatik tahrir (2010-yildan hozirgacha)

2017-yilda **Google Brain** tomonidan **Transformer** modeli ishlab chiqildi, bu esa ChatGPT kabi texnologiyalarga asos yaratdi.

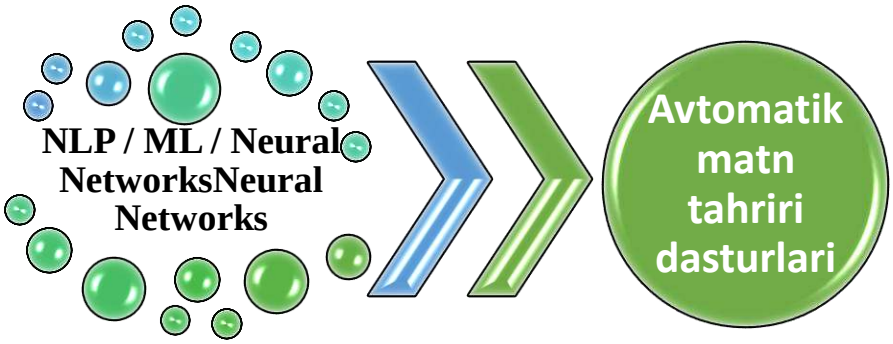
2018-yilda OpenAI **GPT-2** modelini ishlab chiqdi va AI yordamida matn yaratish, tahrirlash va optimallashtirish mumkin bo‘ldi.

- **GPT-3 va ChatGPT (2020-2023)** – sun’iy intellektga asoslangan matn yaratish va tahrirlash imkoniyatlarini taqdim etdi.
- **Quillbot (2017)** – AI yordamida matnni qayta yozish, parafraz qilish va grammatika tahririni amalga oshiruvchi dastur.

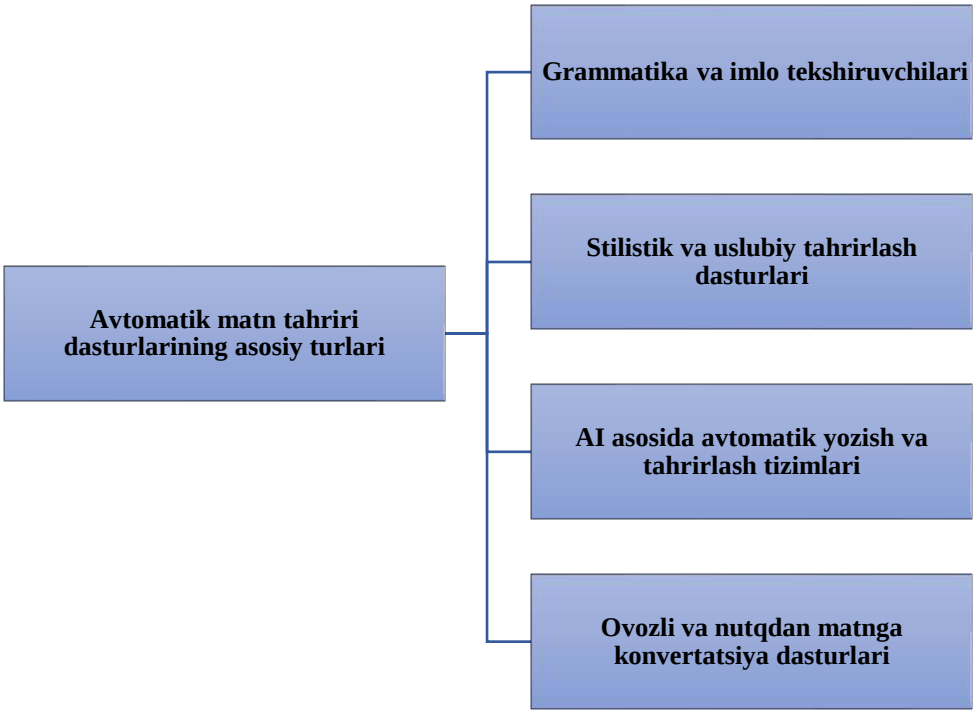
Ushbu dasturlar quyidagi texnologiyalarga asoslanadi:

a. Tabiiy tilni qayta ishlash (NLP) – matnni tushunish va qayta ishlash;
b. Machine Learning (ML) – matndagi xatolarni o‘rganish va yaxshilash;

c. Neyron tarmoqlari (Neural Networks) – matnni kontekstual tushunish va optimallashtirish.



Avtomatik tahrirlovchi vositalarning dastlabki shakllari deterministik, qoidaga asoslangan yondashuvlarga tayangan bo‘lsa, hozirgi zamonaviy tizimlar chuqur o‘rganish (Deep Learning), mashina o‘rganish (Machine Learning), transformator arxitekturalari (masalan, BERT, GPT) kabi ilg‘or texnologiyalar asosida shakllanmoqda. Ushbu vositalar foydalanuvchi tomonidan kiritilgan matnni real vaqt rejimida ko‘p bosqichli analizdan o‘tkazadi: sintaktik va semantik strukturasini tahlil qiladi, grammatik va stilistik nomuvofiqliklarni aniqlaydi hamda to‘g‘rilash bo‘yicha tavsiyalar ishlab chiqadi. Bu jarayon nafaqat tilshunoslik nuqtai nazaridan, balki algoritmik samaradorlik, axborot xavfsizligi va foydalanuvchi tajribasi kabi mezonlar nuqtai nazaridan ham izchil tadqiqotlarni talab etadi.



Avtomatik matn tahrirlash vositalarining turlari

Tahrirlash turi	Tavsifi	Misollar
Grammatik tahrir	Morfologik va sintaktik xatolarni aniqlaydi	Grammarly, LanguageTool
Stilistik tahrir	Matn uslubiga mos bo‘lmagan ifodalarni tuzatadi	ProWritingAid, Hemingway Editor
Kontekstual tahrir	Sun’iy intellekt asosida kontekstga mos variantlarni tanlaydi	Microsoft Editor, ChatGPT Grammarly
Tilga moslashtirish	Ko‘p tilli yoki mashinaviy tarjima matnlarini optimallashtiradi	DeepL Write, Google Translate Editor

Bugungi global amaliyotda Grammarly, LanguageTool, Hemingway Editor, ProWritingAid, Microsoft Editor kabi ilg‘or vositalar nafaqat ingliz tilida, balki ko‘plab boshqa tillarda ham samarali ishlashga moslashtirilgan. Ular o‘z imkoniyatlari bilan jurnalistika, ilmiy maqola yozish, akademik yozuv, tarjima jarayoni va hatto kundalik yozishmalarni sifat jihatidan sezilarli darajada yaxshilamoqda.Avtomatik matn tahrirlash vositalari til texnologiyalarining eng muhim va istiqbolli yo‘nalishlaridan biri sifatida jahon amaliyotida keng qo‘llanilmoqda. Ushbu tizimlar grammatika, imlo, tinish belgilarini to‘g‘rilash, stilistik tahlil va kontekst asosida takliflar ishlab chiqish imkoniyatlari bilan yozma matn sifatini sezilarli darajada oshirmoqda. Ayniqsa, chuqur o‘rganish (Deep Learning), mashina o‘rganish (Machine Learning) va transformer arxitekturalariga asoslangan modellar (BERT, GPT va boshqalar) ushbu vositalarning samaradorligini oshirishda muhim rol o‘ynamoqda.

Maqolada tahlil etilgan Grammarly, LanguageTool, Hemingway, ProWritingAid va Microsoft Editor kabi ilg‘or tahrirlovchi vositalar bugungi kunda ingliz tili va boshqa yirik tillarda keng qo‘llanilmoqda. Ushbu vositalar mashina o‘rganish (ML), chuqur o‘rganish (DL), xususan transformer arxitekturalari asosida ishlovchi model va algoritmlar orqali yuqori aniqlikka ega bo‘lgan, foydalanuvchi ehtiyojiga moslashuvchan va doimiy ravishda o‘zini yangilab boruvchi tizimlardir. Ular o‘quvchilarga, yozuvchilarga, jurnalistlarga, tarjimonlarga, ilmiy tadqiqotchilarga va boshqa mutaxassislarga yozilgan matnni aniq, tushunarli va uslubiy jihatdan to‘g‘ri ifodalash imkoniyatini taqdim etmoqda. Xulosa qilib aytish mumkinki, ushbu avtomatik tahrirlash vositalarining jahon tajribasidagi texnologik asoslarini, lingvistik va texnik imkoniyatlarini, shuningdek, ularning algoritmik ishlash prinsiplarini tizimli ravishda o‘rganish va tahlil qilish orqali tilimiz uchun muqobil tomonlarini andozalash imkoniyatiga ega bo‘lamiz.

O‘zbek tilida imloviy xatolarni avtomatik aniqlash va to‘g‘rilash tizimlarini yaratish murakkab jarayon bo‘lib, bu tilning o‘ziga xos grammatik va fonetik xususiyatlarini inobatga olishni talab etadi. Xususan, o‘zbek tilining qo‘shimchalar tizimi, so‘zlarning turli shakllarda qo‘llanishi, morfologik o‘zgarishlari va fonetik moslashuvlari avtomatik tizimlar uchun qo‘shimcha murakkabliklar tug‘diradi. Shu sababli, bunday tizimni ishlab chiqish nafaqat dasturiy bilimlarni, balki chuqur lingvistik tahlilni ham talab qiladi.

Foydalanilgan adabiyotlar:

1. Peterson, James. Computer Programs for Detecting and Correcting Spelling Errors. 1980.
2. Орехов Б.В., Слободян Е.А. Проблемы автоматической морфологии агглютинативных языков и парсер башкирского языка [Текст] // Информационные технологии и письменное наследие: материалы международной научной конференции (Уфа, 28–31 октября 2010 г.) отв. ред. В. А. Баранов. – Уфа. Ижевск: Вагант, 2010. – С. 167-171.;
3. Zampieri, M.; de Amorim, R.C.Between Sound and Spelling: Combining Phonetics and Clustering Algorithms to Improv Target Word Recovery. Proceedings of the 9th International Conference on Natural Language Processing (PolTAL). Lecture Notes in Computer Science (LNCS). Springer.2014. – P. 438-449.
4. Boltayevich, Elov Botir, Hamroyeva Shahlo Mirdjonovna, and Axmedova Xolisxon Ilxomovna. “Methods for creating a morphological analyzer.” //International Conference on Intelligent Human Computer Interaction. Cham: Springer Nature Switzerland, 2022. https://doi.org/10.1007/978-3-031-27199-1_4
5. Ҳамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари: Филол. фан. бўйича фалсафа докт. (PhD) … дисс. автореф. – Қарши, 2018. – 52 б.
6. Абдурахмонова Н.З. Инглизча матнларни ўзбек тилига таржима қилиш дастурининг лингвистик таъминоти (содда гаплар мисолида): Филол. фан. бўйича фалсафа докт. (PhD)…дисс. –Тошкент, 2018. – 165 б .
7. Abjalova M.A. Tahrir va tahlil dasturlarining lingvistik modullari /monografiya / Toshkent: Nodirabegim, 2020. – 176 b.

UDK: 811.512.133`42

O‘ZBEK TILI MATNLARI USLUBLARARO PEREFRAZ MODELI

Xayatova Zarnigor Marufovna,
doktorant (DSc)
khayatovazarnigor@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada publitsistik uslubdan ilmiy ommabop uslubga, yuqori ilmiy uslubga va badiiy uslubga perefraz qilish jarayonlari ko‘rib chiqiladi. Matnning turli uslublarga o‘zgartirilishi va bu jarayonning tilshunoslikdagi ahamiyati haqida so‘z boradi. Publitsistik uslubdan ilmiy ommabop uslubga o‘zgartirishda ma’lumotlar keng jamoatchilik uchun tushunarli va qabul qilinadigan tarzda taqdim etilishi kerak. Ushbu uslubga o‘tishdagi asosiy maqsad aniq va mantiqiy tuzilishga ega bo‘lgan matn yaratishdir. Badiiy uslubga o‘tkazish esa matnni estetik jihatdan boyitish va hissiy tushunchalarni yetkazishning yangi shakllarini taqdim etishdir.

Abstract. This article discusses the processes of paraphrasing from the journalistic style to the popular scientific style, high scientific style, and literary style. It covers the transformation of texts into various styles and the significance of this process in linguistics. When transforming from journalistic style to popular scientific style, the information must be presented in a manner that is clear and understandable to the general public. The main goal in switching to this style is to create a text with a clear and logical structure. Transitioning to the literary style, on the other hand, aims to enrich the text aesthetically and present new forms of conveying emotions and feelings.

Аннотация. В статье рассматриваются процессы перефразирования с публицистического стиля на научно-популярный, высоконаучный и литературный стили. Обсуждается преобразование текста в различные стили и значение этого процесса в лингвистике. При переходе от публицистического стиля к научно-популярному, информация должна быть представлена в понятной и доступной форме для широкой аудитории. Основная цель перехода к этому стилю — создание текста с ясной и логичной структурой. Переход к литературному стилю, с другой стороны, направлен на эстетическое обогащение текста и представление новых форм передачи эмоций и чувств.

Kalit so‘zlar: *publitsistik uslub, ilmiy ommabop uslub, badiiy uslub, perefraz qilish, matn uslublari, tilshunoslik.*

Kirish. Perefraz qilish jarayoni matnlarni bir uslubdan boshqa uslubga o‘zgartirishda muhim rol o‘ynaydi, bu esa jurnalistika, ilmiy kommunikatsiya va badiiy sohalarda alohida ahamiyatga ega. Perefraz qilish nafaqat axborotni

tushunarli qilishda, balki uning ta’sirini kuchaytirishda ham xizmat qiladi. Ushbu maqolada, publitsistik uslubdan ilmiy ommabop uslubga, yuqori ilmiy uslubga va badiiy uslubga perefraz qilish jarayonlari tahlil qilindi. B.Tayirov o‘z tadqiqotida ilmiy ommabop uslubda “axborotning aniq va to‘liq yetkazilishi”ning ahamiyatiga alohida e’tibor qaratadi va uni “har bir fikrni sodda va tushunarli tarzda ifodalash” sifatida tavsiflaydi [8].

Publitsistik uslubdan ilmiy ommabop uslubga perefraz qilish modeli

Publitsistik uslubdan ilmiy ommabop uslubga perefraz qilishning asosiy maqsadi - ma’lumotlarni keng auditoriyaga, xususan, keng jamoatchilik uchun tushunarli va qabul qilinishini ta’minlashdir. Publitsistik uslubdagi matnlar ko‘pincha hissiy tarzda yoziladi va ommaviy axborot vositalariga moslashtiriladi. Shuning uchun, bu matnlarni ilmiy ommabop uslubga o‘zgartirishda, ularning aniq va mantiqiy tuzilishiga alohida e’tibor qaratiladi [3]. Ilmiy ommabop uslub esa to‘g‘ri va yorqin, bir vaqtning o‘zida juda tushunarli bo‘lishi kerak. Bu usuldagi tekstlar o‘z fikrlarini aniq, sodda va mantiqiy tarzda bildirishga qaratilgan bo‘ladi, shuningdek, maxsus atamalardan foydalanish va ularni tushunarli qilish muhimdir [2].

Masalan, quyidagi publitsistik uslubdagi matnni ko‘rib chiqamiz:

“Bizning jamiyatimizda ta’lim olish huquqi har bir insonning asosiy huquqi hisoblanadi. Bu huquqning amalga oshirilishi jamiyatning kelajagi uchun juda muhimdir. Ta’lim orqali kelajakda yuksak bilim va ma’lumotga ega bo‘lgan insonlar jamiyatni rivojlantirishga yordam beradi. Shuning uchun, ta’lim olish huquqi butunlay amalga oshirilishi kerak.”

Ushbu matnning asosiy maqsadi – ta’lim olish huquqining ahamiyatini ta’kidlash va jamoatchilikni ta’lim olish huquqini amalga oshirishga chaqirishdir. Matnda emotsional ta’sir va ijtimoiy muammoga urg‘u berilgan. Ammo ilmiy ommabop uslubga o‘zgartirishda bu fikrlarni yanada aniq va mantiqiy tarzda ifodalash kerak bo‘ladi.

Ilmiy ommabop uslubga perefraz qilish

Ilmiy ommabop uslubda, matnning maqsadi ma’lumotlarni aniq, sodda va ilmiy tarzda yetkazishdir. Ushbu matnni ilmiy ommabop uslubda quyidagicha o‘zgartirish mumkin:

“Ta’lim olish huquqi har bir fuqaroning asosiy huquqlaridan biridir. Uning amalga oshirilishi jamiyatning barqaror rivojlanishiga xizmat qiladi. Ta’limga kirish imkoniyati, o‘z navbatida, shaxsning ilmiy va professional qobiliyatlarini rivojlantiradi. Shuning uchun ta’lim olish huquqining amalga oshirilishi jamiyatdagi barcha a’zolar uchun zaruriy shartdir.”

Yuqoridagi misolda, publitsistik uslubdagi matnning emotsional va aks ettirilgan ijtimoiy chaqiruvlari ilmiy ommabop uslubda aniq va mantiqiy tarzda ifodalangan. Ilmiy ommabop uslubda, asosan, matnning ma’nosi aniqroq bo‘lib, xususiyatlari va maqsadi haqida qat’iy ilmiy ifodalash kiritilgan. Bu uslubda jamoatni o‘zaro muloqotga yoki jamiyatdagi ijtimoiy faoliyatga jalb qilish o‘rniga, faqatgina faktlar, tushunchalar va izlanishlar asosida fikrlar bildiriladi [3].

Publitsistik uslubdan ilmiy ommabop uslubga perefraz qilishda, maqsad faqatgina ma’lumotni tushunarli qilish emas, balki uni iloji boricha ilmiy, tushunarli va keng jamoatchilik uchun qabul qilinadigan qilib taqdim etishdir. Bu uslubda ishlatiladigan maxsus atamalar va ilmiy ifodalar yordamida matnning aniqligi va ravshanligi oshiriladi. Bunda, keng jamoatchilikning axborot talablarini ham inobatga olish zarur bo‘ladi. Shuningdek, ilmiy ommabop uslubda qaysi terminlarni ishlatish va ularni tushunarli qilib kiritish, muvaffaqiyatli perefraz qilishning asosiy omillaridan biridir [2].

Publitsistik uslubdagi matnni yuqori ilmiy uslubga o‘zgartirish modeli Publitsistik uslubdan yuqori ilmiy uslubga perefraz qilish juda murakkab jarayon hisoblanadi. Bu jarayonda asosiy maqsad - matnning aniq va zamonaviy ilmiy tilga moslashtirilishi, shuningdek, ilmiy terminologiyalar va atamalarni to‘g‘ri tanlashdir [5]. Bu usulda, odatda, shakllangan muammolarga tegishli aniq faktlar va izlanishlar to‘g‘risida yoziladi. Publitsistik uslubdagi emotsional ta’riflar va xulosalar ilmiy uslubda aniq, statistik ma’lumotlar va eksperimental natijalarga asoslanishi lozim [3].

Misol tariqasida, quyidagi publitsistik matnni ko‘rib chiqaylik:

“Ko‘plab yoshlar uchun ta’lim olish imkoniyati, ular o‘z orzulariga erishish yo‘lidagi eng muhim qadamlardir. Jamiyatda ta’limga bo‘lgan qiziqish ortib borayotgan bir paytda, ko‘plab muammolar ham mavjud. Shunday qilib, ta’lim olish huquqi har bir inson uchun, ayniqsa yoshlar uchun, nihoyatida jamiyatning kelajagi uchun nihoyatda ahamiyatli.”

Ushbu matnni ilmiy uslubda perefraz qilishda quyidagi o‘zgarishlarni amalga oshirish kerak:

“Ta’lim olish huquqi barcha fuqarolar uchun fundamental huquqlardan biridir. Ta’lim olish imkoniyatlarining kengayishi, yoshlarning bilim olishga bo‘lgan ehtiyojlarini qondirishga yordam beradi va jamiyatning barqaror rivojlanishini ta’minlaydi. Zamonaviy jamiyatda, ayniqsa yoshlar orasida ta’limga bo‘lgan ehtiyojlar ortganligi kuzatilmoqda va ta’lim tizimining samaradorligi shu asosda keng tahlil qilinishi zarur” [3].

Yuqoridagi misolda, publitsistik uslubdagi matnning umumiy va emotsional ifodalari ilmiy uslubda aniq va ilmiy ma’lumotlar bilan ifodalangan. Bu uslubda

muhokama qilishda ilmiy tilda ishlatiladigan statistik va empirik ma’lumotlar muhim o‘rin tutadi. Bu metodologiya aniq maqsadni ko‘rsatishga qaratilgan bo‘lib, shaxsiy fikrlar va umumiy ifodalardan farqli ravishda, faktlar va eksperimentlar asosida yozilishi kerak.

Ilmiy uslubga o‘zgartirishda asosiy jihatlar

Ilmiy uslubga o‘tishda, matnning to‘liq o‘zgarishi kerak bo‘ladi, bu nafaqat tilni o‘zgartirishni, balki matnning tarkibini ham yanada aniq va batafsil bayon qilishni talab qiladi. Publitsistik uslubdagi umumiy ifodalardan farqli ravishda, ilmiy matnlar aniq va o‘zgaruvchan ilmiy atamalarni o‘z ichiga oladi. Masalan, “ta’lim olish huquqi” o‘rniga “ta’limga teng kirish imkoniyati” deb aniqroq ta’rif berish mumkin.

Publitsistik uslubdagi matnni yuqori ilmiy uslubga o‘zgartirish jarayoni faqatgina tilni o‘zgartirishni emas, balki matnni ko‘proq aniq va tizimli shaklga keltirishni ta’minlaydi. Bu uslubda ishlatiladigan atamalar va terminologiyalarni to‘g‘ri tanlash, ilmiy va statistik ma’lumotlarga asoslanib fikrlar bildirishning alohida ahamiyati bor [3].

Publitsistik uslubdagi matnni badiiy uslubga o‘zgartirish modeli

Publitsistik uslubdagi matnni badiiy uslubga perefraz qilishda asosiy e’tibor matnning emotsiyali va estetik tarzda o‘zaro bog‘lanishini ta’minlashga qaratilgan. Publitsistik uslubdagi tekstlar tezkor va mantiqiy bo‘lib, badiiy uslubdagi matn esa, birinchidan, umumiy va individual his-tushunchalarni yoritishga [1]. Bu modelda, xususan, simbolizm, metaforalar, obrazlar va hissiy tavsiflar muhim ahamiyat kasb etadi. Bu jarayonda xato yoki noto‘g‘ri talqinlarning oldini olish uchun, matnga badiiy elementlarni qo‘shishda ehtiyotkorlik va aniqlik talab qilinadi [7].

Quyidagi publitsistik uslubdagi matnni ko‘rib chiqaylik:

“Yoshlar uchun ta’lim olish imkoniyati hayotlarining kelajagini shakllantiradi. Ta’lim olish orqali ular jamiyatda o‘z o‘rnini topadi va o‘z orzularini amalga oshirish imkoniyatiga ega bo‘ladi.”

Ushbu matnni badiiy uslubda perefraz qilishda, unga quyidagi o‘zgartirishlarni kiritish mumkin:

“Yoshlar o‘z taqdirini o‘zgartirish uchun ta’limni qo‘lga kiritishi kerak. Har bir yangi bilim, ular uchun yangi bir ufq ochadi, ko‘zlarini ochar ekan, dunyo ularni kutmoqda. Har bir dars, bir qadamingizni kelajakka, yorqin orzular tomon qo‘ygan qadamingizdir.”

Yuqoridagi misolda, publitsistik uslubdagi matnning oddiy va mantiqiy xabarini badiiy uslubda, tasviriy elementlar bilan boyitilgan. Badiiy uslubda, so‘zlar orqali his-tuyg‘ularni ifodalashga, obrazlar yaratishga va o‘quvchini tasavvur

qilishga undashga e’tibor beriladi. Misolda, “yangi bir ufq ochadi” va “kelajakka, yorqin orzular tomon” kabi iboralar badiiy uslubdagi tasviriy so‘zlardir [7].

Badiiy uslubda, ayniqsa, simbolizm va metaforalardan keng foydalanish matnni yanada estetik va estetik jihatdan kuchli qiladi. Publitsistik uslubdagi ma’lumotlarni badiiy tarzda aks ettirishda, xato yoki noto‘g‘ri talqinlarning oldini olish uchun ehtiyotkorlik va aniq tushuncha talab etiladi. Badiiy uslubda, matnni o‘qiyotgan o‘quvchi nafaqat faktlar va ma’lumotlarni, balki ular orasidagi ma’nolarni ham his qilishi kerak.

Publitsistik uslubdan badiiy uslubga o‘tish jarayoni faqatgina so‘zlar va iboralarning o‘zgarishini emas, balki butun matnning his-tuyg‘ular va obrazlar orqali kuchayishini talab qiladi. Bu jarayonning asosiy maqsadi, matnni o‘quvchi uchun yanada ta’sirchan va xayoliy qilishdir, shu bilan birga uning estetik ahamiyatini oshirishdir [1].

Matnni publitsistik uslubdan ilmiy ommabop uslubga, yuqori ilmiy uslubga va badiiy uslubga perefraz qilishning turli modellari ko‘rib chiqildi. Publitsistik uslubning ma’lumotlarni tez va aniq tarzda yetkazishiga bo‘lgan zaruriyati, ilmiy va badiiy uslublarning esa ma’lumotlarni aniq, to‘g‘ri va estetik tarzda tasvirlashga qaratilgan. Bu modellarning o‘rtasidagi farqlar va ularning o‘zgarishlari umumiy tilshunoslik va tilimizning ko‘plab sohalarida amaliy ahamiyat kasb etadi [6].

Foydalanilgan adabiyotlar:

1. Arnold Doug, et al. Machine Translation: An Introductory Guide. Blackwell, 1994.
2. Chen Mia, et al. Best Practices for Transformer Models in Translation. *Journal of Artificial Intelligence Research*, vol. 61, 2018, pp. 445–472.
3. Gaspari Federico, John Hutchins. The Evolution and Challenges of Machine Translation. *Translation Spaces*, vol. 6, no. 2, 2017, pp. 167–189.
4. Hutchins John. Introduction to Machine Translation. UPF, 2003, www.hutchinsweb.me.uk/introMT-UPF-2003.pdf.
5. Koehn Philipp. Statistical Machine Translation. Cambridge University Press, 2009.
6. Pym Anthony. Translation Solutions for Many Languages. - Bloomsbury Academic, 2016.
7. Slocum Jonathan. A Survey of Machine Translation: Its History, Current Status, and Future Prospects. / *Computational Linguistics*, vol. 11, no. 1, 1985, pp. 1–17.
8. Tayirov B. O‘zbek tili va yozma kommunikatsiya. 2014.

UDK: 811.512.133`42

PERIFRAZ GENERATSIYA QILISH ALGORITMLARI TAHLILI

Xayatova Zarnigor Marufovna
doktorant DSc
khayatovazarnigor@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada perifraz generatsiyasi algoritmlarining nazariy asoslari va texnik jihatlari o‘rganiladi. Tadqiqot davomida seq2seq, Transformer va ko‘p tilli modellar (masalan, BERT, T5, mBART) asosidagi yondashuvlar taqqoslanadi. O‘zbek tilining morfologik va sintaktik xususiyatlari tahlil qilinib, o‘zbek tilida perifraz yaratishda yuzaga keladigan muammolar va istiqbolli yondashuvlar muhokama qilinadi. Real misollar asosida o‘zbek tilida perifrazlashning samarali modellari tavsiya etiladi. Maqola natijalari tilshunoslik, NLP va ta’lim sohalarida qo‘llanilishi mumkin bo‘lgan amaliy yechimlarni taklif qiladi.

Abstract. This article explores the theoretical foundations and technical aspects of paraphrase generation algorithms. The study compares approaches based on seq2seq, Transformer, and multilingual models such as BERT, T5, and mBART. Special focus is given to the morphological and syntactic characteristics of the Uzbek language, along with challenges and promising methods for generating paraphrases in Uzbek. Effective models are proposed through real examples. The results of the article offer practical solutions applicable in linguistics, natural language processing (NLP), and education.

Аннотация. В данной статье рассматриваются теоретические основы и технические аспекты алгоритмов генерации парафраз. В ходе исследования проводится сравнительный анализ подходов, основанных на seq2seq, Transformer и многоязычных моделях (BERT, T5, mBART). Особое внимание уделяется морфологическим и синтаксическим особенностям узбекского языка, а также трудностям и перспективам генерации парафраз на узбекском языке. На основе реальных примеров предлагаются эффективные модели перефразирования. Результаты статьи могут быть полезны в прикладных задачах лингвистики, обработки естественного языка (NLP) и образования.

Kalit so‘zlar: *perifraz generatsiyasi, tabiiy tilni qayta ishlash, seq2seq model, Transformer, BERT, T5, mBART, ko‘p tilli model, o‘zbek tili, semantik izchillik.*

Kirish. Perifraz (ya’ni, bir fikrni boshqa so‘zlar bilan qayta ifodalash) tabiiy tilni qayta ishlash (Natural Language Processing – NLP) sohasida muhim

muammolardan biri hisoblanadi. Ushbu sohada tadqiqotlar ko‘p jihatdan mashinaviy tarjima, matnni soddalashtirish, plagiat aniqlash va avtomatik savol yaratish kabi qo‘llanmalarga yo‘naltirilgan. Avvalambor, ingliz tilidagi perifraz generatsiyasi algoritmlari bo‘yicha keng ko‘lamli tadqiqotlar olib borilgan. Masalan, tadqiqotlarda chuqur o‘rganilgan seq2seq (sequence-to-sequence) modellari yordamida perifrazlar yaratish mexanizmlari tahlil qilingan [6]. Tadqiqotda LSTM (Long Short-Term Memory) tarmoqlari va e’tibor (attention) mexanizmlarining samaradorligi ko‘rsatib o‘tilgan.

Shuningdek, taklif qilingan “Stacked Residual LSTM” arxitekturasi perifraz sifatini oshirishda muhim rol o‘ynagan [7]. Tadqiqotchilar tarafidan ishlab chiqilgan model ko‘p bosqichli o‘qitish orqali yanada aniq va kontekstga mos jumlalarni generatsiya qilishga erishgan. So‘nggi yillarda Transformer asosidagi yondashuvlar – xususan, BERT (Bidirectional Encoder Representations from Transformers) va T5 (Text-to-Text Transfer Transformer) modellari perifrazlashda inqilobiy natijalarni ko‘rsatdi [7]. Bu modellarning kuchli tomonlaridan biri – ular katta hajmdagi korpuslar ustida o‘qitilib, kontekstni chuqur anglay oladi.

Turli tillarda, xususan, fransuz, nemis va ispan tillarida ham perifrazlash bo‘yicha tadqiqotlar mavjud. Masalan, Mallinsonlar [10] o‘z tadqiqotida ko‘p tilli perifraz korpuslari yordamida neyron tarmoqlarning umumiylashtirish (generalization) qobiliyatini oshirishga erishgan. Ular tomonidan taqdim etilgan “ParaNMT-50M” korpusi ko‘plab neyron modelni qayta o‘qitish va sinovdan o‘tkazishda ishlatilmoqda. Afsuski, o‘zbek tilida perifraz generatsiyasi bo‘yicha tadqiqotlar hozircha cheklangan. Biroq turkiy tillar, xususan, turk tilidagi yondashuvlar ma’lum darajada o‘zbek tili uchun andoza bo‘lishi mumkin. Çelikyilmaz [11] tomonidan turk tilida ishlab chiqilgan periferik matnlar generatsiyasi tizimi, ayniqsa, semantik izchillikni saqlashda ancha samarali ekani qayd etilgan. Shunday qilib, adabiyotlar sharhi shuni ko‘rsatadiki, perifraz generatsiyasi bo‘yicha algoritmlar izchil rivojlanib bormoqda. Transformer arxitekturasi bu borada yangi bosqichni boshlab berdi. Biroq o‘zbek tilida bu sohada hali yetarlicha tadqiqot olib borilmagan bo‘lib, bu yo‘nalishda izlanishlar boshlanishi dolzarb va istiqbolli hisoblanadi.

Algoritmlarning texnik tahlili

Perifraz generatsiyasi bo‘yicha eng ko‘p qo‘llaniladigan yondashuvlar uch asosiy guruhga ajraladi: birinchisi - an’anaviy neyron tarmoqlar, ikkinchisi - Transformer asosidagi zamonaviy arxitekturalar, va uchinchisi - ko‘p tilli modellar. Har bir guruh o‘ziga xos texnik xususiyatlarga, ishlash mexanizmlariga va vazifaga bog‘liq ustunlik yoki cheklovlarga ega.

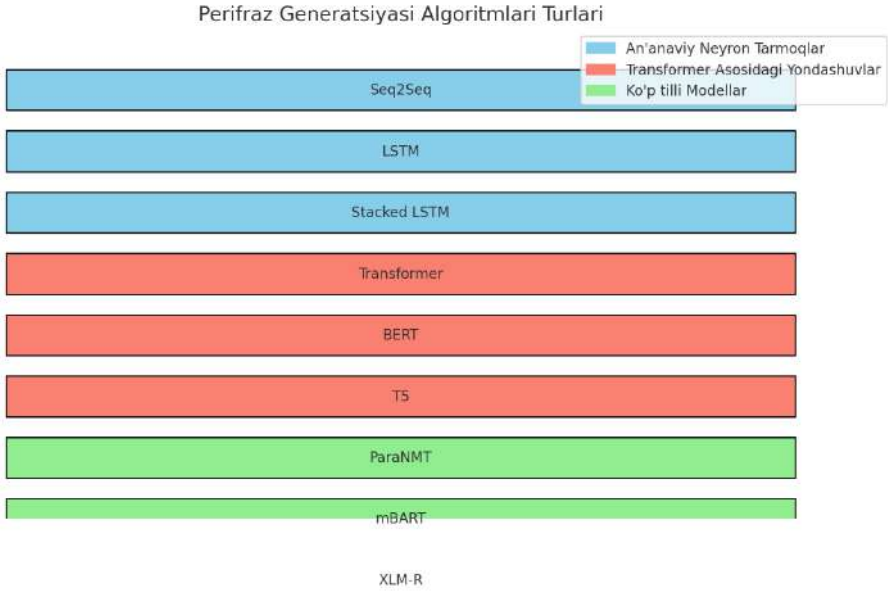
Seq2Seq (sequence-to-sequence) arxitekturasi perifraz generatsiyasida dastlabki samarali yondashuvlardan biri hisoblanadi. Ushbu modelning asosiy tamoyili - berilgan kirish matnini (input sentence) kodlovchi (encoder) blok orqali yashirin holatda ifodalab, uni chiqish matniga (output sentence) dekodlovchi (decoder) blok orqali qayta ishlashdir. Ayniqsa, LSTM (Long Short-Term Memory) hujayralari uzoq muddatli bog‘lanishlarni saqlay olish qobiliyati bilan bu modelga kuchli kontekst anglash imkonini beradi. Biroq ushbu arxitektura ketma-ketlikka asoslanganligi sababli, parallel ishlov berish imkoniyati cheklangan bo‘lib, bu esa katta hajmdagi matnlar bilan ishlashda hisoblash xarajatlarini oshiradi. Bundan tashqari, LSTM modellar murakkab sintaktik yoki semantik o‘zgarishlarni chuqur anglay olishda ko‘pincha qiyinchiliklarga duch keladi [7].

2017-yilda Vaswani va boshqa tadqiqotchilar [8] tomonidan taklif etilgan Transformer arxitekturasi perifraz generatsiyasida yangi bosqichni boshlab berdi. Uning asosiy innovatsiyasi bo‘lgan self-attention mexanizmi modelga matndagi har bir tokenning boshqa tokenlar bilan bog‘liqligini aniqlash va chuqur kontekstni anglash imkonini beradi. Transformerlarning yana bir katta afzalligi — ular ketma-ketlikni ketma-ket emas, balki parallel ravishda qayta ishlaydi, bu esa o‘qitish va inferensiya tezligini sezilarli darajada oshiradi. BERT (Bidirectional Encoder Representations from Transformers) va T5 (Text-to-Text Transfer Transformer) kabi modellarning perifrazlash uchun moslashtirilgan versiyalari mavjud bo‘lib, ular ko‘plab benchmark testlarida yuqori natijalarga erishgan. Masalan, T5 modelining "text-to-text" dizayni uni har xil NLP vazifalariga, jumladan, perifraz generatsiyasiga moslashtirishni osonlashtiradi [3]. Biroq Transformer modellarining katta hajmi va o‘qitish uchun talab etiladigan resurslar ularni cheklangan kompyuter imkoniyatlariga ega tadqiqotchilar uchun qiyinlashtiradi.

So‘nggi yillarda ko‘p tilli (multilingual) modellar perifraz generatsiyasida katta rol o‘ynay boshladi. Ayniqsa, past resursli tillar uchun bu yondashuvlar katta imkoniyatlar yaratmoqda. ParaNMT-50M korpusi yordamida o‘qitilgan modellar turli tillar oralig‘ida o‘xshash ma’nodagi jumlalarni generatsiya qilish imkonini beradi [10]. Bundan tashqari, mBART (Multilingual BART) va XLM-R (Cross-lingual RoBERTa) modellarining ko‘p tilli korpuslarda o‘qitilganligi ularni o‘zbek tiliga o‘xshash tillar uchun ham foydali qiladi. Bu modellar o‘zbek tilida yetarli ma’lumotlar bo‘lmasa ham, boshqa turkiy tillarda o‘qitilgan kontekstni transfer qilish orqali yaxshi natijalarga erishishi mumkin. Ammo bu modellarning asosiy cheklovi — ular o‘zbek tiliga xos bo‘lgan grammatik va stilistik xususiyatlarni chuqur tushunmasligi mumkin, chunki ko‘p tilli korpuslar orasida o‘zbek tilidagi matnlar soni juda kam.

Ushbu texnik tahlil shuni ko‘rsatadiki, perifraz generatsiyasi uchun eng kuchli yondashuvlar Transformer asosidagi modellar hisoblanadi. Biroq har bir modelning o‘ziga xos ishlash doirasi va qo‘llanish sohasi mavjud. Past resursli tillar

uchun ayniqsa, ko‘p tilli modellar yoki transfer learning asosidagi yondashuvlar yanada samarali bo‘lishi mumkin. Shu sababli, algoritm tanlashda til resurslari, vazifa turi va mavjud hisoblash imkoniyatlari e’tiborga olinishi lozim.



1-rasm. Perifraz generatsiyasi uchun qo‘llaniladigan asosiy algoritm turlari

Mana bu diagramma perifraz generatsiyasi uchun qo‘llaniladigan asosiy algoritm turlarini ko‘rsatadi. Ranglar orqali quyidagilar ajratilgan:

- **Ko‘k**: An’anaviy neyron tarmoqlar (Seq2Seq, LSTM)
- **Qizil**: Transformer asosidagi yondashuvlar (BERT, T5)
- **Yashil**: Ko‘p tilli modellar (mBART, ParaNMT)

Algoritmlarning taqqoslov tahlili

Quyidagi jadvalda asosiy uch turdagi perifraz generatsiyasi algoritmlari o‘zaro solishtirilgan. Jadvalda har bir yondashuvning kontekst anglash darajasi, hisoblash murakkabligi, resurs talabchanligi va o‘zbek tiliga moslashuv darajasi keltirilgan.

1-jadval. Uch turdagi perifraz generatsiyasi algoritmlarining o‘zaro solishtiruvi

Algoritm turi	Kontekst anglash	Hisoblash murakkabligi	Resurs talabchanligi	O‘zbek tiliga moslashuv
Seq2Seq + LSTM	O‘rta	O‘rta	Past	Cheklangan
Transformer (BERT, T5)	Yuqori	Yuqori	Yuqori	O‘rta
Ko‘p tilli modellar (mBART, XLM-R)	Yuqori	Yuqori	Yuqori	O‘rta–Yaxshi

Ushbu taqqoslovdan ko‘rinadiki, Transformer asosidagi modellar kontekstni chuqur tushunish qobiliyatiga ega, biroq ularni o‘qitish va ishlatish uchun katta hisoblash resurslari talab etiladi. Ko‘p tilli modellar esa o‘zbek tiliga nisbatan yaxshiroq moslashish imkoniyatiga ega, ayniqsa boshqa turkiy tillar bilan o‘xshashlikdan foydalanilsa.

Perifraz generatsiyasi bugungi kunda ko‘plab real hayotiy vazifalarda samarali qo‘llanilmoqda. Quyida bu algoritmlarning amaliy sohalardagi ishlatilishiga oid bir nechta misollar keltirilgan: Yevropa va AQShdagi ta’lim texnologiyalari kompaniyalari o‘quvchilar uchun matnlarni avtomatik tarzda soddalashtiruvchi ilovalarda perifraz generatsiyasidan foydalanmoqda. Masalan, murakkab ilmiy yoki texnik matnlar oddiy so‘zlar bilan qayta ifodalanadi, bu esa tushunishni osonlashtiradi. Onlayn o‘quv platformalari foydalanuvchilarga test savollarini avtomatik yaratishda perifrazlashdan foydalanmoqda. Masalan, “OpenAI GPT” asosida ishlovchi vositalar berilgan matndan turli xil formulirovkadagi savollar ishlab chiqadi.

Ba’zi akademik platformalar va ilmiy jurnallar perifraz deteksiyasini amalga oshiruvchi tizimlardan foydalanadi. Reklama va marketing sohalarida bir fikrni bir nechta uslubda ifodalash orqali kreativ kontent yaratiladi. Misol uchun, bir necha variantdagi sarlavhalar, postlar yoki reklama matnlari ishlab chiqilishi mumkin. Perifraz generatsiyasi bir nechta tillarda ekvivalent jumlalarni yaratish orqali foydalanuvchiga sinonim variantlar va ifoda shakllarini ko‘rsatadi. Bu yondashuv til o‘rganuvchilarning lug‘aviy boyliklarini oshirishga xizmat qiladi.

O‘zbek tilida perifraz generatsiyasining amaliy jihatlari

Perifraz generatsiyasi algoritmlarini o‘zbek tilida qo‘llashda tilning morfologik boyligi va sintaktik moslashuvchanligi alohida e’tiborni talab qiladi. O‘zbek tili agglyutinativ til bo‘lgani sababli, so‘z yasovchi va qo‘shimcha qo‘shuvchi morfemalarning ko‘pligi perifrazlashda semantik izchillikni saqlab qolish uchun algoritmlardan yuqori darajada kontekstni anglashni talab qiladi.

Masalan, quyidagi jumlani olaylik:

Asl jumla: *Talabalar darsga vaqtida kelishlari kerak.*

Perifraz varianti: *O‘quvchilar mashg‘ulotlarga o‘z vaqtida yetib kelishlari lozim.*

Bu yerda “talabalar” → “o‘quvchilar”, “dars” → “mashg‘ulot”, “kerak” → “lozim” kabi sinonimik almashuvlar mavjud bo‘lsa-da, mazmun o‘zgarmagan. Aynan shunday sintaktik strukturalar asosida qurilgan perifrazlar tilni tushunishga ijobiy ta’sir ko‘rsatadi.

Yana bir misol:

Asl jumla: *Bugun ob-havo iliq va quyoshli bo‘ladi.*

Perifraz variantlari:

- *Havo bugun iliq bo‘lib, quyosh charaqlab turadi.*
- *Quyoshli va iliq kun bugungi ob-havoni belgilaydi.*

Bu misollar ko‘rsatadiki, o‘zbek tilida perifrazlash nafaqat so‘z darajasida, balki gap tuzilmasi, so‘z tartibi va stilistik uslubni o‘zgartirish orqali amalga oshiriladi. Shuning uchun, perifraz algoritmlari faqatgina so‘z almashtirish emas, balki gap tuzilmalarini qayta qurishga qodir bo‘lishi zarur.

O‘zbek tilidagi periferik generatsiyani amalga oshirish uchun quyidagi arxitektura tavsiya etiladi:

1. **O‘zbek tilidagi annotatsiyalangan kichik korpuslar**: masalan, yangilik matnlari, darsliklar va davlat hujjatlari asosida tuzilgan parallellikdagi variantlar.
2. **mBART modelining transfer learning usuli bilan o‘zbek tiliga moslashtirilgan versiyasi**.
3. **Data augmentation (ma’lumotlarni sun’iy ko‘paytirish)** orqali sinonimlar, gap tuzilmalari va stilistik o‘zgarishlar kiritilgan korpuslar.
4. **BLEU va ROUGE kabi avtomatik metrikalar asosida baholash tizimi**.

Ushbu yondashuv yordamida o‘zbek tilidagi perifrazlash vazifalari – masalan, o‘quvchilarga soddalashtirilgan matn berish, avtomatik savol yaratish, tilni o‘rganishga ko‘maklashish va kontent yaratishda takrorlanuvchi fikrlarni stilistik jihatdan boyitish – yuqori sifatda amalga oshiriladi.

Perifraz generatsiyasi algoritmlari zamonaviy tabiiy tilni qayta ishlash sohasida yuqori ahamiyatga ega bo‘lgan yo‘nalishlardan biri sanaladi. Ushbu maqolada perifrazlash algoritmlarining asosiy turlari, jumladan, an’anaviy seq2seq modellari, Transformer arxitekturasi asosidagi yondashuvlar (BERT, T5) hamda ko‘p tilli modellar (mBART, XLM-R) texnik va nazariy jihatdan tahlil qilindi. Shuningdek, o‘zbek tilida perifraz generatsiyasi qanday amalga oshirilishi mumkinligi haqidagi muammolar va istiqbolli yondashuvlar ham ko‘rib chiqildi.

O‘zbek tili o‘zining morfologik boyligi va sintaktik moslashuvchanligi bilan boshqa ko‘plab tillardan farq qiladi. Shuning uchun bu til uchun perifraz algoritmlarini ishlab chiqishda kontekstni chuqur tushunishga asoslangan yondashuvlar zarur. Maxsus tayyorlangan o‘zbek tilidagi korpuslar, sinonimik ma’lumotlar bazalari va stilistik vositalar asosida ishlov beruvchi algoritmlar aynan tilimiz xususiyatlariga moslashtirilgan bo‘lishi kerak.

Tahlillar shuni ko‘rsatadiki, ayniqsa, Transformer va ko‘p tilli modellar yordamida past resursli tillarda ham samarali perifrazlash mumkin. Agar mBART kabi modellar o‘zbek tilida qayta o‘qitilsa va ma’lumotlar bazasi sun’iy yo‘l bilan boyitilsa, tilimizda yuqori sifatli perifrazlash tizimi yaratish mumkin. Masalan, til o‘rganish ilovalari, avtomatik tarjima tizimlari, kontent generatsiyasi vositalari va plagiat aniqlash dasturlarida bu yondashuvlar muhim rol o‘ynaydi.

Shuningdek, maqolada o‘zbek tilidagi real perifraz misollari tahlil qilinib, jumla sathidagi semantik tenglikni saqlab qolish uchun sinonim almashuv, gap tuzilmasini o‘zgartirish va stilistik moslashuvlarning qanday ishlashi ko‘rsatib o‘tildi. Bu jihatlar algoritmik yondashuvlar uchun nazariy asos yaratadi.

Kelgusida o‘zbek tiliga xos perifrazlash tizimlarini yaratishda lingvistik resurslar – lug‘atlar, frazeologik birliklar, semantik tarmoqlar – asosida modellarni o‘qitish va baholash muhim hisoblanadi. Ushbu izlanishlar ilmiy, amaliy va pedagogik sohalarda keng qo‘llanilishi mumkin bo‘lgan natijalarni taqdim etadi.

Foydalanilgan adabiyotlar:

1. Abdullayev A. O‘zbek tilida ekspressivlik ifodalashning sintaksik usullari. – Toshkent, 1983.
2. Rahmatullayev Sh. O‘zbek tili frazeologik lug‘ati. – Toshkent: G‘afur G‘ulom nashriyoti, 2022.
3. Turansky I. Sementicheskaya kategoriya intensivnosti v angliyskom yazike. – Moskva: Vysshaya shkola, 1990.
4. Sharipov M. Hozirgi o‘zbek tilida so‘z birikmalari sintaksisi masalasi. – Toshkent: Fan, 1978.
5. Statista.com. AI usage in education sector worldwide [https://www.statista.com] (murojaat sanasi: 2025-04-11).
6. Chen, X., Liu, X., Sun, M., Lin, Y. Paraphrase generation with deep neural networks. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018. (pp. 448–457). [Chen, 2018:448].
7. Prakash, A., Hasan, S. A., Lee, K., Datla, V., Qadir, A., Liu, J., Farri, O. Neural paraphrase generation with stacked residual LSTM networks. In COLING 2016. (pp. 1243–1252).
8. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Polosukhin, I. Attention is all you need. In Advances in Neural Information Processing Systems, 2017. (pp. 5999–6009).
9. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Liu, P. J. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21(140), 1–67. 2020.

10. Mallinson, J., Sennrich, R., Lapata, M. Paraphrasing revisited with neural machine translation. In Proceedings of the 15th Conference of the European Chapter of the ACL, 2017). (pp. 400–410). [Mallinson et al., 2017:400].

11. Çelikyilmaz, A., Bosselut, A., He, X., Choi, Y. Generating story endings with planning and style-aware reward. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, 2021.

12. Al-Shameri, N., Al-Khalifa, H. A multilingual study of paraphrase generation models across low-resource languages. Data in Brief, Volume 52, April 2024, 109661. Elsevier, 2024.

TINISH BELGILARINING MATNDAGI AHAMIYATI VA VAZIFALARI

Botir Elov Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Sobirova Zarnigor G‘anijon qizi,
tayanch doktorant
sobirovazarnigor1996@gmail.com
ToshDO‘TAU

Annotatsiya. Tinish belgilarining yozma nutqdagi o‘rni beqiyosdir. Og‘zaki nutqda ohang, pauza, urg‘u kabi vositalar orqali fikr ifoda qilinsa, yozma nutqda ushbu vositalar o‘rnini tinish belgilari egallaydi. Tinish belgilarisiz yozilgan matn na mazmunan tushunarli bo‘ladi, na grammatik jihatdan to‘g‘ri. Har bir tinish belgisi o‘ziga xos vazifani bajaradi va jumla hamda matn tuzilishida muhim ahamiyat kasb etadi. Ayniqsa, ilmiy, rasmiy va badiiy uslubdagi matnlarning to‘g‘ri va aniq ifodalanishida tinish belgilarining o‘rni katta. Ushbu maqolada tinish belgilarining matndagi ahamiyati, ularning grammatik vazifalari, o‘qish va tushunishga ta’siri, shuningdek, noto‘g‘ri qo‘llanilishi oqibatida yuzaga keladigan xatoliklar tahlil qilinadi.

Annotation. Punctuation plays an invaluable role in written language. While spoken language conveys meaning through intonation, pauses, and stress, written language relies on punctuation marks to fulfill these functions. A text without punctuation is neither semantically clear nor grammatically correct. Each punctuation mark serves a specific function and holds structural importance in both sentences and overall text. Especially in scientific, formal, and literary styles, the correct and precise use of punctuation is essential. This article analyzes the importance of punctuation in text, its grammatical functions, its impact on reading and comprehension, as well as common errors arising from incorrect usage.

Аннотация. Пунктуация играет бесценную роль в письменной речи. В устной речи смысл выражается с помощью интонации, пауз и ударений, тогда как в письменной речи эту функцию выполняют знаки препинания. Текст, написанный без знаков препинания, будет непонятным по смыслу и некорректным с грамматической точки зрения. Каждый знак препинания выполняет свою уникальную функцию и имеет важное значение в структуре предложения и текста. Особенно велика роль пунктуации в правильной и точной передаче мысли в научном, официальном и художественном стилях. В данной статье рассматриваются значение знаков препинания в тексте, их грамматические функции, влияние на чтение и понимание, а также ошибки, возникающие в результате их неправильного использования.

Kalit so‘zlar: *NLP, yozma nutq, tinish belgilari, sintaktik vazifa, mantiqiy va semantik vazifa, emotsional-intonatsion vazifa, imlo xatosi, uslubiy xato, ASR, BERT.*

Kirish

Tabiiy tilni qayta ishlash (NLP) – bu kompyuterlarga inson tilini qayta ishlash va tushunish imkonini berish uchun tilshunoslik, informatika va sun’iy intellekt tajribasini birlashtirgan fanlararo soha. NLP texnikasining kuchidan foydalangan holda, bizning imlo va grammatika tekshirgichimiz foydalanuvchilarga aniqroq va kontekstdan xabardor xatolarni aniqlash va tuzatish tajribasini taqdim etishga intiladi. NLP asosidagi tekshiruvchilar, ortografik va grammatik xatolarni aniqlash bilan birga, kontekst, sintaksis va semantika tahlilini amalga oshirib, mo‘ljallangan xabarni yaxshiroq tushunib, aniq tuzatishlar va takliflar taqdim etadi.

Tinish belgilar yozma nutqning ajralmas tarkibiy qismi bo‘lib, ular matnning tuzilishini, ohangini, fikrlar orasidagi mantiqiy bog‘liqlikni ifodalashda muhim rol o‘ynaydi. Og‘zaki nutqda ohang, pauza, urg‘u kabi intonatsion vositalar orqali fikr ifodalanadi. Yozma nutqda esa ushbu vazifani tinish belgilar bajaradi. Tinish belgilar fikrning yakunlanganligini, davomiyligini, turlanishini va emotsional tusini ko‘rsatadi. Ular matnni o‘qishni yengillashtiradi, noto‘g‘ri talqin qilinishining oldini oladi.Yozma nutqning ma’nodorligi, mazmunan aniqligi, grammatik tuzilishi va o‘qishga qulayligi ko‘p jihatdan tinish belgilariga bog‘liq. Shuning uchun ham ularni to‘g‘ri qo‘llash savodxonlik belgisi hisoblanadi. Tinish belgilarini bilish va amaliyotda to‘g‘ri qo‘llash yozma nutq madaniyatining ajralmas qismi sanaladi.

Tinish belgilarining funksional vazifalari:

- **Sintaktik vazifasi**

Tinish belgilarining asosiy vazifalaridan biri – sintaktik vazifa, ya’ni matnda gaplarning tuzilishi va ularning bir-biriga bo‘lgan grammatik munosabatlarini belgilashdir. Tinish belgilar yordamida gaplarning chegarasi, bo‘laklarning o‘zaro bog‘liqligi va ularning izchil oqimi ifodalanadi. Masalan, murakkab gaplar tarkibidagi sodda gaplarni vergul yoki nuqtali vergul yordamida ajratish orqali o‘quvchi fikr oqimini aniq anglay oladi.

- **Mantiqiy va semantik vazifasi**

Tinish belgilarning matndagi mantiqiy bog‘liqlikni belgilashda ham katta rol o‘ynaydi. Gaplar orasidagi sabab-oqibat, qarama-qarshilik, izohlilik kabi semantik munosabatlar tinish belgilar orqali oson tushuniladi. Masalan, “lekin”, “ammo”, “shuning uchun” kabi bog‘lovchilar qatnashgan gaplarda vergul qo‘yilishi bu fikrlar orasidagi mantiqiy aloqani yanada ravshanlashtiradi.

- **Emotsional-intonatsion vazifasi**

Og‘zaki nutqda his-tuyg‘ular ohang,pauza va urg‘ular orqali ifodalanadi. Yozma nutqda esa bu vazifani tinish belgilari bajaradi. Undov belgisi (!), so‘roq belgisi (?), ikki nuqta (:), qo‘shtirnoq (“”) kabi belgilar matnga emotsional va ekspressiv tus bag‘ishlaydi. Ular yordamida o‘quvchi matn ohangini sezadi, muallifning munosabatini tushunadi.

Turli uslubdagi matnlarda tinish belgilarining ishlatilishida o‘ziga xosliklar bo‘ladi:

Ilmiy uslub: aniq, lo‘nda va faktlarga asoslangan. Tinish belgilarining qat’iy qo‘llanilishi talab etiladi.

Badiiy uslub: undovlar, iqtiboslar, pauzalar ko‘p bo‘ladi. Tinish belgilarining ko‘plab turi qo‘llaniladi.

Publitsistik uslub: hissiylik va ta’sirchanlik uchun undov va so‘roq belgilaridan tez-tez foydalaniladi.

Tinish belgisi maqsad va nutqiy vaziyatdan kelib chiqqan holda qo‘yiladi.[1]

Imlo qoidalariga, uslubiy me’yorlarga amal qilinmagan hollarda matnda xatolarga yo‘l qo‘yiladi.

Xatolarning quyidagi turlari bor:

1. **Imlo xatolari;**
2. **Punktuatsion xatolar;**
3. **Uslubiy xatolar;**
4. **Mantiqiy xatolar.**

Biz ushbu maqolada aynan tinish belgilari bilan bog‘liq xatolar yuzasidan fikr yuritamiz.

Tinish belgilari xatosi – tinish belgilarini qo‘llash bilan bog‘liq xatolar. Maqsad, fikr og‘zaki nutqda ma’lum ohang bilan, yozuvda esa gapning grammatik qurilishi va tinish belgilari orqali aks ettiriladi.

Tinish belgilari fikrni, maqsadni yozuvda aks ettiruvchi zaruriy vositadir.

Tinish belgilarining turlari va ularning o‘rni

O‘zbek tilida bir necha asosiy tinish belgilar mavjud bo‘lib,ularning har biri o‘ziga xos vazifa va ahamiyatga ega. Quyida ularning eng muhimlari haqida ma’lumot beriladi:

1. **Nuqta (.)**

Nuqta xabar gaplarning oxiriga qo‘yiladi va gapning tugaganligini bildiradi. U matnni mantiqiy ravishda bo‘lib turadi. Masalan: Bugun havo issiq.

2. **Vergul (,)**

Vergul gapdagi bir jinsli bo‘laklarni, qo‘shma gap qismlarini, kirish so‘z va izohlovchi gaplarni ajratish uchun ishlatiladi. Masalan: Biz olma, nok, shaftoli va uzum oldik.

3. **Nuqtali vergul (;)**

Murakkab gap tarkibidagi mustaqil mazmunga ega qismlarni ajratish uchun ishlatiladi. Masalan: U tez gapirardi; men esa uni zo‘rg‘a tushunardim.

4. **Ikkita nuqta (:)**

Izoh, ro‘yxat yoki iqtibos oldidan qo‘yiladi. Masalan: Men quyidagilarni oldim: non, choy, shakar.

5. **So‘roq belgisi (?)**

Savol mazmunidagi gaplar oxirida ishlatiladi. Masalan: Sen bu filmni ko‘rganmisan?

6. **Undov belgisi (!)**

Hayrat, quvonch, buyruq, ogohlantirish va boshqa kuchli emotsiyani ifodalovchi gaplarda qo‘llanadi. Masalan: Juda yaxshi ishlading!

7. **Qo‘shtirnoq (“”)**

Bevosita nutq, iqtibos, ironiya yoki maxsus so‘zlarni ajratish uchun ishlatiladi. Masalan: U menga “Men bu ishni bajaraman”, – dedi.

8. **Qavs (())**

Qo‘shimcha tushuntirish, aniqlashtirish kiritiladigan so‘z yoki iboralar atrofida ishlatiladi. Masalan: Alisher Navoiy (1441-1501) o‘zbek adabiyotining asoschisi hisoblanadi.

9. **Tire (–)**

Izohlovchi gaplardan oldin, ikki so‘z orasidagi bog‘liqlikni ko‘rsatish uchun qo‘llanadi. Masalan: Ona – muqaddas zot.

10. **Ko‘p nuqta (…)**

Mazmunan tugallanmagan gaplar oxirida ko‘p nuqta fikrda uzilish borligini ko‘rsatadi va gapning oxirida yoki, umuman, u qo‘yilgan o‘rinlarda uzun va ba’zan qisqa pauza bo‘ladi.Masalan:

Devorda viloyat xaritasi, tuman hududining rejasi turli-tuman chizgilar, foto hujjatlar... Hammasi ham unib o‘sdi: biri kolxoz raisi, biri rayon madaniyat bo‘limining mudiri, biri deputat... Axir men... qo‘limdan kelmaydi.

Tinish belgilarining noto‘g‘ri qo‘llanilishi oqibatlari

Tinish belgilarining noto‘g‘ri qo‘llanishi yoki ularni umuman ishlatmaslik yozma nutqda bir qator muammolarni keltirib chiqaradi. Bunday xatolar matnning tushunarsiz bo‘lishiga, ma’no buzilishiga, hatto mutlaqo boshqa ma’no paydo bo‘lishiga sabab bo‘ladi.

- ***Ma’no buzilishi***

Tinish belgilarining noto‘g‘ri qo‘llanilishi natijasida gapning asl mazmuni o‘zgarib ketishi mumkin. Masalan:

- To‘g‘ri: U, aka, sizni chaqiryapti.
- Noto‘g‘ri: U aka sizni chaqiryapti.

Birinchi gapda “aka” murojaat sifatida ajratilgan. Ikkinchi gapda esa murojaat anglashilmaydi.

- ***O‘quvchida chalkashlik paydo bo‘lishi***

Tinish belgilar yo‘qligida matnni o‘qish qiyinlashadi, gap qayerda boshlanib, qayerda tugaganini anglash mushkul bo‘ladi. Bu esa o‘quvchining fikrni noto‘g‘ri talqin qilishiga olib keladi. Masalan:

Tinish belgisisiz:

Bugun maktabda bayram tadbiri o‘tkaziladi barcha o‘quvchilar tayyor bo‘lishi kerak

To‘g‘rilangan:

Bugun maktabda bayram tadbiri o‘tkaziladi. Barcha o‘quvchilar tayyor bo‘lishi kerak.

- ***Muloqotdagi uzilishlar***

Yozma muloqotda tinish belgilar fikrlarning aniqligini va izchilligini saqlaydi. Agar ular ishlatilmasa, o‘quvchi yoki tinglovchi matnni to‘liq tushunmasligi mumkin. Ayniqsa, rasmiy hujjatlar, ilmiy matnlar va adabiy asarlarda bu holat jiddiy oqibatlarga olib keladi.

Tinish belgilarini o‘rgatish va o‘zlashtirish muammolari

Tinish belgilarini to‘g‘ri qo‘llashni o‘rgatish — yozma savodxonlikni shakllantirishning muhim qismidir. Biroq amalda bu sohada muayyan muammolar kuzatilmoqda.

Ta’lim tizimidagi yondashuvlar

Boshlang‘ich sinflardan boshlab o‘quvchilarga tinish belgilarini o‘rgatish yo‘lga qo‘yilgan bo‘lsa-da, ko‘pchilikda ushbu bilimlar nazariy darajada qolib

ketmoqda. Darsliklarda mavzu ko‘pincha qoidalarning yodlashiga qaratilgan bo‘lib, amaliy mashg‘ulotlar yetarli emas. O‘qituvchilarning yondashuvi ham bu borada muhim omil sanaladi.

O‘quvchilar duch keladigan muammolar:

- Belgilarning o‘zaro farqini tushunmaslik;
- Qoidalarni amaliyotda tatbiq eta olmaslik;
- Ularni yozma nutqda unutish yoki noto‘g‘ri qo‘llash;
- Murakkab gaplarda tinish belgilarni joylashtirishdagi chalkashliklar.

YECHIM

- Matn tahlillari va xatolar ustida ishlashni yo‘lga qo‘yish;
- Interaktiv o‘quv platformalaridan foydalanish;
- Yozma ishlarni muntazam tahlil qilish orqali o‘quvchilarda mustahkam ko‘nikmalarni shakllantirish;
- Darsliklarni ko‘proq amaliy topshiriqlar bilan boyitish.

Zamonaviy texnologiyalar va tinish belgilar

Bugungi kunda axborot texnologiyalari tez sur’atlar bilan rivojlanmoqda va bu jarayonda til bilan bog‘liq turli sohalar, jumladan, tinish belgilarini avtomatik tarzda tiklash texnologiyalari ham rivoj topmoqda.

Tinish belgilarini tiklashda olimlar eng samarali tizimlar, odatda, prosodik va leksik xususiyatlarni birlashtirgan gibrid modellardan foydalanadilar.[4][6] BERT kabi modellarning paydo bo‘lishiga qadar,leksik xususiyatlar,asosan, takroriy neyron tarmoqlar [4][5]tomonidan qayta ishlangan, shu bilan birga, so‘nggi yondashuvlar transformator [3]arxitekturasidan foydalanadi.Hozirgi zamonaviy metod Courtland va boshq. [5] tomonidan taklif qilingan bo‘lib, bu model oldindan o‘qitilgan BERTdan foydalanadi, bu esa bir xil token uchun bir nechta ehtimollarni yig‘ib, yuqori aniqlik va sezilarli darajadagi parallelizmga erishishga imkon beradi.

Sun’iy intellekt texnologiyalarida “punctuation restoration” — ya’ni tinish belgilarni tiklash muhim yo‘nalish hisoblanadi. Ushbu sohada BERT, RoBERTa, T5, XLNet kabi transformator modellaridan foydalaniladi. Ular yozilgan matnni segmentlarga ajratadi va ularning tuzilishiga qarab to‘g‘ri tinish belgilarini joylashtiradi. Bu texnologiyalar avtomatik tarjima, chatbotlar, hujjatlar tahlili kabi ko‘plab sohalarda qo‘llaniladi.

Nutqni matnga aylantirish tizimlarida (ASR)

Sun’iy intellekt asosidagi nutqni matnga aylantiruvchi tizimlar (Google Speech-to-Text, Yandex SpeechKit, Whisper va boshqalar) intonatsion ohangdan kelib chiqib matndagi tinish belgilarni aniqlashga harakat qiladi.[2] Bu tizimlar hali ham mukammal emas, shuning uchun matnga keyingi ishlov berish (post-processing) zarur bo‘ladi.

Avtomatik nutqni tanib olish (ASR) tizimlari, odatda, tinish belgilarisiz segmentatsiya qilinmagan transkriptlarni chiqaradi. Tinish belgilarini tiklash transkriptsiyalangan nutqni qayta ishlashda muhim qadamdir. Tündik va boshq. [4] transkriptlarda tinish belgilarining yo‘qligi o‘qish osonligini sezilarli darajada pastlashtirishini, bu esa katta so‘z xatoliklar darajasi kabi ta’sir qilishini ko‘rsatdi. Neyron mashina tarjimasi, his-tuyg‘ular tahlili va ma’lumotlarni olish kabi quyidagi vazifalar ham ifodaviy chegaralarga ega bo‘lishdan foyda ko‘radi.

Xulosa

Tinish belgilar yozma nutqning ajralmas qismi bo‘lib, u nafaqat grammatik me’yorlarni, balki fikrning aniq va ravon ifodalanishini ta’minlaydi. Ularning sintaktik, semantik va emotsional vazifalari tufayli matnlar tushunarli, mazmunli va mantiqiy bo‘lib chiqadi. Yozma madaniyatni shakllantirish, savodxonlikni oshirish, ilmiy, badiiy va rasmiy nutqlarda tiniq ifodani ta’minlashda tinish belgilar alohida o‘rin tutadi.

Tinish belgilarining matndagi ahamiyati juda katta. Ular matnning tuzilishini belgilaydi, fikrni aniqlashtiradi, o‘quvchi uchun qulay o‘qish muhitini yaratadi. Har bir tinish belgisi o‘ziga xos sintaktik va semantik yuklama bilan matnga xizmat qiladi. Yozma nutqda fikrning to‘g‘ri va aniq yetkazilishi, aynan, tinish belgilarining o‘z o‘rnida va maqsadga muvofiq qo‘llanilishiga bog‘liq. Shu boisdan, tinish belgilarining o‘rganilishi va amaliyotda to‘g‘ri qo‘llanilishi har bir o‘quvchi va yozuvchi uchun zaruriy ko‘nikmadir.

Texnologik taraqqiyot sharoitida ham tinish belgilar o‘z dolzarbligini yo‘qotgani yo‘q. Aksincha, sun’iy intellekt, avtomatik tarjima va ovozli tizimlarda ularni to‘g‘ri tiklash va qo‘llash yanada muhimlashmoqda. Shu sababli, kelajakda ham tinish belgilarni o‘rganish, o‘rgatish va to‘g‘ri qo‘llash masalalari doimiy e’tiborda bo‘lishi lozim.

Foydalanilgan adabiyotlar:

1. O.Madayev, A.Sobirov, Z.Xolmanova, Sh.Toshmirzayeva, G.Ziyodullayeva, M.Shamsiyeva. Yozma ish turlari: insho, bayon, diktant,2017.
2. Păiș et al., 2021 shared a pretty comprehensive survey on punctuation and casing restoration. It is interesting to see how the model proposed by Susanto et al., is still at the forefront of the SOTA models, 2016.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N.,Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems,2017– pp. 5998–6008

4. Tündik, M.A., Szaszák, G., Gosztolya, G., Beke, A.: User-centric evaluation of

5. automatic punctuation in asr closed captioning,2018.

6. Tündik, M.Á., Tarján, B., Szaszák, G.: A bilingual comparison of maxent-and

7. rnn-based punctuation restoration in speech transcripts. In: 2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom). IEEE,2017–pp.000121–000126.

8. Garg, B., et al.: Analysis of punctuation prediction models for automated transcript generation in mooc videos. In: 2018 IEEE 6th International Conference

9. on MOOCs, Innovation and Technology in Education (MITE). IEEE, 2018–pp. 19–26.

TIL MODELLARI VA CHUQUR O‘RGANISH(DEEP LEARNING) USULLARI:UMUMIY TAVSIF

Sobirova Zarnigor G‘anijon qizi,
tayanch doktorant
sobirovazarnigor1996@gmail.com
ToshDO‘TAU

Annotatsiya. Hozirda sun’iy intellekt, xususan chuqur o‘rganish (deep learning) yutuqlari asosida yaratilgan til modellari inson tilini tushunish, qayta ishlash va yaratish borasida katta yutuqlarga erishmoqda. Til modellari – bu kompyuterlar yordamida inson tilini tushunish va undan samarali foydalanishga imkon beruvchi algoritmlar bo‘lib, ular matnlarni tahlil qilish, tarjima qilish, chatbot yaratish va boshqa ko‘plab vazifalarda ishlatiladi. Ushbu maqolada til modellarining NLPdagi o‘rni, ularning tarixiy rivojlanishi, texnik arxitekturasi, qo‘llanilish sohalari va ayniqsa, o‘zbek tiliga oid tadqiqotlar tahlil qilinadi.

Annotation. Nowadays, language models developed based on artificial intelligence, particularly deep learning advancements, have achieved significant progress in understanding, processing, and generating human language. Language models are algorithms that enable computers to comprehend and effectively utilize human language. They are applied in various tasks such as text analysis, translation, chatbot creation, and many others. This article analyzes the role of language models in NLP, their historical development, technical architecture, areas of application, and especially research related to the Uzbek language.

Аннотация. В настоящее время языковые модели, разработанные на основе достижений искусственного интеллекта, особенно глубинного обучения (deep learning), достигли больших успехов в понимании, обработке и генерации человеческой речи. Языковые модели — это алгоритмы, которые позволяют компьютерам понимать человеческий язык и эффективно его использовать. Они применяются в таких задачах, как анализ текста, перевод, создание чат-ботов и многих других. В данной статье рассматриваются роль языковых моделей в области обработки естественного языка (NLP), их историческое развитие, техническая архитектура, области применения, а также исследования, посвящённые узбекскому языку.

Kalit so‘zlar: *NLP, til modellari, statistik yondashuvlar, RNN, LSTM, transformer, UzBERT, BERTbek, ASR*

Kirish. So‘nggi yillarda sun’iy intellekt va chuqur o‘rganish texnologiyalaridagi yutuqlar natijasida tabiiy tilni qayta ishlash (Natural Language Processing, NLP)

sohasi jadal rivojlanmoqda. NLPning asosiy maqsadi kompyuterga inson tilini tushunish, tahlil qilish va yaratish qobiliyatini singdirishdir. Bu maqsadga erishishda asosiy vositalardan biri – til modellari (language models) hisoblanadi. Til modellari matnni tushunish va yangi matn yaratish uchun mo‘ljallangan neyron tarmoqlar bo‘lib, ular turli arxitekturalarda – rekurrent neyron tarmoqlar (RNN), ularning takomillashtirilgan turlari LSTM va GRU, hamda eng yangi Transformer arxitekturasi kabi modellar ko‘rinishida namoyon bo‘ladi.

Til modeli (Language Model) nima?

Til modellari inson yozgan yoki aytgan matnlar asosida **til qonuniyatlarini o‘rganadigan** va yangi mazmunli matnlar yaratishga qodir algoritmik tizimlardir. Ular matndagi so‘zlar ketma-ketligiga asoslanib keyingi so‘zni bashorat qilish, gapning ma’nosini aniqlash, tarjima qilish, savollarga javob berish kabi ko‘plab murakkab vazifalarni bajaradi.

Til modellari dastlab oddiy **statistik yondashuvlar** (masalan, N-gram modellar) asosida qurilgan bo‘lsa, keyinchalik **neyron tarmoqlarga** asoslangan chuqur modellar — **RNN**, **LSTM**, va ayniqsa **Transformer** arxitekturasi asosidagi yirik modellar (masalan, BERT, GPT, T5) orqali ulkan yutuqlarga erishildi.

Language Modeling

1. Statistical Language Models

- Linguistically motivated
- Adaptive
- Exponential
- Continuous Space
- Decision Tree
- N-Gram

2. Neural Language Models

- Speech Recognition
- Machine Translation
- Sentiment Analysis
- Text Suggestions
- Parsing Tools

3. Pre-trained Language Models

- BERT
- DeBERTa
- biLSTM
- T5
- Switch Transformers

4. Large Language Models

- ChatGPT
- GPT-3
- GPT-4 API
- PaLM
- Galactica
- Minerva
- GLAM
- AlexaTM
- LLaMA
- PanGu-Σ
- Cerebras-GPT
- BloombergGPT
- Falcon

Bugungi kunda til modellari **mashinaviy tarjima, chatbotlar, nutqni matnga aylantirish, matn generatsiyasi, xulosa chiqarish** kabi bir qator amaliy masalalarda asosiy yechim sifatida xizmat qilmoqda.

1. Chatbotlar

- Misollar: ChatGPT, LaMDA
- Tavsif: Tabiiy dialog uchun o‘rgatilgan model

2. Mashinaviy tarjima

- Misol: Google Translate
- Tavsif: Transformer asosida ishlaydi

3. Matn yaratish

- Misollar: GPT-3, GPT-4
- Tavsif: Hikoya, maqola, kod yozish

4. Nutqni matnga aylantirish

- Misollar: Whisper, ASR
- Tavsif: Audio kirishini matnga o‘zgartiradi

5. Savol-javob tizimi

- Misollar: BERT, T5
- Tavsif: Matn asosida savollarga javob beradi

6. Tilni avtomatik aniqlash

- Model: LID

- Tavsif: Matn qaysi tilda yozilganini aniqlaydi

Til modellari bugungi kunda quyidagi sohalarda keng qo‘llanilmoqda:

- **Tibbiyot:** klinik yozuvlarni avtomatik tahlil qilish;
- **Huquq:** hujjatlarni avtomatik tahlil va saralash;
- **Talim:** avtomatik tarjima, esse tekshirish, chat-botlar;
- **Dasturlash:** kod yozuvchi AI tizimlar (masalan, GitHub Copilot).

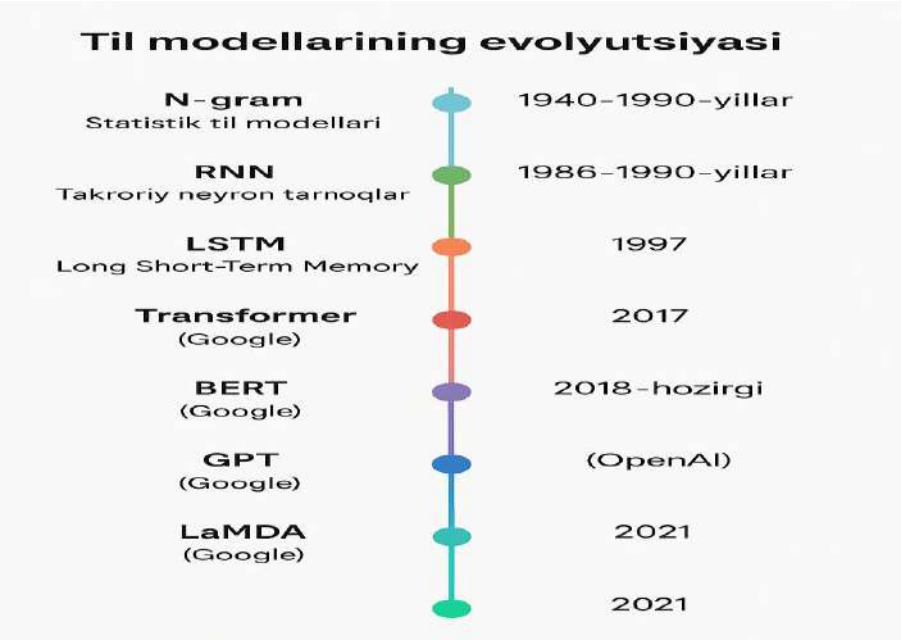
Soha	Model misoli	Izoh
Mashinaviy tarjima	Google Translate	Transformer asosida ishlaydi
Chatbotlar	ChatGPT, LaMDA	Tabiiy dialog uchun o‘rgatilgan model
Matn yaratish	GPT-3, GPT-4	Hikoya, maqola, kod yozish
Nutqni matnga aylantirish	Whisper, ASR	Audio kirishni matnga o‘zgartiradi
Savol-javob tizimi	BERT, T5	Matn asosida savollarga javob beradi
Tilni avtomatik aniqlash	LID model	Matn qaysi tilda yozilganini aniqlaydi

Til modellari tarixi va evolyutsiyasi

Til modellarining rivojlanishi dastlabki statistik modellardan boshlangan. 1980 –1990-yillarda **N-gram** modellar keng qo‘llanilgan bo‘lib, ular berilgan so‘zlar ketma-ketligiga asoslangan ehtimolliklarni hisoblaydi. Biroq bu yondashuvlar uzoq kontekstni hisobga ololmas edi. Bugunga kelib an’anaviy n-gram modellardan boshlab, chuqur neyron tarmoqlarga asoslangan yondashuvlargacha til modellari sezilarli darajada rivojlandi.

Keyinchalik neyron tarmoqlarga asoslangan modellar, xususan **rekurrent neyron tarmoqlar (RNN)**, **Long Short-Term Memory (LSTM)** va **Gated Recurrent Unit (GRU)** kabi modellar paydo bo‘ldi. Ular ketma-ketlikdagi so‘zlar orasidagi murakkab bog‘liqliklarni aniqlash imkonini berdi.

2017-yilda **Transformer** modeli[1] taqdim etilishi bilan yangi bosqich boshlandi. Transformer arxitekturasi parallel hisoblash imkoniyati, self-attention mexanizmi va samaradorligi bilan til modellarining rivojiga inqilobiy turtki berdi.



Til modellar turlari

A) **N-gram modeli (Statistik model):** so‘zlar ketma-ketligi asosida keyingi so‘zni ehtimol asosida aniqlaydi. Masalan, trigram model: $P(w_3 \mid w_1, w_2)$ Faqat uzoq kontekstni tushunmaydi.

B) ***RNN (Recurrent Neural Network):*** *Har bir so‘z ketma-ketlikda o‘rganiladi , har bir qadamda* ***hidden state*** *yangilanadi. Uzoq kontekstni eslab qolishda qiyinchilik bor (*vanishing gradient *muammosi).*

C) ***LSTM (Long Short-Term Memory):*** *RNN asosida yaratilgan.* ***Forget gate, Input gate, Output gate*** *kabi eshiklar bilan uzoq kontekstni eslab qoladi. Matn tarjima, nutq tanish kabi sohalarda ishlatilgan.*

D) ***Transformer :*** *Parallel ishlaydi, self-attention mexanizmi bilan har bir so‘z boshqa so‘zlarga “qaraydi”. Asosiy komponentlar:*

- **Self-attention**
- **Multi-head attention**
- **Position encoding**

GPT, BERT, T5 kabi modellar shu asosda yaratilgan.

Til modellarining asosiy komponentlari:

- **Plan** – Reja tuzish bosqichi.

- **Do** – Amalga oshirish bosqichi.
- **Check** – Tekshirish yoki nazorat qilish bosqichi.
- **Act** – Harakat qilish yoki takomillashtirish bosqichi.

O‘zbek tili uchun tabiiy tilni qayta ishlash (NLP) yo‘nalishida so‘nggi yillarda jadal izlanishlar boshlanib, bir qator resurs va modellar yaratilmoqda.Shu bilan birga, *O‘zbek tili hozircha resurslar jihatidan “low-resource” (resursi cheklangan) til* deb qaraladi – ya’ni kattaroq korpuslar, belgilangan (annotated) ma’lumotlar bazalari va yirik o‘qitilgan modellar kamligi sababli til texnologiyalari sohasida hali mavjud bo‘shliqlar bor[2].

Zero, ilgari O‘zbek tili uchun katta hajmli matn korpuslari va belgilangan ma’lumotlar to‘plamlarini yaratish ustida yetarlicha ish olib borilmagani ta’kidlanadi [3].

Biroq, ayni paytda bu bo‘shliqni to‘ldirish yo‘lida dastlabki qadamlar qo‘yilmoqda – masalan, 2022-yilda O‘zbek tilining morfologik va sintaktik jihatdan belgilangan korpusini yaratish bo‘yicha loyiha doirasida maxsus *POS (so‘z turkumi) va sintaksis* anotatsiya tizimi ishlab chiqilib, birlamchi belgilangan korpus tuzila boshlandi. Shuningdek, O‘zbek tilining Universal Dependencies formatidagi birinchi sintaksis daraxtkorpusi ham yaratilgan bo‘lib, unda 500 ga yaqin jumla sintaktik tahlil bilan belgilangan (ushbu resurs hozircha kichik hajmda bo‘lsa-da,kelgusida kengaytirilmoqda).

Matn korpusi mavjudligi chuqur o‘rganish modellari uchun nihoyatda muhim – chunki yirik **til modellari**ni sifatli o‘qitish uchun katta hajmdagi matn ma’lumotlari talab etiladi. O‘zbek tilida asosiy ochiq matn manbalaridan biri – **O‘zbekcha Vikipediya** bo‘lib, unda taxminan 124 ming maqola mavjudligi haqida xabar berilgan [4], biroq Vikipediya tarkibining bir qismi avtomatik tarjima qilingan yoki rasmiy ensiklopedik matnlardan to‘g‘ridan-to‘g‘ri ko‘chirilganligi tufayli til modeli o‘qitish uchun ma’lum cheklovlar mavjud – jumladan, maqolalarning uslubi qisqa va telegrafik bo‘lishi, yoki ayrim matnlarda kirill-lotin aralash yozuv belgilari uchrashi kabi muammolar aniqlangan. Shu sabab, tadqiqotchilar Vikipediya ma’lumotlarini tozalash va qo‘shimcha manbalar bilan boyitishga e’tibor qaratmoqdalar.

Xususan, yaqinda O‘zbek tilida keng qamrovli **yangiliklar korpusi** shakllantirildi – masalan, Daryo.uz saytining ~200 ming yangilik maqolasidan iborat korpus tuzilib, undagi jumlalar tuzilishi va til jihatdan Vikipediya ma’lumotiga qaraganda ancha xilma-xil va toza ekani qayd etildi. Bunday sifatli korpuslar yirik til modellari uchun asos bo‘lib xizmat qilmoqda.

UzBERT haqida

So‘nggi yillarda O‘zbek tilida ham **chuqur o‘rganishga asoslangan til modellari** paydo bo‘laboshladi. Dastlabki ishlardan biri – Mansurov B. va Mansurov A. tomonidan taqdim etilgan **UzBERT** modeli bo‘lib, u 2021-yilda e’lon qilingan[5].

UzBERT – **BERT arxitekturasi** asosida O‘zbek tilida oldindan o‘qitilgan ilk yirik model hisoblanadi. Tadqiqotchilar ushbu modelni O‘zbek tilidagi matnlarda (kirill va lotin yozuvlaridagi ma’lumotlar bilan) o‘qitib, natijada UzBERT **ko‘p tilli mBERT modeliga nisbatan** O‘zbek matnini tushunish vazifalarida ancha yuqori natijalarni ko‘rsatganini ma’lum qilganlar .Ya’ni, ayni paytgacha O‘zbek tilida alohida model bo‘lmagani sabab mBERT (100+ tilni qamragan ko‘p tilli BERT) ishlatilar edi, biroq monolingual UzBERT o‘sha vazifalarda aniq ustunlikka ega ekani empiric ravishda ko‘rindi.

UzBERT modeli ochiq manbada e’lon qilinib, MIT litsenziyasi ostida hamjamiyatga taqdim etilgan.

Yana bir yondashuv – 2023-yilda Elmurod Kuriyozov va hamkorlari tomonidan yaratilgan **BERTbek** modeli bo‘lib, bu ham O‘zbek tilidagi matnlarda pre-training qilingan Transformer arxitekturali modeldir. BERTbek modelini yaratish uchun tadqiqotchilar maxsus matn korpusi tuzdilar: O‘zbekcha Vikipediya maqolalarining tozalangan nusxasi va yangilik saytlaridan olingan 200 mingdan ortiq maqolalar jamlanmasida model **Masked Language Modeling** vazifasida oldindan o‘qitildi. So‘ngra BERTbek turli vazifalarga moslab fine-tuning qilinib, natijalari baholandi. Xususan, **O‘zbek tilidagi matnlarni toifalash (klassifikatsiya) bo‘yicha yangi ma’lumotlar to‘plami** ustida turli modellarning solishtirma tahlili o‘tkazildi Ushbu tahlilda an’anaviy **qoida asosida** ishlovchi usullar zamonaviy chuqur o‘rganish modellaridan ancha ortda qolishi ko‘rindi – **RNN** va **CNN** kabi neyron tarmoqlar bazaviy usullardan ancha yaxshiroq natija ko‘rsatdi. Ayniqsa, oldindan o‘qitilgan til modeli sifatida BERTbekni qo‘llagan yechim eng yuqori aniqlikka erishdi va boshqa modellardan ustun chiqdi. Bu natijalar O‘zbek tilida **pretraining+fine-tuning** yondashuvi samaradorligini ko‘rsatib, kelgusidagi tadqiqotlar uchun tayanch bo‘lib xizmat qilmoqda. Ta’kidlash joizki, O‘zbek tilidagi modellar faqat matn bilan cheklanmay, ovozli nutq uchun ham rivojlanmoqda: masalan, 2020-yillarda O‘zbek tili uchun bir necha soatlik nutq korpuslari tuzilib, **Automatic Speech Recognition (ASR)** tizimlarida *Deep Learning* modellari qo‘llanila boshlandi. Bunday dastlabki loyihalar O‘zbek nutqini matnga avtomatik o‘girishda ijobiy natijalar bermoqda. Umuman olganda, so‘nggi paytda O‘zbek tilini raqamli muhitda qo‘llashni osonlashtirish uchun **lug‘atlar**, **morfologik tahlilchilar**, **matn korpuslari** va **til modellari** borasida bir qator ochiq loyihalar paydo bo‘ldi. Bu jarayon hali boshida turgani bois, mavjud modellarni

yanada takomillashtirish va katta resurslar yaratish ustida izlanishlar davom etmoqda.

Xulosa

Xulosa qilib aytganda, chuqur o‘rganishga asoslangan til modellari so‘nggi yillarda jadal rivojlanib, *rekurrent arxitekturalar*dan *attention*ga asoslangan Transformer arxitekturalariga o‘tish natijasida sifat va samaradorlik yangi bosqichga ko‘tarildi. Bu modellarning mashinaviy tarjima, dialoglashuv tizimlari (chatbotlar), matn yaratish, tilni avtomatik aniqlash kabi ko‘plab sohalarda muvaffaqiyatli qo‘llanilayotgani kuzatilmoqda. Ayniqsa, yirik *pretrained* modellarning kichik *fine-tuning* bilan moslashuvi turli topshiriqlarda inson darajasiga yaqin natijalarga erishish imkonini bermoqda.

O‘zbek tili kabi resurslari cheklangan tillar uchun ham chuqur o‘rganish asosida ilk yirik modellarning yaratilgani va ularning muloqot, klassifikatsiya kabi vazifalarda ijobiy natijalari mahalliy NLP rivoji uchun muhim poydevor bo‘lib xizmat qilmoqda. Albatta, hali oldinda O‘zbek tilida katta korpuslar tuzish, modellarning yanada takomillashtirilgan versiyalarini yaratish va keng qo‘llash vazifalari turibdi.

Modelni o‘rgatish va optimallashtirishning ilg‘or usullari – attention mexanizmlari, turli optimizatorlar, transfer learning yondashuvlari – esa bu jarayonda tadqiqotchilarga asosiy qurol bo‘lib qoladi.

Zamonaviy ilmiy tadqiqotlar shuni ko‘rsatmoqdaki, chuqur o‘rganishdagi yondashuvlar uyg‘unligi tufayli til modellari tez suratlarda kuchayib, nafaqat global tillar, balki o‘z tilimiz – O‘zbek tilida ham ajoyib natijalar sari siljimoqda. Til modellari tabiiy tilni kompyuter orqali qayta ishlashda muhim rol o‘ynaydi. Ularning arxitekturasi (RNN, Transformer), o‘qitish metodlari (pretraining, attention, optimizatorlar) va qo‘llanilish sohalari har tomonlama chuqur tadqiq etilmoqda. O‘zbek tili uchun ham ilk yirik modellar yaratilmoqda va resurslar boyitilmoqda. Kelgusida O‘zbek tilida ko‘p funksiyali, ochiq manbali va kuchli til modellarining paydo bo‘lishi kutilmoqda.

Foydalanilgan adabiyotlar

1. Vaswani, A., et al. (2017). *Attention is All You Need*.
2. Sanatbek Matlatipov, Hulkar Rahimboeva, Jaloliddin Rajabov, and Elmurod Kuriyozov. 2022. Uzbek sentiment analysis based on local restaurant reviews
3. Maksud Sharipov, Jamolbek Mattiev, Jasur Sobirov, and Rustam Baltayev. 2022. Creating a morphological and syntactic tagged corpus for the uzbek language. CEUR Workshop Proceedings, 3315:93 – 98.
4. https://dumps.wikimedia.org/uzwiki

5. B Mansurov and A Mansurov. 2021b. Uzbert: pretraining a bert model for uzbek. arXiv preprint arXiv:2108.09814.

6. Devlin, J., et al. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT.

7. Brown, T., et al. (2020). *Language Models are Few-Shot Learners*.

8. Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing*. Pearson.

MATNNI PERIFRAZ QILISH

Elov Botir Boltayevich,
Texnika fanlari falsafa doktori, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Ibragimova Mohitob Eldor qizi,
Kompyuter lingvistikasi yo‘nalishi IV kurs talabasi
mohitobibragimova@gmail.com
ToshDO‘TAU

Annotatsiya. Maqolada perifraz tushunchasi va uning lingvistik xususiyatlari keng yoritilgan. Perifraz – so‘z yoki iboralarning semantik jihatdan qayta shakllantirilishi bo‘lib, u matnning ifodalilik darajasini oshirish, plagiatdan qochish va tushunarli bayon qilish uchun ishlatiladi. XX asr tilshunosligida perifraz fenomeni turli yondashuvlar orqali tadqiq qilingan bo‘lib, uning nominativ, ekspressiv, tasviriy va uslubiy funksiyalari aniqlangan. Shuningdek, maqolada perifraz qilish usullari, akademik yozuvda plagiatdan qochish yo‘llari hamda ilmiy manbalarda to‘g‘ri iqtibos keltirish zarurligi batafsil ko‘rib chiqilgan. Perifraz usullari sinonimlar orqali perifraz, gap tuzilishini o‘zgartirish, passiv va faol gaplar orasidagi almashish, mazmunni qisqartirish va kengaytirish kabilarni o‘z ichiga oladi. Maqolada shuningdek, perifraz qilishda foydalanish mumkin bo‘lgan onlayn vositalar haqida ham ma’lumot beriladi. Ushbu tadqiqot perifrazani lingvistik muammo sifatida o‘rganish, uni amaliy yozuvlarda samarali qo‘llash imkoniyatlarini kengaytirishga qaratilgan.

Kalit so‘zlar: *perifraza, XX asr, tasviriy, lingvistika, akademik yozuv, uslub, tilshunoslik, APA uslubi, MLA uslubi, plagiat, iqtibos, faol gaplar(aniq nisbat), passiv gaplar(majhul nisbat), Specific gap(aniq gaplar), general gap(umumiy gaplar), sinonimlar, tasviriy ifodalar, frazeologizmlar, QuillBot*

PERIFRAZ yoki PERIFRAZA – (gr. reriphrasis – yaqin gapiraman so‘zidan) ko‘chim turlaridan biri bo‘lib, unda narsa yoki inson o‘z nomi bilan emas, balki unga xos asosiy belgilar, xususiyatlar nomi bilan yuritiladi. Perifraza ko‘pincha she’riy nutqqa tasviriylik va ta’sirchanlik bag‘ishlaydi[1].

Umuman, dunyo tilshunosligida perifrazalar talqiniga bag‘ishlangan juda ko‘plab monografik tadqiqotlar olib borilgan, barcha izlanishlarda bu atama “perefras”, “perifrazis”, “perifrasis”, “perifraza” “parafraz”, “parafraza”, “parafrazis” kabi fonetik hodisalarning variantlashuvi asosida izohlanganligiga guvoh bo‘ldik. “Parafraza” va “perifraza” atamalarida “para” ham, “peri” ham yunonchadan olingan bo‘lib, “para” – yonida, atrofida, yonma-yon ma’nolarida, “peri” – haqida, to‘g‘risida, xususida ma’nolarida qo‘llaniladi. O‘zbek tilida

perifraza terminini qo‘llashni mantiqan to‘g‘ri deb hisoblaymiz[2]. H.Shamsiddinov ham “perifraza” atamasini qo‘llashni ma’qullaydi[3].

Perifraz qilish - bu biror matnni o‘zgartirish yoki boshqalarning g‘oyalarini o‘z so‘zlaringiz bilan ifodalashdir. Bu jarayon matnning ma’nosini saqlagan holda, so‘zlar va gaplar shaklini o‘zgartirishni o‘z ichiga oladi. Matnni perifraz qilish orqali biz manbaning mazmunini yanada tushunarli va aniqroq ifodalashga erishamiz.

XX asrning 50-yillarida perifrazalar dunyo tilshunosligida, ayniqsa, rus tilshunosligida I.Z.Ilina (1954), V.P.Utkina (1959)lar tomonidan lingvistik jarayon sifatida boshqa hodisa yoki vositalar (istiora (metafora), metonimiya, sinekdoxa, majoz (allegoriya), kinoya, sifatlash (epitet), kichraytirish (litota), mubolog‘a (giperbola)) bilan bog‘liq holda tasvirlangan. Perifrazalar emotsional-ekspressivlik ma’nolarini o‘zida mujassamlashtirgan, so‘zlovchining xarakter xususiyatini ochib beruvchi, tasviriy obrazlar orqali ifodalangan erkin ibora yoki butun jumla shaklida ma’noga ega bo‘lgan tilning uslubiy vositalaridan biri sifatida talqin qilinib kelingan [2]. Perifraza hodisasi jahon tilshunosligida atroflicha tadqiq etilganligi va unga turlicha ta’riflar berilganligiga e’tibor qaratar ekanmiz, tilshunoslikda XX asrning 70- yillaridan boshlab perifrazalarning xususiyatlarini o‘rganishga qiziqish bir qadar oshganligini kuzatamiz. Perifraza lingvistik muammo sifatida birinchilardan bo‘lib V.P.Grigorev tadqiqotlarida so‘z-obraz-denotat munosabatlari nuqtai nazaridan o‘rganildi. Olim uni “poetik nutq hosil qilishning eng yorqin, qiziqarli va istiqbolli usullaridan biri”, deb tavsifladi[4]. I.P.Galperin perifrazaning vazifalarini obyektlar nominatsiyasi, uning predmeti va boshqa nutqiy iboralar o‘rtasidagi munosabatlariga ko‘ra ajratadi. Galperin ta’kidlaydiki, perifraz quyidagi funksiyalarga ega: nominative, tasviriy, boyitish, ekspressiv, muqobil, hissiy-baholash, uslubni shakllantirish, ko‘tarish. I.A. Sidorov perifrazaning funksiyalarini o‘rganar ekan, perifrazaning nominativ, ekspressiv, cheklangan uchta funksiyasini ajratishni taklif qiladi. Biroq, bu funksiyalarning soni turli so‘zlarning turli xil tomondan tahlil qilish uchun yetarli emas. L. V. Grehneva, V. P. Moskvina, M. A. Sirivli asarlari orqali perifraza tushunchasi tushuntirish sifatida ko‘rsatilgan, ular ham ma’lumot beruvchi, boyitish va to‘plash deb nom berishgan. P.Sh.Allayorova perifrazani bu mumkin bo‘lgan qisqa va sodda so‘z o‘rniga uzunroq iborani ishlatishni o‘z ichiga olgan usul hisoblanadi deb ta’kidlaydi[5]. Fransuz tilshunosi Sh.Balli perifrazalarga nisbatan “fikr (jumla)ni takrorlashning kengaytirilgan shakli” sifatida baho berishi munozaralarga sabab bo‘ldi[6]. Yana qator manbalarda perifraza sodda tushunchani murakkab usullar orqali ta’riflash (I.V.Arnold), bir tushunchani bir necha so‘zlar orqali atash (M.V.Lomonosov), fenomenning mantiqiy ifodasini hissiy ifodaga almashtirish (A.A.Potebnya), obrazli sinonim (A.I.Efimov), oddiy so‘zni tavsiflovchi birikma bilan almashtirish (O.S.Axmanova), so‘zni kengroq ifoda etish (Ө.Aytbayұli) sifatida tushuntiriladi[2]. B.Dal perifraza

yunoncha “siqilgan nutqning batafsil, keng tushuntirilgani” degan xulosaga keladi[7].

Perifraz qilishning ahamiyati turli yozuvlarda, ayniqsa, akademik va ijodiy yozuvda katta rol o‘ynaydi. Bu nafaqat boshqa manbalar bilan ishlashni talab qiladi, balki yozuvchiga o‘z ovozini saqlab qolgan holda mazmunni yaratishga imkon beradi. Quyida perifrazning muhimligini ko‘rsatuvchi asosiy sabablar keltirilgan:

1.Plagiatdan qochish uchun - matnni perifraz qilish orqali siz boshqa manbalardan foydalanasiz, lekin ayni vaqtda o‘z so‘zlaringiz bilan ifodalaysiz. Bu esa o‘zingizni plagiatdan himoya qilishning eng samarali usullaridan biridir. Perifraz asl matn fikrlarni ko‘rsatishga imkon beradi, shu bilan birga manbaga hurmat saqlanadi.

2.Mazmunni aniqroq yetkazish - ba’zan asl matn murakkab yoki biz uchun tushunarsiz bo‘lishi mumkin. Perifraz qilish matnni soddalashtirish va bizga osonroq tushunadigan shaklda taqdim etishga yordam beradi.

3. Asosiy fikrlarni qisqartirish - perifraz qilish orqali matnning muhim qismlarini ajratib, ularni qisqacha bayon qilishingiz mumkin. Bu, ayniqsa, katta hajmdagi ma’lumotlarni qisqa va tushunarli shaklda taqdim etishda foydalidir.

4. Til ko‘nikmalarini rivojlantirish - perifraz qilish lug‘at boyligini oshiradi va yozma nutq ko‘nikmalarini rivojlantiradi. Bu jarayon davomida yangi so‘zlarni o‘rganasiz va gaplarni turli usullarda shakllantirishni mashq qilasiz.

5. O‘rganuvchilar bilan bog‘lanish - boshqalarga mos tilda va uslubda yozish uchun perifraz qilish muhim. Masalan, murakkab ilmiy ma’lumotlarni oddiy va aniq so‘zlar bilan qayta ifodalash orqali turli auditoriyalar uchun tushunarli bo‘lishni ta’minlaydi.

Quyida samarali perifraz qilish uchun asosiy usullar keltirilgan(1-jadvalga qarang):

1-jadval. Perifraz qilish usullari

Usul	Izoh	Asl matn	Perifraz qilingan shakli
Sinonimlar, tasviriy ifodalar, frazeologizmlardan foydalanish	Bir so‘z yoki iborani o‘xshash ma’noli boshqa so‘z bilan almashtirish.	“Universitetimiz katta yutuqlarga erishmoqda”.	“Universitetimiz ulkan yutuqlarga erishmoqda”.
So‘z shaklini o‘zgartirish	So‘zlarning turkumini yoki grammatik shaklini o‘zgartirish. Masalan, otni fe’lga, fe’lni sifatga almashtirish.	“Xususiylashtirish bir qancha masalalarda muvaffaqiyatsizlikka uchradi(Bayliss va McKinley (2007)).”	“Bayliss va McKinley (2007) ta’kidlashicha, xususiylashtirishning muvaffaqiyatsizligi ko‘plab omillarga bog‘liq.”

Faol va passiv gaplar o‘rtasida almashish(Aniq nisbatdagi gaplarni majhul nisbatdagi gapga aylantirish, ingliz tilida)	Faol gapni passiv gapga yoki aksincha o‘zgartirish orqali matnga yangilik qo‘shish.	“O‘qituvchi talabalarga yangi mavzuni tushuntirdi.”	“Yangi mavzu talabalarga o‘qituvchi tomonidan tushuntirildi.”
So‘z tartibini o‘zgartirish	Jumlalar ichidagi so‘zlarning joylashuvini o‘zgartirish orqali yangi tuzilma yaratish.	“Talabalar yangi texnologiyalardan foydalanishni o‘rganmoqda.”	“Yangi texnologiyalardan foydalanishni talabalar o‘rganmoqda.”
Jumlalarni qisqartirish yoki kengaytirish	Jumla uzunligini o‘zgartirish orqali matnni yangicha ifodalash.	“Ilm-fan yutuqlari odamlar hayotini sezilarli darajada yaxshilaydi.”	Qisqartirish: “Ilm-fan yutuqlari hayotni yaxshilaydi.” Kengaytirish: “Zamonaviy ilm-fan yutuqlari insonlar turmush sifatini sezilarli ravishda oshirishga xizmat qiladi.”
O‘z so‘zlaringiz bilan ifodalash	Matnni o‘zingizning so‘zlaringiz bilan yozib, asl matn mazmunni saqlash.	“Iqlim o‘zgarishi global muammo bo‘lib, jiddiy choralarni talab qiladi.”	“Global darajadagi iqlim o‘zgarishi muammosi katta e’tibor va qat’iy choralarga ehtiyoj sezadi.”
Specific gapdan general gapga yoki aksincha o‘tish	Maxsus ma’lumotlarni umumlashtirish yoki umumiy ma’lumotni aniqlashtirish.	“Smartfon foydalanuvchilari ijtimoiy tarmoqlarda kuniga o‘rtacha uch soat vaqt o‘tkazadi.”	Umumiylashtirish: “Ko‘pchilik texnologiyalardan kun davomida aktiv foydalanadi.” Aniqlashtirish: “Odamlar ijtimoiy tarmoqlarni smartfon orqali muntazam tekshiradi.”
Hissiyotni o‘zgartirish yoki neytrallashtirish	Matnning hissiyot darajasini o‘zgartirish yoki uni neytral shaklda ifodalash.	“Bu ish juda qiziqarli va hayratlanarli!”	Neytral shakl: “Bu ish juda qiziq.”

Matnni perifraz qilganimizda iloji boricha plagiatdan qochishimiz kerak. Perifraz qilish boshqalarning fikrlarini o‘z so‘zlaringiz bilan ifodalashni anglatadi, ammo bu o‘zingizning g‘oyangiz emas, balki boshqa manbaga asoslanganligini ko‘rsatish uchun manba keltirish zarur. Nima uchun perifraz qilishda manba keltirish kerak?

Manba egasiga hurmat ko‘rsatish uchun: Perifraz orqali sizning yozganingiz aslida boshqa bir muallifning fikri ekanligini tan olasiz.

Plagiatdan qochish uchun: Manba ko‘rsatilmagan perifraz plagiat hisoblanishi mumkin.

To‘g‘ri manbaga yo‘naltirish uchun: O‘rganuvchilar siz taqdim etgan fikrni chuqurroq o‘rganishni xohlasa, manba orqali to‘g‘ri ma’lumotga ega bo‘ladi.

Manba shakli odatda matnda ishlatilayotgan havola uslubiga bog‘liq. Misollar (2-jadvalga qarang):

2-jadval. Perifrazda manba keltirish usullari

Usul	Izoh	Misol
APA uslubi	Matnda muallifning familiyasi, nashr etilgan yili va sahifa raqami ko‘rsatiladi.	“Kompyuter texnologiyalari ish unumdorligini sezilarli darajada oshiradi” (Smith, 2020).
MLA uslubi	Matnda faqat muallifning familiyasi va sahifa raqami ko‘rsatiladi.	“Texnologiyalar ishlab chiqarishda inqilob qildi” (Smith 45).
Havolani jumlaga kiritish		Smith (2020) ta’kidlaganidek, kompyuter texnologiyalari ko‘plab sohalarda samaradorlikni oshirishda muhim rol o‘ynaydi.

Qachon iqtibos keltirish shart bo‘lmaydi? Agar siz umumiy bilim yoki keng ma’lum bo‘lgan faktlar haqida yozayotgan bo‘lsangiz, iqtibos talab qilinmaydi. Masalan:

- Quyosh sharqdan chiqadi.
- Suvning qaynash harorati 100°C ga teng.

Plagiat – bu boshqalarning g‘oyalarini, so‘zlarini yoki asarlarini o‘zlashtirib, o‘z nomidan taqdim etish. Bu odatda ilmiy, adabiy yoki boshqa yozma ishlar kontekstida yuzaga keladi va axloqiy, huquqiy yoki akademik qoidalarni buzadi.

Quyida plagiatning asosiy turlarini ko‘ramiz.

To‘g‘ridan-to‘g‘ri nusxa ko‘chirish: Boshqa birovning matnini iqtibos keltirmasdan yoki manba ko‘rsatmasdan to‘liq yoki qisman ko‘chirish.

Parafrazani noto‘g‘ri ishlatish: Asl manbadan ma’lumotni qayta yozish jarayonida yetarli darajada o‘zgarish kiritmaslik yoki manbani keltirmaslik.

O‘z-o‘zidan plagiat: O‘zining oldingi yozgan ishlarini, yangilik yaratmasdan, yangi vazifalarda qayta ishlatish.

Manbalarni yashirish: Boshqa manbalar asosida yozilgan fikrlarni o‘z nomidan taqdim etish, ammo haqiqiy manbalarni ko‘rsatmaslik.

Vizual plagiat: Rasmlar, grafikalar, jadval va boshqa vizual materiallarni iznsiz ishlatish.

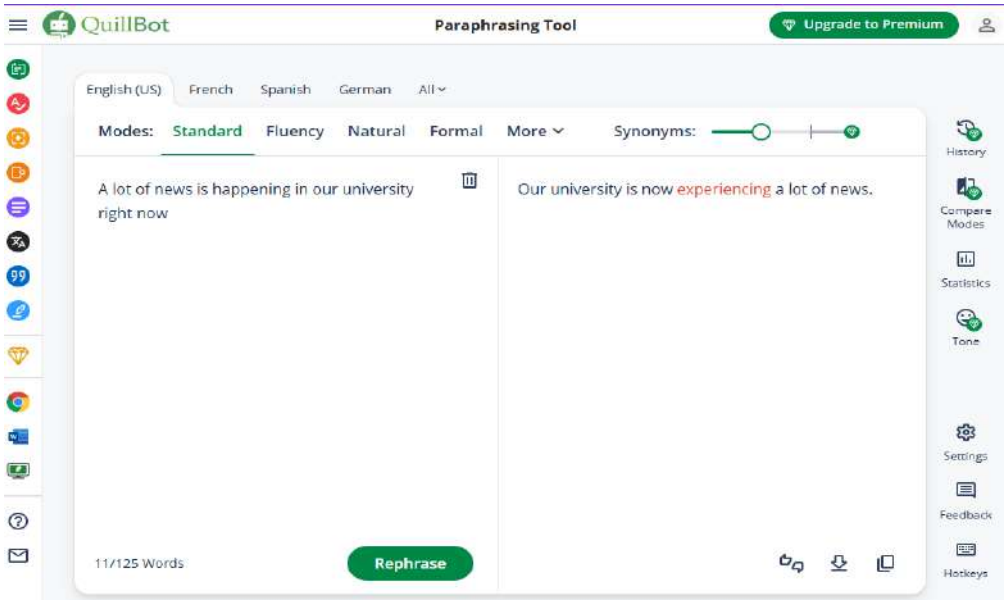
Yarim plagiat: Matnning ba’zi qismlarini o‘zgartirib, lekin asl fikr va g‘oyalarni o‘zgartirmasdan o‘zlashtirish.

Quyidagi perifraz qiluvchi saytlaning barchasi ingliz va ba’zi tillar uchun juda yaxshi natija olish mumkin, ammo hozirgi vaqtda o‘zbek tili uchun perifraz qiluvchi samarali saytlar mavjud emas.

1.QuillBot(1-rasm)

Juda ko‘p maqtovga sazovor bo‘lgan vosita bo‘lib, bepul va pullik versiyalari mavjud.

Xususiyatlari: 8 xil paraphrasing rejimlari (Formal, Creative, Academic va boshqalar), sinonimlar tanlash imkoniyati va Google Docs kabi platformalar bilan integratsiya. Narxi: Bepul; Premium versiyasi $8.33/oydan boshlanadi.



1-rasm. ***QuillBot interfeysi***

2.Grammarly Paraphraser

Matnni qayta yozish bilan birga grammatik va uslub xatolarini tuzatadi.

Xususiyatlari: Matn ravonligini oshirish va plagiyatni oldini olish, interfeysi qulay va oson foydalanish mumkin. Narxi: Bepul versiyasi mavjud; Premium versiyasi $12/oydan boshlanadi.

3. Wordtune

Turli xil yozish ehtiyojlari uchun mos keladigan, juda foydali vosita.

Xususiyatlari: Jumlalarni qisqartirish, kengaytirish va moslashuvchan o‘zgartirish imkoniyatlari. Narxi: Bepul; Premium versiyasi $6.99/oydan boshlanadi.

4.Paraphraser.io

Qulay va ko‘p tilli platforma, o‘quvchilar va yozuvchilar uchun mos.

Xususiyatlari: Bir nechta paraphrasing rejimlari va plagiyat tekshiruv integratsiyasi. Narxi: Bepul; Premium versiyasi $7/oydan boshlanadi.

5.Jasper AI

Ilmiy va biznes yozuvlari uchun ajoyib vosita.

Xususiyatlari: Mantiqiy va kontekstual tahlil bilan matnni yuqori sifatli qayta yozadi. Narxi: $39/oydan boshlanadi.

Perifraz qilish – bu matnni qayta ishlash san’ati bo‘lib, uning yordamida har qanday ma’lumotni yangi ko‘rinishda ifodalash mumkin. Bu usul yozuvchining ijodiy yondashuvi va tilni chuqur tushunishini talab qiladi. Perifraz qilish ko‘p hollarda plagiatdan qochish vositasi sifatida qaraladi, ammo bu jarayonning ahamiyati bundan ancha kengroq.

Birinchidan, perifraz qilish orqali murakkab matnni o‘rganuvchi uchun osonroq va tushunarliroq qilish mumkin. Ushbu usul ma’lumotni turli auditoriyalarga moslashtirishda yordam beradi, bu esa yozuvchining asarlarini kengroq o‘rganuvchilar guruhiga yetkazadi.

Ikkinchidan, bu usul ijodiy va ilmiy yozuvlarda fikrlarni qayta ifodalash, ularni kengaytirish yoki qisqartirish imkonini beradi. Gaplarning shaklini o‘zgartirish, sinonimlardan foydalanish yoki muayyan fikrlarni yangi uslubda taqdim etish orqali har bir matn o‘ziga xos ko‘rinish kasb etadi.

Perifraz qilishda usullar muhim ahamiyatga ega. Sinonimlarni tanlash, gap tartibini o‘zgartirish, grammatik shakllarni moslashtirish kabi usullar matnni boyitadi va mazmunni yangi qirralaridan ko‘rsatadi. Bu usullar orqali yozuvchi o‘z mahoratini namoyon qilishi, o‘rganuvchini matnga jalb qilishi va asl g‘oyalarni yanada o‘tkirroq ifodalashi mumkin. Shuningdek, perifraz qilish orqali yozuvchi matnni plagiatdan xoli qilish bilan birga, manbalar bilan munosib munosabatni saqlab qoladi. Manbalarni to‘g‘ri ko‘rsatish va iqtiboslar bilan ishlash orqali yozuvning ilmiy va axloqiy standartlari ta’minlanadi. Perifraz qilish faqat akademik sohada emas, balki kundalik hayotda, ijodiy yozuvlarda va hatto biznes yozishmalarida ham keng qo‘llaniladi. Bu usul o‘z fikrlarini boshqalarga ta’sirli va aniq ifodalashni xohlagan har bir inson uchun zarur ko‘nikma hisoblanadi.

Yakunida aytish mumkinki, perifraz qilish har bir yozuvchining ijodiy va professional rivojlanishi uchun muhim vositadir. U nafaqat yozuv sifatini oshiradi, balki yozuvchini kengroq auditoriya bilan bog‘laydi, murakkab fikrlarni

soddalashtiradi va har qanday mazmunni yodda qolarli tarzda ifodalashga yordam beradi.

Foydalanilgan adabiyotlar:

1. D. O‘rayeva,D. Quvvatova Jahon adabiyoti terminlarining izohli lug‘ati. –T., 2019. - B. 117
2. Ahmadova U. “Perifraza” atamasi va uning nomlashdagi o‘rni. 2021. - B. 435
3. Шамсиддинов Ҳ. Перифраза хусусида айрим мулоҳазалар // Ўзбек тили ва адабиёти. 1993, № 4. – Б. 29-35. (Shamsiddinov H. Perifraza xususida ayrim mulohazalar // O`zbek tili va adabiyoti. 1993, № 4. – B. 29-35.)
4. Григорьев В.П. К проблеме перифразы // Сб. докл. и сообщений лингвистического общества. – Калинин, 1974. – С.158. (Grigoryev V.P. K problem perifrazi // Sb. dokl. i soobsheniy lingvisticheskogo obshestva. – Kalinin, 1974. – B. 158.)
5. Shahnoza Allayorova Primqulovna Perifrazaning tilishunoslikdagi vazifalari va xususiyatlari. 2023.
6. Балли Ш. Французкая стилистика: Пер. с фр. – Москва, 1961. – С. 126. (Balli Sh. Franzuskaya stilistika: Per. s fr. – Moskva, 1961. – S. 126.)
7. Даль. В.И. Толковый словарь живого великорусского языка. 3-том. – Москва: Издание книгопродавца-типографа М.О. Вольфа, 1882. – 576 с. (Dal V.I. Tolkoviy slovar jivogo velikorusskogo yazika. 3-tom, M., 1882. – 576 s; Shu manba. – 14 s.)
8. https://www.futurelearn.com/info/courses/english-for-study-intermediate/0/steps/35241
9. https://write.smumn.edu/wp-content/uploads/sites/38/2020/09/Paraphrasing-Techniques.pdf
10. https://wisc.pb.unizin.org/esl117/chapter/two-techniques-for-paraphrasing/
11. https://www.lib.sfu.ca/about/branches depts/slc/writing/sources/techniques-paraphrasing
12. https://researcher.life/blog/article/top-7-tried-and-tested-paraphrasing-techniques/
13. https://www.scribbr.com/working-with-sources/how-to-paraphrase/
14. https://uq.pressbooks.pub/academicwritingskills/chapter/paraphrasing-skills/
15. https://www.grammarly.com/blog/summarizing-paraphrasing/paraphrasing-a-sentence/
16. https://quillbot.com/guides/paraphrasing
17. https://libguides.uta.edu/researchprocess/paraphrasing

18. https://owl.purdue.edu/owl/research_and_citation/using_research/quoting_paraphrasing_and_summarizing/paraphrasing.html
19. https://www.verbling.com/ru/articles/post/6-easy-steps-to-paraphrasing-in-english-559ed5e6448aa5ca55a11572
20. https://prowritingaid.com/how-to-paraphrase
21. https://www.grammarly.com/blog/summarizing-paraphrasing/paraphrasing-a-poem/
22. https://blog.wordvice.com/how-to-paraphrase-in-research/
23. https://www.uts.edu.au/current-students/support/helps/self-help-resources/academic-skills/how-paraphrase-effectively
24. https://aithor.com/blog/how-to-paraphrase-a-sentence-effectively-techniques-and-examples
25. https://www.elegantthemes.com/blog/business/best-paraphrasing-tools
26. https://www.papertrue.com/blog/paraphrasing-tools/
27. https://wordvice.ai/blog/best-online-paraphrasing-tools
28. https://quillbot.com/paraphrasing-tool

UDK: 811.512.133`42

KORPUS LINGVISTIKASIDA PARSERLARNING UMUMIY TAVSIFI

Abdullayeva Oqila Xolmo‘minovna,
doktorant DSc
abdullayevaoqila@gmail.com
ToshDO‘TAU

O‘tkirova Fotima Baxtiyor qizi,
Kompyuter lingvistikasi yo‘nalishi IV bosqich talabasi
otkirovafotima@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada korpus lingvistikasining taraqqiyoti va jahonda yaratilgan parserlarning tavsifi, turlari, ularning ishlash prinsiplari, yaratilish maqsadi, hosilalar strukturasi hamda tilshunoslikdagi ahamiyati haqida so‘z boradi. Rivojlanib borayotgan soha lingvistikasida parserlarning o‘rni haqida batafsil ma’lumot beriladi.

Abstract. This article discusses the development of corpus linguistics and the description, types, operating principles, purpose of creation, structure of derivatives, and significance in linguistics of parsers created in the world. Detailed information is provided about the role of parser in the linguistics of the emerging field.

Аннотация. В данной статье рассматривается развитие корпусной лингвистики и описание, типы, принципы работы, цель создания, структура производных и значение созданных в мире парсеров в лингвистике. Предоставляется подробная информация о роли анализаторов в развивающейся отраслевой лингвистике.

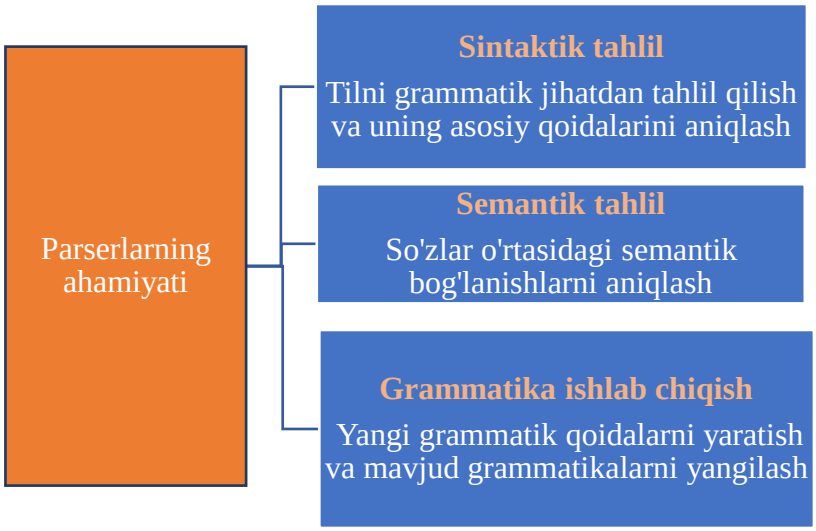
Kalit so‘zlar: *Korpus lingvistikasi, korpus, parser, parser turlari, matn, tahlil, sintaktik tahlil, semantik tahlil, grammatik tahlil, treebank.*

XXI asrning global muammolaridan biri tabiiy tillarning milliy xususiyatini saqlab qolishdan iborat. Dunyo tillarining elektron korpuslarini yaratish va rivojlantirishda NLP hamda til texnologiyalariga doir tadqiqotlarni izchil ravishda olib borish dolzarb vazifaga aylandi[1]. Jahon tilshunosligida korpus lingvistikasining metodologiyasi yoki tilshunoslikning matn korpusini yaratish va tahlil qilish bilan shug‘ullanuvchi bo‘limi til tadqiqotlarini o‘tkazishda keng qo‘llaniladi. Korpus lingvistikasining rivojlanishi bilan matn materiallarining katta ma’lumotlari bazalarini izchil yaratish va kompyuter xotirasida saqlash mumkin

bo‘ldi. Dunyoda tilshunoslikning taraqqiyoti talabi bilan jahon korpus tilshunosligida olib borilayotgan nazariy hamda amaliy izlanishlar bizning yurtimizda ham soha rivoji hamda mamlakat ravnaqi uchun kerakli va zarur nuqta ekani kundan kun ayon bo‘lmoqda.

Korpus lingvistikasi – bu tilni matn korpusi orqali o‘rganishning empirik usuli. Korpuslar – bu ma’lum bir til xilma-xilligini ifodalashga qaratilgan nutq yoki yozuv matnining muvozanatli, ko‘pincha, qatlamli to‘plamlari. Bugungi kunda korpuslar, odatda, mashina tomonidan o‘qiladigan ma’lumotlar to‘plami deya tadqiqotlarda eslatiladi. Ushbu soha XX asrning 60-yillarida kompyuter lingvistikasining bo‘limi sifatida shakllangan tarmog‘i bo‘lib, kompyuter texnologiyalaridan foydalangan holda lingvistik korpuslarni (matnlar korpuslarini) qurish va ulardan foydalanishning umumiy tamoyillarini ishlab chiqish bilan shug‘ullanadi. Lingvistik korpus yoxud til korpusi muayyan lingvistik muammolarni hal qilishga mo‘ljallangan, maxsus saralangan va teglangan elektron shakldagi matnlar majmuasi hisoblanadi va pragmatik ahamiyatga ega eng yirik elektron baza sanaladi. Tilshunoslikda tilni o‘rganish uchun foydalaniladigan matematik va kompyuter asosidagi metodlar kompyuter lingvistikasini tashkil etadi. Bu soha tilshunoslikning boshqa yo‘nalishlaridan farqli ravishda empirik va statistik yondashuvlar bilan tahlil qilishni talab qiladi.

Soha mashinaga asoslangan tahlillarni amalga oshirishda bir qancha dasturlardan foydalanadi. Ana shunday soha texnologiyalarining asosiy vositalaridan biri parserlar (tilni qayta tahlil qilish dasturlari) hisoblanadi. Bunday dasturlar bir necha xil bo‘lib, tilning tuzilishini, grammatikasini va semantik xususiyatlarini tushunish uchun kerak bo‘lgan tizimlarni yaratadi. Korpus lingvistikasi matnlarni qayta tahlil qilishda empirik usullarga tayangani uchun korpusdagi mavjud so‘z, gap va ular orasidagi o‘zaro sintaktik aloqalarni aniqlashda parserlar asosiy vosita sanaladi. Korpuslar – tilni qayta real ishlatishdagi barcha xususiyatlarini o‘zida jamlagan matnlar majmuyi. Tahlil qiluvchi dasturlar matnlarni sintaktik analiz qilib, grammatik modellarni qayta quradi. Korpus lingvistikasi yordamida tilshunoslar grammatik tuzilmalar, so‘z birikmalari, kollokatsiyalar va boshqa til xususiyatlarini tahlil qilishadi. Matnlarning grammatik tuzilishini aniqlashda parserlar muhim dastur sanaladi. Bunday dasturlar, ko‘pincha, quyidagi maqsadlarda qo‘llaniladi:

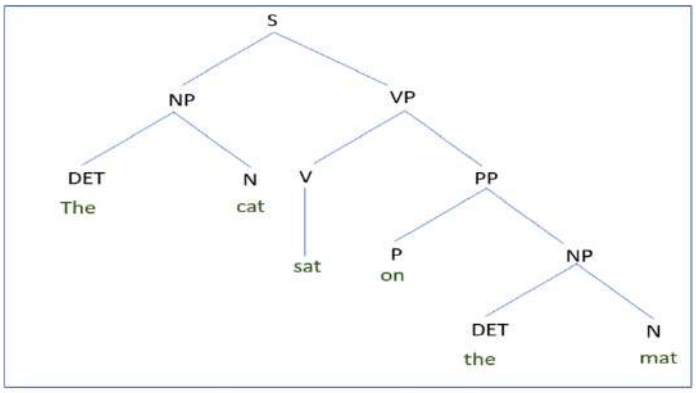


1-rasm. Parserlarning ahamiyati

Parserlarning turi va ularning tavsifi

Matnni grammatik jihatdan tahlil qilib, uning sintaktik strukturasini aniqlovchi dasturiy ta’minot vositalariga **parserlar** deyiladi. Bugungi kunda parserlarning quyidagi turlari mavjud:

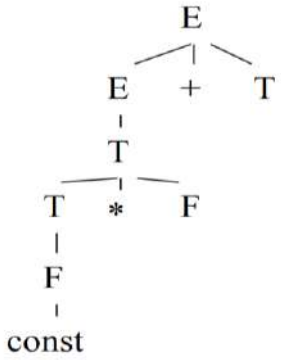
1. **Sintaktik parserlar** kiritilgan gaplarning sintaksis strukturasini aniqlaydi va ushbu strukturaga mos keluvchi sintaktik sxema quradi. Dastlab, jumlalarni sintaktik jihatdan tahlil qiladi. Matnlarni ma’lum bir til grammatikasiga moslashtirib, iboralar va tilda gaplarni tuzish uchun qo‘llaniladigan yaxlit birliklarni ajratib oladi. Sintaktik tahlilning maqsadi qanday qilib tarkibiy qismlar gaplarni shakllantirish uchun ma’lum bir sintaktik tartibda joylashtiriladi, degan savolga javob berish hisoblanadi. Bunday tahlil dasturlarida gaplardagi mavjud so‘zlarning o‘zaro munosabati hamda bog‘liqligi yakuniy natijasini amalga oshirishda muhim rol o‘ynaydi. Zamonaviy tilshunoslikda parsing, asosan, kompyuter yordamida sintaktik tahlil qilishni anglatadi(2-rasm).



2-rasm. Sintaktik parser natijasi

Bu turdagi parserlar ikki turga bo‘linadi:

Top-down (yuqoridan pastga) parserlar – gaplarni grammatik qoidalarga tayanib tahlil qiladi. Bunday parserlar yuqoridan pastga usulidan foydalanadi hamda tahlil sxemasi boshlang‘ich belgidan ildiz sifatida pastki qismdagi terminal belgilarga qarab o‘sib borishi tarzida shakllanadi(3-rasm). Oddiy tillar bilan ishlashda bunday tahlilni amalga oshirish va tushunish pastdan yuqoriga qarab shakllangan tahlildan osonroq.

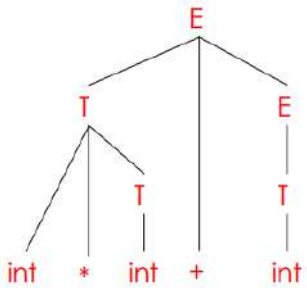


3-rasm. Top-down parserning tahlil sxemasi

Bunday tahlil qilishning yana-da samarali varianti ham bor bo‘lib, u **LL-Parsing** deb ataladi. Bunday tahlilda *Lookhead* belgisi hisobga olinib, keyingi qoidani qo‘llash kerakligini oldindan aytib turish imkoni bo‘ladi. LL sinfidagi parserlar top-down parserlar singari tahlil jarayonida ortga qaytishi shart bo‘lmaydi.

Recursiv Descent Parserlar Top-down parserlarning yana bir turi bo‘lib, o‘ziga xos yuqoridan pastga yo‘natirilgan tahlil dasturi sanaladi. Bu turdagi tahlil usuli rekursiv ravishda ishlaydi va qoidalarni bosqichma-bosqich amalga oshiradi.

Bottom-up (pastdan yuqoriga) parserlar – berilgan matndagi so‘zlardan tortib, grammatik qoidalar yordamida yaxlit sintaktik struktura hosil qiladi. Dastlab, dastur matndagi jumlalarni so‘z birikmalariga ajratib oladi. Keyin ushbu birikmalar umumiy grammatik tuzilishga asoslangan holda guruhlanadi. Oxirida butun jumlaning sintaktik strukturasi hosil qilinadi(4-rasm).

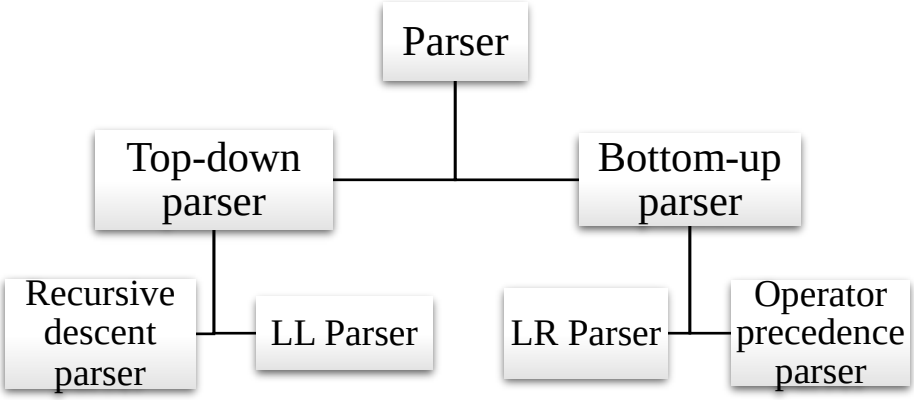


4-rasm. Bottom-up parser sxemasi

Bottom-up parserlarning **LR Parser (Left to right)** va **Operator Precedence Parser** kabi 2 turi mavjud. LR Parserlar jumlani chapdan o‘ngga qarab tahlil qiladi. Bunday dastur tahlili bilan murakkab bo‘lgan grammatik qoidalarni tushunish

mumkin bo‘ladi. Tez ishlaydi hamda kuchli parser hisoblanadi. Operator Precedence Parser, asosan, operatorlarning ustuvorlik darajasini aniqlash uchun qo‘llaniladi. Bunday dastur turlari oddiyroq bo‘lgan grammatik qoidalarni tahlil qilish uchun juda samarali deb baholanadi.

Bugungi kunda jahon tilshunosligida yaratilgan sintaktik tahlil dasturlari tilshunoslikka ko‘plab tahlil usullari va grammatik strukturalarni taqdim etdi. Bu kabi aqlli parserlarning yaratilishi o‘ziga xos bo‘lgan tahlil metodlari tufayli ko‘plab istisnoli sintaktik vaziyatlarga to‘g‘ri yechimlar berdi(5-rasm).



5-rasm. Parserlar klassifikatsiyasi

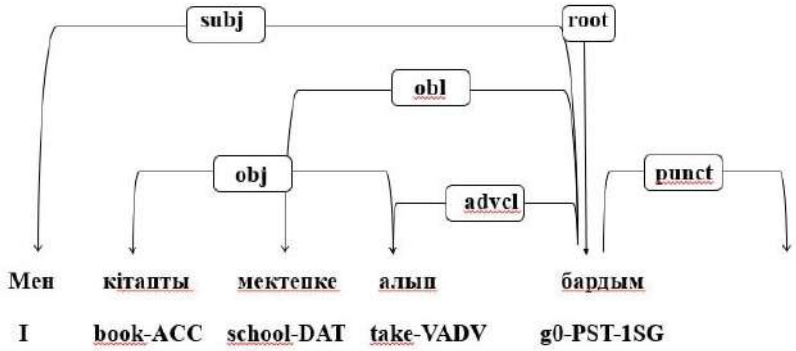
2. Semantik Parserlar matnning ma’no jihatidan tahlilini amalga oshiradi. Ular gaplardagi mavjud leksemalarning o‘zaro bog‘liqligini aniqlash maqsadida ishlab chiqiladi. Semantik parserlar so‘zlarning ma’nosini matnning umumiy mazmuniga aloqador holatda aniqlaydi. Masalan, “Dugonam gulni hidladi” va “Dugonam gulni uzdi” jumlalaridagi *gul* so‘zi har ikki gapda ham turli rol o‘ynagan. Semantik parser yordamida bu farqni aniqlash osonlashadi. Bir qancha muammolarga qaramasdan, bugungi kunda yuqori aniqlikda ishlovchi semantik parserlar yaratilgan. Bu kabi muammo turlaridan ayrimlari tabiiy tildagi gaplarning noaniq bo‘lishi, gaplar murakkab tuzilmalar va munosabatlarga ega bo‘lishidir.

3. Stoxastik va Statik parserlar. Bu kabi parser turlarini ishga tushirish uchun bir nechta inson omili bilan bog‘liq jarayonlar talab etiladi. Jumladan, katta hajmdagi ma’lumotlar to‘plamidan foydalanib, dasturga muayyan matnni to‘g‘ri tahlil qilish o‘rgatiladi hamda eng ehtimoliy modellar asosida tahlil qila boshlaydi. Bu usulda yaratilgan tahlilchi dasturlar bir necha million misollarni tahlil qilib, ma’lum grammatik strukturalarni o‘rganib chiqadi.

Statik parserlar avvaldan belgilangan grammatik qoidalarga asoslangan bo‘lib, ayni matnni aniq sintaktik tuzilishga ajratadi. Bunday tahlil dasturlarida barcha grammatik qoidalar oldindan belgilanadi.

4. Dependency (bog‘liqlik) parserlarda har bir so‘z boshqa biror so‘zga bog‘lanadi va bu bog‘lanish tizimi biror “root” so‘zi atrofida tashkil topadi. Root so‘zi – gapdagi asosiy so‘z bo‘lib, barcha boshqa so‘zlar shu asosda bog‘lanadi. Masalan, “Men kitob o‘qidim” degan jumlada “o‘qidim” – root so‘z bo‘lib, “men”

va “kitob” so‘zlari unga bog‘langan. Natijada bog‘lanishlar yo‘naltirilgan chiziqlar shaklida ifodalanadi. Har bir bog‘lanish bir so‘zni boshqasiga bog‘laydi va bu bog‘lanish yo‘naltirilgan bo‘ladi. Ya’ni, bir so‘z ikkinchisiga bog‘lanib, uni boshqaradi. Bunday munosabatlar grammatik jihatdan gapning to‘g‘riligini tekshiradi. Bunday tahlil turida gapning strukturasini tasvirlash uchun turli diagrammalar (dependency trees) ishlatiladi. Har bir so‘z node (tugun) sifatida ifodalanadi va ular o‘rtasidagi bog‘lanishlar yo‘naltirilgan chiziqlar bilan ko‘rsatiladi. Diagrammalarni ko‘rish orqali tilshunoslar va kompyuter tizimlari grammatik tuzilmalarni osongina tahlil qilishi mumkin(6-rasm).



6-rasm. Dependancy parsing natijasi

Xulosa qilib aytganda, Dependency parsing (bog‘lanish tahlili) va, umuman, parser dasturlari tabiiy tilni qayta ishlash (NLP) tizimlarining, masalan, savol-javob tizimlari yoki mashina tarjimasi kabi ko‘plab sohalarida muhim bosqich hisoblanadi. o‘zbek tili morfologik jihatdan boy va grammatik jihatdan murakkab til bo‘lib, uni avtomatik tarzda qayta ishlashni yanada qiyinlashtiradi. O‘zbek tiliga oid NLP vositalari va resurslarining cheklanganligi bu vazifani yanada murakkablashtiradi. Qoidalarga asoslangan hamda ma’lumotlarga asoslangan chuqur o‘rganish (deep learning) modellari parserlarni ishlab chiqish va amaliyotda qo‘llashda natijalarni ko‘rsatmoqda. Biroq, chuqur o‘rganishga asoslangan tizimlar yaxshi natijalarga erishish uchun katta hajmdagi ma’lumotlarga muhtoj bo‘ladi. Bu boradagi tadqiqotlar chuqur o‘rganilishiga ehtiyoj mavjudligini ko‘rsatadi.

Foydalanilgan adabiyotlar:

1. Abdurahmonova N., Urdishev K. Corpus based teaching Uzbek as a foreign language. / Journal of Foreign Language Teaching and Apllied Linguistics (J-FLTAL). 2019.
2. Salvatella M.L. Parsing ad Evaluation. Improving dependency grammars accuracy. PhD. diss – Universitat de Barcelona, 2016.
3. Ruket Cakici. Wide Coverage parsing for Turkish. –M., 2009.

4. Collins Michael, James Brooks. Prepositional attachment through a backed-off model. –Cambridge., 1995.
5. Foth K., Menzel W. Learning the contraints weights of a dependency grammar using genetic algorithms // 3th International conference on Domain Decomposition Methods, 2001. –P 243-247.
6. Abjalova M. Korpus lingvistikasi/uslubiy qo‘llanma. –Toshkent, 2022.
7. https://www.geeksforgeeks.org/types-of-parsers-in-compiler-design/

AVTOMATIK IMLO TUZATISH BO‘YICHA TADQIQOTLAR TAHLILI: TAVSIF VA MUAMMOLARI

Ahmedova Maftuna Kaxramon qizi,
magistrant
maftunaAhmedova1997@gmail.com
ToshDO‘TAU

Annotatsiya. Matndagi imloni avtomatik to‘g‘rilashga qaratilgan tadqiqotlar uchta bosqichma-bosqich murakkablashib boruvchi muammoga e’tibor qaratgan: so‘z bo‘lmagan xatoliklarni aniqlash; so‘z xatolarini tuzatish; kontekstga bog‘liq so‘zlarni to‘g‘rilash. Birinchi muammoda berilgan so‘zlar ro‘yxatida uchramaydigan qatorlarni aniqlash uchun samarali andoza moslashtirish (pattern-matching) va n-gramm tahlil usullari ishlab chiqilgan. Ikkinchi muammoda umumiy va maxsus ilovalarga mo‘ljallangan turli xil imlo tuzatish usullari ishlab chiqilgan bo‘lib, ularning ayrimlari imlo xatolari qoliplarini batafsil o‘rganishga asoslangan. Uchinchisida esa tabiiy tilni qayta ishlash vositalari yoki statistik til modellaridan foydalangan holda bir nechta tajribalar o‘tkazilgan.

Ushbu maqolada imlo xatolari qoliplari bo‘yicha tadqiqot natijalari tahlil qilindi, matndagi avtomatik xatolarni tuzatishning barcha yo‘nalishi bilan bog‘liq tadqiqot masalalari muhokama qilindi.

Abstract. Research on automatic spelling correction in text has focused on three progressively more complex problems: detecting non-word errors; correcting word errors; and correcting words in context. In the first problem, efficient pattern-matching and n-gram analysis methods were developed to detect strings that do not appear in a given list of words. In the second problem, various spelling correction methods were developed for general and specific applications, some of which are based on a detailed study of spelling error patterns. In the third, several experiments were conducted using natural language processing tools or statistical language models.

This article analyzes the results of research on spelling error patterns and discusses research issues related to all areas of automatic text correction.

Аннотация. Исследования в области автоматического исправления орфографии в тексте были сосредоточены на трех постепенно усложняющихся проблемах: обнаружение несловесных ошибок; исправление орфографических ошибок; Исправление слов в зависимости от контекста. В первой задаче были разработаны эффективные методы сопоставления с образцом и анализа n-грамм для выявления строк, которые не встречаются в заданном списке слов. Во второй задаче были разработаны различные методы исправления орфографии для общего и частного применения, некоторые из

которых основаны на детальном изучении закономерностей орфографических ошибок. Третий этап включал в себя несколько экспериментов с использованием инструментов обработки естественного языка или статистических языковых моделей.

Kalit so‘zlar: *imlo qoidalari, noleksik birliklar, avtomatik imlo tuzatish, tinish belgilari.*

Imlo tekshiruvchi va tuzatuvchi dasturlar – bu berilgan matndagi grammatik va kontekstual imlo xatolarini aniqlaydigan va ularni maxsus algoritm yoki qoidalar to‘plami asosida to‘g‘rilaydigan dasturiy ilovadir. Ba’zi imlo tekshiruvchi dasturlar noto‘g‘ri yozilgan so‘zlar uchun to‘g‘ri muqobil variantlar ro‘yxatini yoki so‘zlar ketma-ketligi bo‘yicha takliflarni taqdim etadi. Avtomatik imlo tekshirish aksariyat tillarda so‘z protsessorlari (matn muharrirlari) uchun keng tarqalgan xususiyat hisoblanadi. Shuningdek, deyarli barcha veb-brauzerlarda ichki imlo tekshirgichlari mavjud. Avtomatik imlo tuzatish sohasi so‘nggi yillarda ilmiy tadqiqot e’tiborini tortib kelmoqda. Tadqiqot ishlarida ushbu mavzuga oid qisqacha kirish qismi mavjud bo‘lsa-da, hozirgacha nazariy asoslarni umumlashtiruvchi va shu kungacha ishlab chiqilgan yondashuvlarga umumiy ko‘rinish beruvchi tadqiqot yetishmaydi.

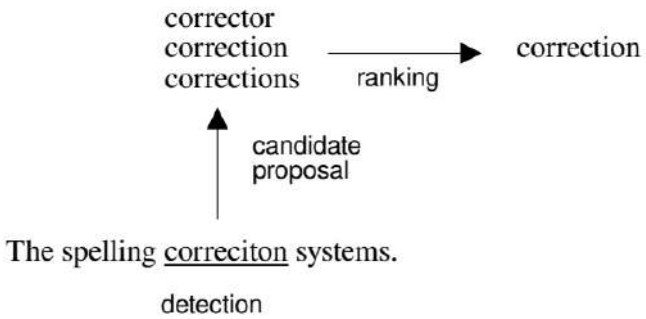
O‘zbek tili kam resursga ega bo‘lgan tillardan biri hisoblanib, unga oid tadqiqotlar, ma’lumotlar va vositalar hali rivojlanish bosqichida. Ma’lumotlar bazasida katta hajmdagi matnlarning yetishmasligi yangi muammolarni keltirib chiqaradi. Avtomatik imlo tuzatish tizimi tabiiy tilni qayta ishlash tizimining (NLP) bir qismi hisoblanadi. Matnlarni interaktiv tarzda tuzatish juda qiyin bo‘lganligi sababli, ma’lumotlar bazasidagi matnlar avtomatik ravishda to‘g‘rilanishi kerak. Imlo tuzatish tizimi oldingi va keyingi matn kontekstiga asoslanib, eng mos tuzatish variantini tanlaydi. Imlo tekshiruvchi dasturlar bo‘yicha tadqiqotlar 1950-yillarning oxirlariga borib taqaladi va hozirda ular ingliz, nemis va xitoy tillari kabi ko‘plab tillar uchun yaxshi rivojlangan. 1992-yilda Kukich[1]tomonidan taqdim etilgan izlanish keng qamrovli tadqiqotlarni boshlab berdi. Imlo xatolarini avtomatik tuzatish bo‘yicha ko‘p tadqiqotlar mavjud. Mitton tomonidan yozilgan kitob imlo xatolari turlarini tahlil qilib, avtomatik imlo tuzatish tizimini yaratish yondashuvlarini tasvirlab bergan [2]. Yannakoudakis va Favtroplar aksariyat imlo xatolari fonologik va ketma-ketlikka asoslangan maxsus qoidalarga bo‘ysunishini yozgan. Bu mualliflar maqolasida imlo xatolarini uchta asosiy kategoriya bo‘yicha tasniflab, 1377 ta imlo xatosi shakllarining tahlil natijalarini taqdim etgan[3]:

1. Undosh harflarga oid xatolar
2. Unli harflarga oid xatolar
3. Ketma-ketlik xatolari

Ba’zi yozuv tizimlari: arab, vetnam yoki slovak kabi yozuvlarda turli xil harf variantlaridan foydalanadi, bu esa so‘zning ma’nosini o‘zgartirishi mumkin. Gimenes va Roman braziliya-portugal tilida diakritik belgilarning tushirib

qoldirilishi imlo xatolarining keng tarqalgan turi ekanligini tasdiqlagan[4]. Zamonaviy adabiy arab tilida matnlar odatda diakritik belgilarsiz yoziladi[5]. Bu tipografik xatolik bo‘lib, muallif qo‘shimcha belgilardan foydalanmaydi va o‘quvchi asl ma’noni o‘zi tushunishini kutadi. Yo‘qolgan diakritik belgilar odatda qisqa unli tovushlarni yoki harflarning o‘zgarishini bildiradi. Ular harflarning tepasiga yoki tagiga joylashtiriladi. Arab matnlariga unli tovushlarni va boshqa diakritik belgilarni qo‘shish jarayoni diakritizatsiya yoki vokalizatsiya deb ataladi[6]. Azmi va Almajed arab tilidagi diakritizatsiya muammosiga e’tibor qaratib, uning baholash mezonini taklif qilgan[7]. Asahiah va boshqalar[8] esa arab tilidagi diakritizatsiya texnikalari bo‘yicha sharh nashr etgan.

Avtomatik imlo tuzatish tizimi imlo xatolarini aniqlaydi va tuzatish uchun nomzod variantlar to‘plamini taklif qiladi (1-rasmga qarang).



1-rasm. Avtomatik imlo tuzatish jarayoni[9]

So‘z yangi yoki shunchaki kam uchraydigan, kam tanilgan nom (NER) bo‘lishi yoki boshqa tilga tegishli leksema bo‘lishi mumkin. Biroq to‘g‘ri yozilgan so‘z gapda semantik jihatdan noto‘g‘ri bo‘lishi ham ehtimoldan xoli emas. Avtomatik imlo tuzatish tizimini ishlab chiqish uchun imlo xatosining paydo bo‘lish jarayonini tushunish zarur. Kukich imlo xatolarini to‘g‘ri so‘zlar lug‘atiga asoslanib quyidagi guruhlarga ajratgan:

Haqiqiy so‘z xatolari – so‘z noto‘g‘ri yozilgan bo‘lsa-da, uning to‘g‘ri shakli so‘zlar lug‘atida mavjud bo‘ladi.

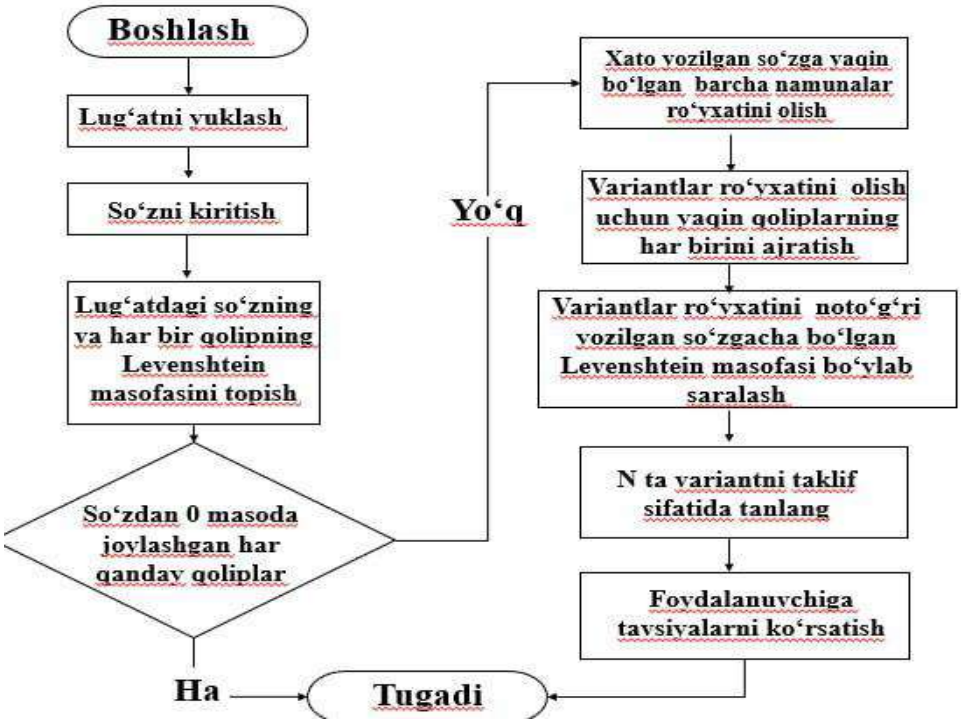
So‘z bo‘lmagan xatolar – noto‘g‘ri yozilgan so‘z shakli to‘g‘ri so‘zlar lug‘atida mavjud emas.

Ko‘pgina imlo tuzatish tizimlari so‘z bo‘lmagan xatolarni aniqlash uchun so‘zni to‘g‘ri so‘zlar lug‘atidan qidiradi. Bu bosqich so‘zlarni qidirish va tahlil qilish uchun mo‘ljallangan hash jadvali (hash table) yoki qidiruv daraxti kabi tezkor qidiruv usullarini talab qiladi. Ko‘pgina so‘z bo‘lmagan xatolarni tuzatish tizimlari ochiq kodli oldindan belgilangan imlo tuzatish tizimlaridan foydalanadi, masalan, **Aspell** yoki **Hunspell** xatolarni aniqlash, tuzatish nomzodlarini yaratish va dastlabki nomzodlarni saralash uchun qo‘llaniladi.

Kukich[10]hamda Pirinen va Linden[11] ushbu jarayonni uch bosqichga ajratgan:

1. Xatoni aniqlash
2. Tuzatish variantlarini yaratish
3. Tuzatish variantlarini tartiblash

Tadqiqotlarda avtomatik imlo tekshiruvchi dasturlarning ishlash tizimining Blok-sxema algoritm bosqichlarining umumiy ko‘rinishlari taklif qilingan. 2-rasmda taklif qilingan imlo tekshirish algoritmining blok-sxemasini ko‘rish mumkin.



2-rasm. Taklif qilinadigan imlo tekshirish algoritmining blok-sxemasi

Ba’zi tadqiqotlarda Avtomatik imlo tekshiruvchi dasturlarning noto‘g‘ri birlikni aniqlash va to‘g‘ri variantini taklif etish tizimi farqli ekanligini uchratish mumkin. To‘g‘ri tuzilgan ma’lumotlar bazasi, matnlar resursi yetarli bo‘lmagan tillar uchun xatolari oldindan belgilangan qoidalar va modellar asosida aniqlaydigan va tuzatadigan tizimlar taklif qilinadi. Bu tizim **A priori imlo tuzatish tizimi** deb nomlanadi. Qoidalarga asoslangan bu tizim – **Lug‘at asosida ishlaydi**, ya’ni so‘zlarning to‘g‘ri yoki noto‘g‘ri ekanligini oldindan tuzilgan lug‘atga qarab aniqlaydi, shuningdek, qoidalar asosida xatolarni tuzatadi: tilga xos grammatik va orfografik qoidalar asosida tuzatishlar kiritadi. Aynan mana shu tizim o‘zbek tilida ham imloni tuzatuvchi dasturlarning ishlab chiqilishida ilk bosqich hisoblanadi. Ammo qoidalarga asoslangan modellarda lug‘at bazasi cheklanganligi va so‘zni individual tahlil qilganligi sababli kontekstga ko‘ra tuzatishda xatoliklarga sabab bo‘ladi. Dastlab Google yoki Microsoft Word ning dastlabki imlo tekshiruvchilari ham lug‘at asosida ishlagan, lekin hozirgi zamonaviy tizimlar kontekstga asoslangan neyron tarmoqlar yordamida yanada rivojlangan.

Xulosa sifatida shuni aytish mumkinki, avtomatik imlo tuzatish ham mashina tarjimasiga o‘xshash jarayon hisoblanadi. Xatolarni o‘z ichiga olgan matn eng ehtimoliy to‘g‘ri shaklga “tarjima qilinadi”. Ushbu yondashuv butun natijaviy jumlani hisobga oladi. Bugungi kunda ko‘pchilik norasmiy xat va hujjatlardagi imloviy xatolarni tekshirish uchun ko‘p vaqt sarflaydi. Imloni tekshiruvchi tizimlarning yaratilishi esa barcha shu xizmatlarni yagona kutubxona sifatida taqdim etadi, shuningdek, foydalanuvchilar uni o‘z dasturlariga qo‘shishlari mumkin.

Bu quyidagi afzalliklarni ta’minlaydi:

1. **Vaqtni tejash:** Taklif etilgan algoritm kompyuterda bir soniyada minglab so‘zlarning imlosini tekshirishi va har soniyada o‘nlab so‘zlar uchun tuzatish tavsiyalarini berishi mumkin.
2. **Aniqlik:** Algoritm 70-80% hollarda so‘zni to‘g‘ri yozilishini birinchi tavsiya sifatida taklif qiladi, 90% hollarda esa to‘g‘ri yozilgan so‘zning variant yuqori beshlik ichida bo‘ladi.
3. **Xarajatlarni tejash:** Imlo tekshirish jarayonini avtomatlashtirish natijasida kompaniyalar,nashiryot va davlat tashkilotlari qo‘lyozmalarni tekshiruvchi xodimlar uchun xarajatlarini kamaytirishi mumkin.

Foydalanilgan adabiyotlar:

1. Kukich, K. Techniques for automatically correcting words in text. Acm Comput. Surv. 1992, 24, 377–439.
2. Mitton, R. English Spelling and the Computer; Longman Group: Harlow, Essex, UK, 1996; p. 214
3. Yannakoudakis, E.J.; Fawthrop, D. The rules of spelling errors. Inf. Process. Manag. 1983, 19, 87–99
4. Gimenes, P.A.; Roman, N.T. Spelling error patterns in Brazilian Portuguese. Comput. Linguist. 2015, 41, 175–184
5. Zitouni, I.; Sarikaya, R. Arabic diacritic restoration approach based on maximum entropy models. Comput. Speech Lang. 2009, 23, 257–276.
6. Zitouni, I.; Sarikaya, R. Arabic diacritic restoration approach based on maximum entropy models. Comput. Speech Lang. 2009, 23, 257–276.
7. Azmi, A.M.; Almajed, R.S. A survey of automatic Arabic diacritization techniques. Nat. Lang. Eng. 2015, 21, 477–495.
8. Asahiah, F.O.; Odéjobi, O.A.; Adagunodo, E.R. A survey of diacritic restoration in abjad and alphabet writing systems. Nat. Lang. Eng. 2018, 24, 123–154
9. Daniel Hládek, Ján Staš, Matúš Pleva. Survey of automatic Spelling correction. Electronics 2020, 9, 1670; doi:10.3390/electronics9101670.
10. Kukich, K. Techniques for automatically correcting words in text. Acm Comput. Surv. 1992, 24, 377–439.

11. Pirinen, T.A.; Lindén, K. State-of-the-art in weighted finite-state spell-checking. In Computational Linguistics and Intelligent Text Processing, Proceedings of the CICLing 2014, Kathmandu, Nepal, 6–12 April 2014; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8404, Part 2, pp. 519–532.

O‘ZBEK TILI MATNLARIDA TINISH BELGILARNI AVTOMATIK QO‘YISH

Matyakubova Noila Shakirjanovna,
Tayanch doktorant
matyakubovanoila@navoiy-uni.uz
ToshDO‘TAU

Avloqulova Iqbol,
Kompyuter lingvistikasi yo‘nalishi 3-bosqich talabasi
iqbolaavloqulova2@gmail.com
ToshDO‘TAU

Annotatsiya. Tinish belgilarini avtomatik tarzda tiklash – matnning mazmunini to‘g‘ri anglash, gaplar chegarasini aniqlash va nutqni o‘qish jarayonini yengillashtirishda muhim ahamiyat kasb etadi. Ayniqsa, nutqni matnga aylantirish, ijtimoiy tarmoqlardagi punktuatsiyasiz yozuvlar va raqamlashtirilayotgan hujjatlar bilan ishlashda tinish belgilarining avtomatik aniqlanishi tabiiy tilni qayta ishlash (NLP) sohasining dolzarb yo‘nalishlaridan biri hisoblanadi. Ushbu maqolada o‘zbek tili matnlarida tinish belgilarni avtomatik tiklash muammosi chuqur yoritilib, mavjud usullar tahlil qilingan. Xususan, transformer arxitekturasiga asoslangan BERT va ELECTRA modellarining samaradorligi, shuningdek, statistik va qoidaviy yondashuvlarning afzalliklari ko‘rib chiqilgan. Tadqiqotda modal so‘zlar va so‘roq yuklamalari ishtirokidagi gaplar asosida Python dasturlash tilida amaliy tajriba o‘tkazilib, NLP jarayonida tilning semantik va sintaktik xususiyatlarini hisobga olgan holda tinish belgilarini aniqlashga qaratilgan fikrlar ilg‘or surilgan. Mazkur yondashuv natijalari esa kelgusidagi tadqiqotlar uchun mustahkam asos bo‘lib xizmat qilishi mumkin.

Аннотация. Автоматическое восстановление знаков препинания имеет важное значение для обеспечения точного понимания содержания текста, определения границ предложений и упрощения процесса чтения. Особенно в случаях преобразования речи в текст, пунктуационно небрежных записей в социальных сетях и работы с оцифровываемыми документами автоматическое определение знаков препинания стало одним из актуальных направлений в области обработки естественного языка (NLP). В данной статье глубоко анализируется проблема автоматического восстановления знаков препинания в текстах на узбекском языке, а также рассматриваются существующие методы. В частности, оценивается эффективность моделей на основе трансформеров, таких как BERT и ELECTRA, а также преимущества статистических и правилосообразных подходов. В исследовании проводились практические эксперименты на языке Python, сосредоточенные на

предложениях с участием модальных слов и вопросительных суффиксов, и предлагались передовые идеи по определению знаков препинания с учетом семантических и синтаксических особенностей языка в процессах NLP. Результаты данного подхода могут послужить прочной основой для будущих исследований.

Annotation. Automatic punctuation restoration plays a crucial role in ensuring accurate understanding of text content, identifying sentence boundaries, and simplifying the reading process. Particularly in speech-to-text conversion, punctuation-less writings on social media, and the digitization of documents, the automatic detection of punctuation marks has become one of the pressing directions in the field of natural language processing (NLP). This article deeply examines the issue of automatic punctuation restoration in Uzbek language texts and analyzes existing methods. Specifically, it evaluates the effectiveness of transformer-based models such as BERT and ELECTRA, as well as the advantages of statistical and rule-based approaches. The study includes practical experiments conducted in Python, focusing on sentences involving modal words and interrogative suffixes, and proposes advanced insights for identifying punctuation marks by considering the semantic and syntactic features of the language in NLP processes. The results of this approach can serve as a solid foundation for future research.

Kalit so‘zlar: *Tinish belgilari, tinish belgilarini avtomatik tarzda qo‘yish (APR), so‘roq belgisi, vergul, modal so‘zlar, so‘roq yuklamasi.*

Hozirgi axborot texnologiyalari rivojlangan davrda matnlarni avtomatik qayta ishlash tobora muhim ahamiyat kasb etmoqda. Ayniqsa, o‘zbek tili kabi milliy tillardagi matnlarni qayta ishlash, tahlil qilish va tuzatish tizimlarini yaratish dolzarb vazifalardan biri hisoblanadi. Matnlarni qayta ishlashning muhim bosqichlaridan biri tinish belgilarini avtomatik tiklash jarayonidir. Tinish belgilari matn ma’nosini to‘g‘ri ifodalashda muhim rol o‘ynaydi. Ular jumlalar o‘rtasidagi chegaralarni belgilaydi, matn qismlarini mantiqan ajratadi va fikrning aniq ifoda etilishini ta’minlaydi.

Bugungi kunda ijtimoiy tarmoqlar, messenjerlar va boshqa raqamli platformalarda yozilgan ko‘plab matnlar tinish belgilarisiz yoki noto‘g‘ri tinish belgilari bilan yaratilmoqda. Bundan tashqari, og‘zaki nutqni yozma ko‘rinishga o‘tkazishda, tarixiy hujjatlarni raqamlashtirishda va katta hajmdagi matnlarni qayta ishlashda tinish belgilarini avtomatik tiklash muhim masalalardan hisoblanadi.

Tinish belgilari tabiiy tilni qayta ishlashda (NLP) muhim rol o‘ynaydi, chunki ular matnning sintaktik va semantik tuzilishini belgilashga yordam beradi. Ko‘pchilik NLP ilovalari, masalan, nutqni matnga aylantirish tizimlari, asl nutq tarkibida tinish belgilarini saqlamaydi. Shu sababli, matnni keyinchalik tushunish va qayta ishlash uchun tinish belgilari avtomatik tarzda tiklanishi kerak.

Ushbu maqolada o‘zbek tili matnlarida tinish belgilarni avtomatik tiklash muammosining dolzarbligi, mavjud yechimlar va ularning samaradorligi, hamda kelajakda ushbu yo‘nalishda olib borilishi kerak bo‘lgan ilmiy-tadqiqot ishlari haqida so‘z yuritiladi.

So‘nggi yillarda chuqur o‘rganish (Deep Learning) usullari yordamida Avtomatik Nutq Tanib Olish (ASR) tizimlarining aniqligi sezilarli darajada oshdi. O‘zbek tilidagi matnlarda tinish belgilarni tiklash masalasi bugungi kunda zamonaviy texnologiyalar va sun’iy intellekt yordamida ko‘plab tadqiqotlar obyektiga aylangan. Ushbu sohada xalqaro tajribalar, xususan, ingliz va venger tillari uchun ishlatilgan usullar orqali samarali natijalarga erishilgan. BERT modellari asosida avtomatik tinish belgilari tiklash usullari Attila Nagy va boshqa tadqiqotchilar tomonidan taqdim etilgan. Ularning tadqiqoti ingliz tili uchun Ted Talks[1] ma’lumotlar bazasi va venger tili uchun Szeged Treebank ma’lumotlar bazasida o‘tkazilgan bo‘lib, samaradorlik 79.8% (ingliz tili) va 82.2% (venger tili) ko‘rsatkichlariga erishilgan. Courtland va boshqalar transfer arxitekturasini ishlatgan holda BERT modellari ko‘p so‘zli bashoratlar yordamida aniqlikni oshirish usullarini taklif qilgan [1,2]. Ushbu usullar o‘zbek tilida ham grammatik qoidalarga asoslangan katta ma’lumotlar bazasini shakllantirish orqali qo‘llanilishi mumkin.

Shuningdek, Viet Dac Lai va tadqiqotchilar tadqiqotlarida livestream videolari transkriptlarida tinish belgilari tiklashning o‘ziga xos qiyinchiliklari ta’kidlangan[3]. Ular Behance platformasidan olingan videolar uchun yangi annotatsiyalangan korpus (BehancePR)ni taqdim etganlar, bu esa spontan nutq, ko‘p sonli so‘z xatolari va hissiy ifodalar (masalan, undov belgisi) kabi xususiyatlarni o‘z ichiga oladi.

O‘zbek tilidagi matnlarda tinish belgilari tiklash masalasini o‘rganishda zamonaviy sun’iy intellekt texnologiyalari muhim rol o‘ynaydi. Martin Polacek va tadqiqotchilarning 2023-yilda nashr etilgan ishida avtomatik tinish belgilari tiklash (APR– Automatic Punctuation Restoration) yondashuv taklif qilingan, bu esa faqat matnli kiritishdan foydalanadi va ELECTRA-Small modeli asosida ishlaydi. Ularning tadqiqoti real vaqt rejimida televidenya va radio transkriptlarida tinish belgilari (nuqta, vergul, so‘roq belgisi) tiklashda o‘rtacha 71.2% va 69.4% natijalarini ko‘rsatgan [4]. Bu ish o‘zbek tilida ham real vaqtli transkripsiyada tinish belgilari tiklash uchun samarali algoritmlarni ishlatish imkoniyatini ko‘rib chiqishga asos berishi mumkin. Ushbu tadqiqot shuni ko‘rsatadiki, transformator modellari (masalan, ELECTRA) va kichik kontekstli kiritish bilan past kechikish (uch so‘z) va yuqori aniqlikni ta’minlash mumkin. Bu o‘zbek tilidagi matnlarda tinish belgilari tiklashda yangi yechimlar izlash uchun muhim manba hisoblanadi.

Adebayo Mustapha Bakare va boshqa tadqiqotchilar (2023)[5] ijtimoiy tarmoqlardagi matnlarda tinish belgilarni tiklash uchun transformator modeliga

asoslangan algoritm taklif qilishgan. Ular Amazon (34,659 sharh) va Telekom Malaysia (5000 sharh) ma’lumotlaridan foydalanib, turli BERT modellarini sinovdan o‘tkazishgan. Natijada, RoBERTa-Large modeli ikki yo‘nalishli LSTM bilan birgalikda eng yuqori aniqlikni ko‘rsatgan: Amazon uchun 97% va Telekom uchun 90%. Bu tadqiqot ijtimoiy tarmoqlardagi noto‘g‘ri punktatsiyali matnlarni tahlil qilishda tinish belgilarni tiklashning sentiment tahliliga foydasini ta’kidlaydi . O‘zbek tilidagi ijtimoiy media matnlari uchun ham ushbu yondashuv moslashtirilishi mumkinligi ko‘rsatiladi.

Tinish belgilarini avtomatik tiklash modellari.

Nutqni avtomatik matnga aylantirish (ASR – Automatic Speech Recognition) tizimlarida tinish belgilarini tiklash (punctuation restoration) – bugungi kunda tabiiy tilni qayta ishlash sohasining dolzarb va chuqur o‘rganilayotgan masalalaridan biridir. Gap tuzilmasining aniqligini ta’minlash, matnni o‘qish va tushunishni osonlashtirish va hatto keyingi bosqichdagi mashinali tarjima yoki matn tahlilini aniq bajarish uchun tinish belgilarining to‘g‘ri qo‘yilishi muhim ahamiyatga ega.

Transformer modellar asosidagi ilg‘or yondashuvlar. Tinish belgilarini tiklashda transformer arxitekturasiga asoslangan zamonaviy modellar – BERT va ELECTRA samaradorligi solishtirilib tahlil qilingan[4]. Taklif etilgan tizim, ayniqsa, real vaqt rejimidagi nutq transkripsiyasi uchun moslashtirilgan bo‘lib, matn asosida ishlovchi (text-only) ELECTRA-Small modeli 11 millisekund kechikish bilan ishlashga muvaffaq bo‘lgan. U model nuqta (F1 = 73.8), vergul (F1 = 79.2) va so‘roq belgisi (F1 = 42.1)larni aniqlashda yuqori natijalarga erishgan [4]. Bunday real vaqt yondashuvlar ayniqsa jonli subtitrlar va translyatsiyalar uchun dolzarb hisoblanadi.

Statistika va n-gram modellarga asoslangan yondashuvlar. Gravano, Jansche va Bacchiani (2009) tomonidan olib borilgan tadqiqotda esa katta hajmdagi korpuslardan foydalangan holda statistik yondashuvlar orqali punktuatsiyani tiklashga urinishlar amalga oshirilgan. Xususan, 58 milliondan[6] ortiq tokenlar asosida qurilgan n-gram modellar yordamida vergul va nuqta singari belgilarni muvaffaqiyatli aniqlashga erishilgan. Biroq kam uchraydigan belgilar – so‘roq belgisi, tire va nuqtali vergul singari tinish belgilar bu yondashuvlarda nisbatan past aniqlik bilan tiklangan.

O‘zbek tilidagi matnlarda APRning qo‘llanilishi.

Tinish belgilari o‘zbek tilida yozma nutqning asosiy tarkibiy qismi bo‘lib, ularning qo‘llanilishi nutqning semantik, grammatik xususiyatlariga qat’iy bog‘liq. Fitratning “O‘zbek tili qoidalari to‘g‘risida bir tajriba” asariga asoslanib, tinish belgilari “turish belgilari” sifatida tasniflanadi va ularning ma’nosi ularning vazifasi, yozuvdagi o‘rni, shuningdek, gapning sintaktik tuzilishi va ma’no yukiga bog‘liq[7].

Masalan, nuqta jumlani tugatishni, vergul qo‘shimcha ma’lumotlarni ajratishni, so‘roq va undov belgilari esa gapning ohang va maqsadini ifodalaydi.

Grammatik jihatdan bir xil shakllangan va bir xil leksik vositalarga ega bo‘lgan quyidagi besh gapni olib ko‘raylik, bularda besh xil ma’no birligi tinish belgilaridan bilinadi. Masalan:

1. Biz fazoni o‘rganmoqdamiz. (Xabar ma’nosi)
2. Biz fazoni o‘rganmoqdamiz? (So‘roq ma’nosi)
3. Biz fazoni o‘rganmoqdamiz! (Undov ma’nosi)
4. Biz fazoni o‘rganmoqdamiz... (fikrning bo‘linishi)
5. Biz fazoni o‘rganmoqdamiz!.. (Kuchli emotsiya – sevinch – fikrning bo‘linishi ma’nosi)[7].

Yuqoridagi misollar shuni ko‘rsatadiki, tinish belgilarini avtomatik tiklash nafaqat matnni grammatik jihatdan to‘g‘ri tuzishda, balki uning mazmunan to‘liq anglanishida ham muhim rol o‘ynaydi. Ayniqsa, o‘zbek tilida bir xil grammatik tuzilishga ega gaplar turli tinish belgilar yordamida turlicha ma’nolarni ifodalashi mumkin. Bu esa avtomatik tizimlar uchun jiddiy muammo bo‘lib, tinish belgilarini to‘g‘ri joylashtirish orqali matnning semantik va pragmatik ma’nosini tiklash zarurligini ko‘rsatadi. Yuqorida keltirilgan misollar, ya’ni “Biz fazoni o‘rganmoqdamiz” gapining har xil tinish belgilar bilan ifodalanishi (xabar, so‘roq, undov, noaniqlik va emotsional kuchlanish) — bu jarayonning nozikligini va sun’iy intellektga asoslangan tizimlarda ehtiyotkorlik bilan yondashish lozimligini anglatadi. Shuning uchun ham avtomatik tinish belgilarini tiklash modellarini o‘zbek tilining intonatsiya va sintaktik xususiyatlariga moslashtirish dolzarb masalalardan biri hisoblanadi.

Shu bilan bir qatorda matndagi xatoliklarni tuzatish muammolari orasida tinish belgilarni qo‘yish ham alohida e’tiborni talab qiladi. Tinish belgilar bo‘yicha ikki turdagi masala ko‘riladi. Birinchisi, matnda mavjud tinish belgilarni to‘g‘ri qo‘yilganligini tekshirish. Ikkinchisi esa, matnda tinish belgilari umuman mavjud emas vaziyatlarda ularni to‘g‘ri qo‘yish [8].

Ushbu maqolada matnda mavjud bo‘lmagan tinish belgilarni tiklash masalasiga alohida e’tibor qaratilgan. Xususan, modal so‘zlar va so‘roq yuklamalarining matndagi vazifasi, ularning gap ma’nosi va ohangiga ta’siri nuqtayi nazaridan tahlil qilingan. Ushbu birliklarning grammatik va semantik xususiyatlariga tayangan holda, tinsh belgilarini aniqlashda qanday xizmat qilishi amaliy jihatdan o‘rganilgan.

Amaliy ish jarayonida modal so‘zlar va so‘roq yuklamalari ishtirokidagi gaplarni tahlil qilish hamda ularni tegishli tinish belgilar bilan to‘ldirishga oid

tajribalar o‘tkazilgan. Ushbu yondashuv tinish belgilarni avtomatik tiklashda lingvistik yondashuvning muhimligini ko‘rsatishga xizmat qilish mumkin.

Quyidagi jadvalda **modal so‘zlar va so‘roq yuklamalari ishtirokidagi gaplar namunalarini jadvallar shaklida ko‘rsatiladi.** Jadvalda ba’zi modal so‘roq yuklamalarining matndagi vazifasi, unga mos tinish belgisi ko‘rsatiladi (Qarang:1 va 2-jadvalga).

1-jadval. Modal so‘zlar va ularning turlari.

Modal so‘z turlari	Modal so‘zlar	Misollar
Rostlik, tasdiq-ishonch, qat’iylikni bildiruvchi so‘zlar	darhaqiqat, haqiqatan, albatta, shubhasiz, so‘zsiz, shaksiz.	Darhaqiqat, bu holat ko‘p kuzatiladi.
Taxmin, gumon kabi ma’nolarni ifodalovchi so‘zlar	shekilli, chog‘i, chamasi, ehtimol, balki, aftidan, taxminan.	Bu ishni tugatish uchun, taxminan, qancha vaqt ketadi?
Fikrning o‘zaro munosabati	xullas, demak, binobarin, qisqasi, umuman, masalan, chunonchi, jumladan, aksincha, avvalo, avvalambor, nihoyat.	Men ko‘plab mevalarni yaxshi ko‘raman, masalan, olma, banan va uzum.
Ta’kid	xususan, ayniqsa.	Xususan, ingliz tili o‘quvchilarga eng ko‘p qiyinchilik tug‘diradi.

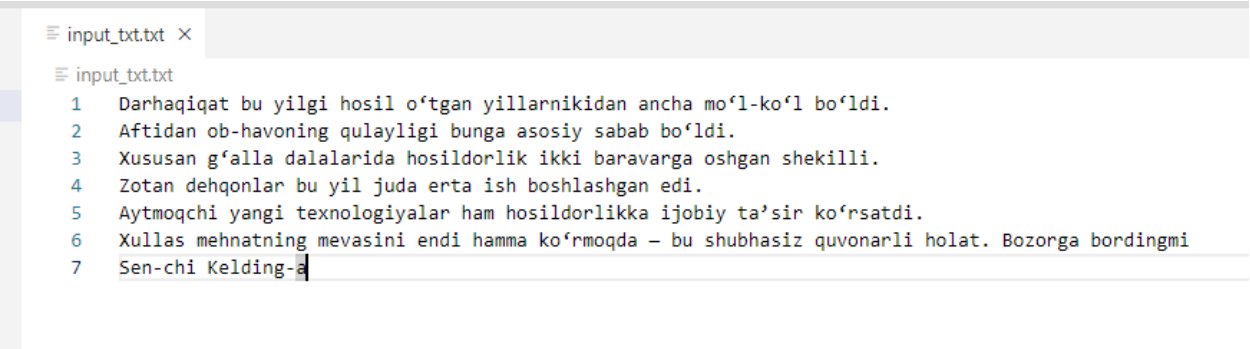
2-jadval. So‘roq yuklamalari.

So‘roq yuklamalari	Misollar
-mi	Ertaga biznikiga kelasan**mi**?
-chi	Men kitobni o‘qib tugatdim, sen-**chi**?
-ya	Nahotki shu gaplar rost bo‘lsa-**ya**?
-a	Bugunoq qaytasan-**a**?

Yuqoridagi jadval asosida tanlab olingan modal so‘zlar va so‘roq yuklamalari ishtirokidagi gaplar amaliy ish jarayonida dasturiy tahlil qilingan. Matnni tahlil qilish va tinish belgilarini avtomatik qo‘shish jarayonida Python dasturlash tilidan va uning *nltk* (Natural Language Toolkit) kutubxonasidan foydalaniladi. Matnlar tokenlarga ajratish va so‘zlarni aniqlashda, shuningdek, tinish belgilarini qo‘yishda *re-kutubxonasi* yordamida muntazam ifodalardan (regular expressions) foydalaniladi. Bu yondashuv yordamida modal so‘zlar va so‘roq yuklamalarini aniqlash va ular bilan ishlash ko‘zda tutilgan.

Amaliy tajriba quyidagi bosqichlarda amalga oshirildi:

1. Matn tayyorlandi va .txt formatida faylga joylashtirildi (qarang: 1-rasmga).



1-rasm. Dasturga matn kiritish jarayoni.

2. Dasturiy tahlil uchun Python kodi yozildi. Ushbu amaliy ishda matnlarni tahlil qilish va tinish belgilarini avtomatik qo‘shish jarayonida Python dasturlash tilidan va uning *nltk* (Natural Language Toolkit) kutubxonasidan foydalanildi. Ushbu kodda matnlarni tokenlarga ajratish uchun *word_tokenize* va *sent_tokenize* funksiyalari hamda *uznatcorpora*, shuningdek, matnni birlashtirish va qayta tiklash uchun *TreebankWordDetokenizer* funksiyasidan foydalanildi. Tahlil jarayonida shuningdek, *re-kutubxonasi* yordamida maxsus tartibga solish va o‘zgarishlar kiritish amallari bajarildi.

Dasturiy tahlil uchun Python kodi:

```
def modal_sozlarga_vergul_qosh(matn):
    gaplar = sent_tokenize(matn)
    yangi_gaplar = []

    for gap in gaplar:
        sozlar = sozlarni_ajrat(gap)
        oxirgi_simvol = gap[-1]
        yangi_sozlar = []
        i = 0
        while i < len(sozlar):
            soz = sozlar[i]
            soz_clean = re.sub(r'[.,!?]', '', soz.lower()).strip()
            is_modal = soz_clean in modal_sozlar
            # print(soz_clean, is_modal)

            if is_modal:
                original_soz = soz
                next_soz = sozlar[i + 1] if i < len(sozlar) - 1 else
None
                prev_soz = yangi_sozlar[-1] if yangi_sozlar else None

                if i == 0:
```

```
                    yangi_sozlar.append(original_soz + ',')
                    i += 1
                    continue

                elif i > 0 and i < len(sozlar) - 1 and next_soz not in
['.', '!', '?']:
                    old_soz = yangi_sozlar.pop() if yangi_sozlar else
''
                    if old_soz:
                        yangi_sozlar.append(old_soz + ' ,')
                    yangi_sozlar.append(original_soz + ',')
                    i += 1
                    continue

                else:
                    if i > 0 and prev_soz and next_soz is None:
                        old_soz = yangi_sozlar.pop()
                        yangi_sozlar.append(old_soz + ', ' +
original_soz )
                    else:
                        yangi_sozlar.append(original_soz)
                    i += 1
                    break
            else:
                yangi_sozlar.append(soz)
            i += 1

        yangi_gap = detokenizer.detokenize(yangi_sozlar)
        if oxirgi_simvol not in ['.', ',', '?', '!', ]:
            oxirgi_simvol = ''
        yangi_gaplar.append(yangi_gap + oxirgi_simvol)
    return '\n'.join(yangi_gaplar)
```

3. Modal so‘zlar va so‘roq yuklamalari aniqlangach, ularning o‘rniga qarab quyidagi qoidalar bo‘yicha mos tinish belgi tiklanadi:

– Modal so‘zlar gap boshida kelgan bo‘lsa, vergul (,) belgisi: ***Xullas**, bu yilgi mavsumda o‘zgarishlar kutilmoqda, shuning uchun tayyor bo‘lishimiz kerak*;

– Modal so‘zlar gap o‘rtasida kelsa, modal so’zning ikkala tomonidan vergul (,) belgisi: *Ilmiy tadqiqotlar, **jumladan**, tajribalar ham bu jarayonni yaxshilashga yordam beradi;*

– Gap oxirida kelsa, modal so‘z oldidan vergul (,) belgisi qo‘yiladi: *Bu yilgi sayohat juda qiziqarli bo‘ladi, **shubhasiz**.*

4. So‘roq yuklamalari gapni oxirida qo‘llanilsan so‘roq (?) belgisi qo‘yiladi.

Misol uchun, *Sen hali kelmading**mi**?*

5. Natijalar.txt formatida fayl ko‘rinishida chiqarib beriladi (qarang: 3-rasmga).

```
output.txt
1  Darhaqiqat, bu yilgi hosil o‘tgan yillarnikidan ancha mo‘l ko‘l bo‘ldi.
2  Aftidan, ob-havoning qulayligi bunga asosiy sabab bo‘ldi.
3  Xususan, g‘alla dalalarida hosildorlik ikki baravarga oshgan, shekilli.
4  Zotan, dehqonlar bu yil juda erta ish boshlashgan edi.
5  Aytmoqchi, yangi texnologiyalar ham hosildorlikka ijobiy ta’sir ko‘rsatdi.
6  Xullas, mehnatning mevasini endi hamma ko‘rmoqda bu, shubhasiz, quvonarli holat.
7  Bozorga bordingmi? Sen-chi? Kelding-a?
```

3-rasm. Amaliy natijadan namuna.

Shu o‘rinda ta’kidlash joizki, Python dasturlash tilida yozilgan kod faqat modal so‘zlar va so‘roq yuklamalari asosida ishlaydi. Bu yondashuv tinish belgilarini tiklashda ayrim istisno holatlar uchun yuqori samaradorlik ko‘rsatkichni bermasligi mumkin. Biz bilamizki modal so‘zning tinish belgisiz shakli ham to‘g‘ri deb qabul qilinishi mumkin. Masalan: *“To‘g‘ri, yo‘llar haydovchilar uchun muhim.”*, *“To‘g‘ri, yo‘llarda ham qiyinchiliklar bo‘ladi.”* va *“To‘g‘ri yo‘llar haydovchilar uchun muhim.”*, *“To‘g‘ri, yo‘llarda ham qiyinchiliklar bo‘ladi.”* kabi gaplarda faqat bitta vergul qo‘yilishi natijasida gapning ma’nosi tubdan o‘zgaradi. Bunday nozik holatlarni avtomatik tarzda to‘g‘ri aniqlash esa hali ham dolzarb muammo bo‘lib qolmoqda. Bu esa mazkur sohada yanada chuqurroq lingvistik tahlil va zamonaviy algoritmlarni joriy qilish zaruratini ko‘rsatadi.

Xulosa. Yuqoridagi tadqiqot shuni ko‘rsatadiki, o‘zbek tilida tinish belgilarini avtomatik tiklash masalasi sun’iy intellekt modellari, katta hajmdagi ma’lumotlar bazasi va domenlar o‘rtasidagi farqlarni hisobga olgan holda yondashishni talab qiladi. Tinish belgilarining grammatik-sintaktik strukturadagi o‘rni beqiyos bo‘lib, ularning avtomatik aniqlanishi NLPning ko‘plab sohalari uchun dolzarb masaladir. Chuqur o‘qitish modellari va lingvistik qoidalarni uyg‘unlashtirish bu boradagi samarali usullardan biri bo‘lsa-da, matn konteksti noaniqligi, domenga xoslik va o‘zbek tiliga xos grammatik xususiyatlar kelajakda maxsus korpuslar asosida hal etilishi mumkin. Ayniqsa, modal so‘zlar va so‘roq yuklamalarining kontekstga ko‘ra o‘zgaruvchanligi ularni alohida o‘rganishni talab etadi va istiqbolli tadqiqot yo‘nalishi sifatida qaralmoqda.

Foydalanilgan adabiyotlar:

1. Nagy A., Bial1 B., Ács J. *Automatic punctuation restoration with BERT models.* 2021. arXiv:2101.07343v1 [cs.CL] 18 Jan 2021.
2. Elov B., Alayev R., Abdullayev A., Aloyev N. Yuqori n-gram modellarini o‘zbek tili matnlariga qo‘llash. Raqamli Transformatsiya va Sun’iy Intellekt ilmiy jurnali. Volume 2, Issue 5, October 2024. – P. 152-162.

3. Lai V., Veyseh A., Dernoncourt F., Nguyen T. Punctuation Restoration. Dublin, Ireland, 20-24 August 2023. – P. 446-450.

4. Poláček M., Cerva P., Zdansky J., Weingartova L. Online Punctuation Restoration using ELECTRA Model for streaming ASR Systems. In Interspeech. (2023).

5. Bakare A., Anbananthen K., Muthaiyah S., Krishnan J., Kannan S. Punctuation Restoration with Transformer Model on Social Media Data. Appl. Sci. 2023, 13, 1685

6. Gravano A., Jansche M., Bacchiani M. Restoring punctuation and capitalization in transcribed speech. Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2009, 19-24 April 2009, Taipei, Taiwan.

7. Ishmurzayeva L. O‘zbek tilshunosligida tinish belgilarining ilmiy tadqiqi. Journal of Innovations in Scientific and Educational Research, Volume-6, Issue-4, DOI: 10.5281/zenodo.7878232. (2023).

8. Ochilov M., Jurayev D., Narzullayev O. Matndagi tinish belgilarni tiklash muammolari va yechimlari. Kompyuter Lingvistikasi: Muammolar, Vechim, Istiqbollar, Vol. 1, No. 01, 2024.

VI SHO‘BA. KOMPYUTER LINGVODIDAKTIKASI

ҚАЖЫҒАЛИ МҰХАНБЕТҚАЛИҰЛЫ ШЫҒАРМАЛАРЫНЫҢ КӨРКЕМДІК ЕРЕКШЕЛІКТЕРІ

Миров М.О.,
Филология ғылымдарының кандидаты, доцент
mirov-agu@bk.ru
Жұмамұратова А.С.,
Магистрант
aminazumamuratova@gmail.com
Қ. Жұбанов атындағы Ақтөбе өңірлік университеті
Қазақстан, Ақтөбе

Аннотация. Бұл мақалада қазақ прозасының көрнекті өкілі Қажығали Мұханбетқалиұлының шығармашылық ерекшеліктері мен көркемдік шеберлігі талданады. Жазушының «Өмір» және «Қысқы түн» әңгімелері негізінде адам жанының терең иірімдері, рухани ізденіс, ұлттық болмыс және қоғам мен тұлға арасындағы байланыс көркемдік құралдар арқылы сараланып көрсетілген. Автордың стильдік тәсілдерге (метафора, эпитет, монолог, символика) ерекше мән беруі оның прозасын терең мазмұнмен, философиялық оймен байытады.

Аннотация. В статье рассматриваются особенности творчества и художественного мастерства известного казахского прозаика Кажыгали Муханбеткалиулы. На основе рассказов «Өмір» и «Қысқы түн» анализируются внутренние переживания героев, духовные поиски, национальный менталитет и отношения между личностью и обществом. Особое внимание уделяется стилистическим приёмам (метафора, эпитет, монолог, символика), которые придают произведениям глубину и философскую насыщенность.

Abstract: This article analyzes the literary style and artistic mastery of prominent Kazakh prose writer Qazhygali Mukhanbetkaliuly. Based on his short stories "Ömir" and "Qysqy tün", the article explores the psychological depth of characters, spiritual quest, national identity, and the relationship between individual and society. The writer's use of stylistic devices such as metaphors, epithets, monologues, and symbolism enriches the narratives with philosophical insight and emotional depth.

Кілт сөздер: *қазақ прозасы, көркемдік тәсіл, метафора, эпитет, монолог, символика, рухани ізденіс, адамгершілік, ұлттық болмыс.*

Қазақ әдебиетінде шынайы болмысты терең психологиялық тұрғыда бейнелеп, ұлттық мінез бен адамгершілік мәселелерін көркем тілмен жеткізген

қаламгерлердің бірі – Қажығали Мұханбетқалиұлы. Ол шығармаларында қарапайым адамдардың ішкі жан әлемін, өмірлік күресін, рухани ізденісін шынайы әрі әсерлі суреттейді. Жазушының прозалық туындылары терең мазмұнымен, идеялық тұтастығымен және көркемдік шеберлігімен ерекшеленеді.

Қажығали Мұханбетқалиұлының көркем әлемінде адам тағдыры, ар мен ұждан, қоғам мен тұлға арақатынасы басты тақырыптар ретінде алдыңғы қатарға шығады. Ол кейіпкерлердің ішкі күйзелісін, сана иірімдерін терең ашып, оқырманды өмір туралы ойлануға жетелейді. Шығармаларында шынайы өмір көріністері мен ұлттық болмыстың сыр-сипаты табиғи түрде тоғысып жатады.

Жазушының шығармашылық ерекшелігіне орай Шархан Қазығұл былай деп баға береді:

«Қажығали ағаның бойында өмір бойы қалтқысыз қалыптасқан таза сана мен ақыл-парасаттың апогейі осы шығармасының 62 тамырын бүлкілдетіп тұр. Қажығали Мұханбетқалиұлы өзінің интеллектік, эмоционалдық, рухани, керек десеңіз, физикалық күштерін үйлесімділік пен жарасымдылық биігіне көтере білуінің нәтижесінде эксклюзивті рух диалектикасын қалыптастыру арқылы осы іргелі туындыны табандылықпен жазып шыққанына бек сенімдімін.» [1]

Бұл пікір қаламгер шығармаларындағы терең рухани қуатты, интеллектуалдық тұтастықты және көркемдік шеберлікті дәл сипаттайтынына дәлел бола алады.

Жазушының «Өмір» және «Қысқы түн» атты әңгімелері осы көркемдік ерекшеліктердің нақты көрінісі болып табылады. Ол бұл әңгімелер арқылы адам жанының нәзік қырларын, қоғамдағы өзгерістердің адам бойына әсерін терең философиялық және көркемдік деңгейде жеткізе білген. Әсіресе кейіпкерлердің ішкі толғаныстары мен өмірлік таңдаулары арқылы олардың рухани әлеміне үңілуге мүмкіндік туады. Әңгімелердегі символдық бейнелер мен детальдар шығарманың мазмұнын тереңдете түседі. Бұл туындылар Қажығали Мұханбетқалиұлының сөз өнеріндегі шеберлігін ғана емес, оның гуманистік көзқарасын да айқын аңғартады.

Қаламгердің қазақ прозасында жоғары бағаланған, оқырманға ерекше ой салған туындысы - «Өмір» әңгімесі. Бұл шығармада қарттық, өлім, өмірдің мағынасы сынды ұлы ұғымдар күнделікті ауыл тұрмысының, табиғат көрінісінің және қарапайым кейіпкерлердің шынайы әңгімесі арқылы әсем бейнеленеді. Әңгіме өзінің лирикалық-философиялық сипатымен, көркемдік-стильдік әдістерінің байлығымен ерекшеленеді.

Шығарманың негізгі идеясы — өмірдің өткінші екенін түсіну, қарттық пен өлімді табиғи заңдылық ретінде қабылдау, сондай-ақ сол жолда абыроймен ғұмыр кешудің маңызын көрсету. Елекең мен Тәйкеннің арасындағы әңгіме — сырт қарағанда қарапайым ауыл қарияларының сөзі болғанымен, шын мәнінде тіршілік философиясына айналған. Елекеңнің «ертелі-кешті бәріміз де сол жерге бармаймыз ба... тек абыроймен алса болды дағы» деген сөзі — өлімді үрей емес, қабылдауға тиіс заңды ақиқат ретінде түсіну керектігін білдіреді.

Бұл әңгіме арқылы автор оқырманын өмірге шүкіршілікпен қарауға, әр кезеңнің қадірін білуге үндейді. Жас жігіттің қариялар әңгімесіне таңырқауы мен терең ойға батуы арқылы жазушы ұрпақтар сабақтастығын да айқындайды. Жас ұрпақ естелік арқылы, қарттар өмір тәжірибесімен өзара байланысқа түсіп, өмір мәніне жаңа қырынан қарай бастайды.

Сондай - ақ, шығарманың мәнді, мазмұнды жеткізілуіне, оқырманның жүрегінен орын алуына стильдік ерекшеліктердің қолданылуы маңызды. Жазушы табиғатынан әр әңгімені өзгеше толғаныспен әдеби бейнелеу құралдары арқылы баяндайды. “Өмір” шығармасының қуатын арттырып тұрған басты тәсілдердің бірі — табиғат көрінісі мен адам көңіл-күйінің үндестігі. Әңгіме басындағы:

- *«Күн ұясына байыды. Бірақ батыс жақ алаулай қызарып, біразға дейін тұрып алды...»* — деген сөйлемде табиғат суреті арқылы өмірдің соңғы кезеңі, яғни қарттық символикалық түрде беріледі. Күннің батуы — өмірдің бітімін білдірсе де, оның «алаулай қызарып» тұрып алуы — өмірдің әр сәтінің өз мағынасы, өз жылуы барын аңғартады. [2,96]

Автор психологиялық тереңдікке де ерекше мән береді. Елекең мен Тәйкеннің естеліктері — тек қарттық туралы емес, өмірдің құндылығы, жастықтың қадірі, отбасылық қатынас пен қарияға деген құрметтің өзгеруі туралы ойларға жетелейді. Мысалы, Елекеңнің Би ата туралы әңгімесі:

- *«Марқұм сөйтіп итшілікпен жүріп дүние салып еді...»* — деген жолдарда жазушы қоғамдағы қарияның орны, ұрпақтан көрер мейірімнің азаюы секілді әлеуметтік астарлы мәселелерді көтереді. Бұл — шығармадағы идеялық салмақты арттырып тұрған тұстардың бірі. [2,176]

Кейіпкерлердің бейнесі портреттік детальдармен толық ашылады. Тәйкеннің «қайрала-қайрала жұқарып қалған қазан пышақ» кейпіндегі бейнесі мен Елекеңнің «қызғылт реңді дөңгелек жүзі» — бір-біріне қарама-қарсы, бірақ өмірді бірдей түсінген екі түрлі тұлғаны сипаттайды. Ал жап-жас студент жігіттің бейнесі — келешек өмірге сеніммен қарау, бірақ әлі тәжірибеден бейхабар жастықтың көрінісі.

Шығармада диалог тәсілі арқылы кейіпкерлердің ішкі ойы мен дүниетаным көзқарастары ашылады. Бұл диалогтар тек әңгімені дамытып қана қоймай, өмір мен өлім, қадір мен қасірет, шүкір мен өкініш туралы терең толғамдарға жол ашады. Елекең мен Тәйкеннің сөзі — халық даналығының көркем жалғасы іспетті беріледі.

Сонымен қатар, көркемдегіш құралдардың ішінде метафора мен эпитет көп қолданылған. Мысал келтіретін болсам:

- *«Жүз жасаған кәрі еменді түп-тамырымен қопарып алып, басқа жерге отырғызғанмен, мәуелер ме?!»* — мұндағы кәрі емен қартайған адамға (нағашысына) теңелген. Ескі жерінен айырылған қарияның жаңа ортаға бейімделе алмайтыны ағаштың тамырымен қопарылып, басқа жерге отырғызылғанымен мәуе бермейтіндігі арқылы астарлы бейнеде берілген. [2,175]
- *«Сол үйдің төбесіне төнген қырсықтың бұлты...»* — бұл жерде қырсықтықтың бұлты сөзі қиындық пен сәтсіздік теңеліп, образды түрде сипатталған. [2,175]
- *«Сақал шіркін де қартая ма дейім...»* — сақалдың қартайғаны арқылы адамның табиғи әлсіреуі, тіпті бейнесі де уақыттың әсерінен өзгереді деген терең ой берілген. [2,176]

Шығарманың көркемдік тәсілдері — табиғат суреттері, метафоралар, портрет пен диалог, ішкі монолог — осы әңгіменің идеясын терең жеткізуге септігін тигізді.

Қажығали Мұханбетқалиұлы шығармаларының ішінде тағы бір әңгімесі - «Қысқы түн». Әңгіме ауыл өмірінің тыныс-тіршілігін, қартаюдың ауыртпалықтарын және адам мен табиғат арасындағы қатынасты суреттейді. Мұнда ауылдың қарапайым өмірін, тұнжыраған қарт адамның ішкі әлемін, оның жалғыздық сезімін және заманның өзгерісін көрсету арқылы қоғамдағы рухани құлдырау мен адам мен табиғат арасындағы үйлесімділік мәселелері көтерілген. Шалдың өмірі мен ойлары терең символикалық мәнге ие, ол жалғыздықта өткен уақытты, өткен дәуірдің құндылықтарын аңсайды, бірақ осы аңсау оған әрі мұң, әрі қуаныш әкеледі. Әңгіме оқырманды қарттықтың қиындықтары мен елдің дамуындағы өзгерістер туралы ойлануға итермелейді.

Шығармада стильдік ерекшеліктер мен көркемдік тәсілдер терең ойластырылып, оқырманды ерекше әсерге бөлейді. Әңгімелерінде қолданылған эпитет, метафора, және монолог арқылы жазушының дара стилі айқын көрінеді. Осы тәсілдерді нақты мысалдармен талдай отырып, жазушының шығармашылық ерекшеліктерін ашуға болады.

Академик Р.Сыздық жай анықтауыш заттың қалыпты белгісін білдірсе, эпитет өзі атайтын белгіні бейнелеп көрсететінін және оның қалыпты табиғи белгі болмайтынын айтады. [5,15] Эпитет — бұл қандай да бір объектінің, құбылыстың немесе кейіпкердің қасиетін, сипатын немесе ерекшелігін айқындап, көркем түрде сипаттайтын сөз немесе сөз тіркесі. Мысалы, Қажығали Мұханбетқалиұлының шығармасында эпитетті қолданудың бір мысалы: «*Қартайған шалдың көздері жарқырап тұрды»*. Бұл жерде *«жарқырап тұрған»* эпитеті шалдың көзін тек физикалық тұрғыда сипаттаудан гөрі, оның ішкі жан дүниесінің, өмірге деген белсенді көзқарасының белгісі ретінде қолданылады. Шал қартайса да, оның көзі жанының жарқырап тұрғанын, өмірге деген көзқарасындағы жігерін көрсетеді. Эпитет «жарқырап» сөзі арқылы кейіпкердің қарттық кезеңінде де өмірге деген құштарлығы мен сезімталдығын көрсетеді, әрі оқырманға оны тек қарт ретінде ғана емес, жан дүниесі бай, күрделі адам ретінде суреттейді. [3, 214]

Көне заманнан бері қарай метафора жалпы ауыстыру ретінде кең мағынада, кейде троптың бір түрі ретінде тар мағынада түсіндірілуде. Метафораны «қысқарған теңеу» деуге болмайды, өйткені оларды семасиологиялық тұрғыдан қарасақ, теңеуге тура, метафораға ауыс мағына тән. [4, 13] Әңгімеде «***қарттың жүрегі***» метафорасы ерекше назар аударады. Жазушы мұнда жүректі тек дененің бір бөлігі ретінде емес, адам жан дүниесінің символы ретінде көрсетеді. Мысалы, оның жүрегі ауырлай бастаған сайын, қаламгер қарттың ішкі дағдарысын, уақыттың өтумен бірге, оның рухани күйінің өзгеретінін сипаттайды. Бұл метафора арқылы жүректің ауырлығы, оның өмірге деген көзқарасындағы қиындықты білдіреді. Жазушының метафорасы тек сөздің әсемдігіне емес, оның терең философиялық мағынасына да қызмет етеді.

Шығарманың стильдік ерекшеліктерінің бірі — монологтардың тереңдігі мен табиғилығы. Қажығали Мұханбетқалиұлы кейіпкерлерінің ойларын және сезімдерін монолог арқылы шынайы көрсетеді. Мысалы, Дайрабай шалдың монологы өте әсерлі. Оның баласы Тұрғалиға деген өкініші мен ойлары оның ішкі жан дүниесін ашып көрсетеді. Шалдың өз-өзімен сөйлесуі арқылы оның өмірге деген көзқарасы мен қарттықты қабылдауы айқындалады:

«Тек енді Тұрғалижанның екі жылы тез бітіп, оралса болды дағы**...**» [3, 212]

Бұл монологта шалдың жалғыздық сезімі, өтіп кеткен өмірі туралы ойлары, баласының оралуы арқылы соңғы үмітінің тірелуі сезіледі. Мұнда жазушы адамның ішкі тыныштығы мен сыртқы әлемге деген көзқарасындағы

қарама-қайшылықты терең көрсетеді. Жазушының монологты қолдануы оның кейіпкерінің рухани күйін толық ашуға көмектеседі.

Қорытындылай келе, Қажығали Мұханбетқалиұлы — қазақ прозасының көркемдік деңгейін биіктеткен, адам жанының терең иірімдерін шынайы суреттей білген суреткер. Оның шығармаларындағы кейіпкерлер тек көркем бейне ғана емес, ұлттық болмыстың, рухани ізденістің, қоғаммен қарым-қатынастың айнасы іспетті. Әңгімелерінде өмір мен өлім, ар мен ұждан, табиғат пен адамның ішкі дүниесі бір-бірімен үндесіп, терең философиялық ойлармен өріледі.

Жазушы қолданған метафора, эпитет, монолог, символика секілді көркемдік құралдар туынды мазмұнын байытып, оқырманды терең сезім мен ойға жетелейді. Ол әр кейіпкері арқылы ұлттық дүниетанымды, адамгершілік құндылықтарды, өмірлік шындықты көркем тілмен өрнектейді.

Мархабат Байғұттың: *«Қажығали Мұханбетқалиұлы — тегеурінді топтың, Әбдіжәміл Нұрпейісов ағамыз айтқандай, әдебиетімізге “жаңаша тіл, жаңаша ойлау, жаңаша элегия әкелген” талантты топтың өзгеше өкілі. Қажығалидың қай-қай туындысы да оның ұлт суреткері екенін айқындап тұрады»* — деген пікірі жазушының көркемдік әлемін дәл сипаттап, талдауға негіз болған тұжырымдарды нақтылай түседі.

Қажығали Мұханбетқалиұлының шығармашылығы — қазақ әдебиетінің рухани байлығын толықтырған, ұлттық әдебиеттің көркемдік кеңістігін тереңдеткен құнды қазына ретінде бағаланады.

Пайдаланылған әдебиеттер:

1. https://zhasalash.kz/news/olmejtugyn-artynda-soz-qaldyrgan-21648/
2. Қажығали Мұханбетқалиұлы. Ақтөбе кітапханасы. Ақтөбе, 2014. 21 Т. – 2014. – 336 бет.
3. Мұханбетқалиұлы Қ. М 82 Сынық терезе. Әңгімелер мен хикаяттар. – Астана: «Елорда», 2003. – 272 бет.
4. Қазақ тілінде сөздердің метафоралы қолданылуы: мұғ.арн. көмекші құрал / Б.Хасанов;. – Алматы: Мектеп, 1966. – 207 б.
5. Қалижан Бекхожин поэзиясындағы теңеу мен эпитеттің қолданыс ерекшеліктері: филол. Ғылымд. канд. ғылыми дәрежесін алу үшін дайынд. дис. Автореф.: 10.02.02 / Қ.Т. Маликов. – Астана: [б.ж.], 2005. – 24, [5] б. – 100. – Б.ц.

NEUROLINGUISTICS IN THE PHILOLOGICAL CONTEXT: RESEARCH APPROACHES AND EDUCATIONAL PRACTICES

Mirov M.,
Candidate of Philological Science
mirov-agu@bk.ru

Zhumagul D.,
Master of Arts in Education
zhumaguldilara@gmail.com
Zhubanov Univeristy, Aktobe, Kazakhstan

Abstract: This article explores the integration of neurolinguistic research into the field of philology as a means of deepening our understanding of language processes and enhancing the effectiveness of language education. It analyzes key directions in neurolinguistics, research methods, and their applied value for language teaching, literary analysis, and the development of students' metacognitive skills. Educational strategies grounded in neuroscience are proposed, taking into account the cognitive characteristics of learners.

Keywords: *neurolinguistics, philology, language learning, cognitive processes, interdisciplinary approach, neuroeducation*

Contemporary philology is undergoing a re-evaluation of its methodological boundaries. In the age of digital technologies and interdisciplinary research, the need to integrate the humanities with advances in natural and cognitive sciences becomes increasingly evident. Neurolinguistics — the study of the neural mechanisms of human language activity — is one such point of intersection [*Zherebilo,* 2019. 45].

Traditionally, philology has been associated with textual analysis, cultural interpretation, and historical-linguistic research. However, in today's educational landscape, there is a growing need not only for humanistic but also for cognitive literacy — the ability to understand how individuals process language, what occurs in the brain during reading, writing, or verbal communication, and what pedagogical conclusions can be drawn from this knowledge.

The aim of this article is to present neurolinguistics as a relevant field for philologists, to outline its main research approaches, and to examine educational practices based on neurolinguistic data.

Neurolinguistics as a Scientific Discipline: Theoretical Foundations and Methods

Neurolinguistics emerged at the intersection of linguistics, neurobiology, and psychology. It studies which areas of the brain are involved in language activity, how

information is encoded, stored, and retrieved [*Friederici,* 2017. 24], how speech develops in children, and how it is disrupted in pathologies.

Methods include fMRI, EEG, TMS, and behavioral experiments [*Rodden& Stemmer,* 2008. 57].

Key questions in neurolinguistics include:

- What are the neural bases of language comprehension and production?

- How does the brain distinguish and process phonemes, morphemes, and syntactic structures?

- What happens in the brain during second language acquisition?

- How are emotion, memory, and language interconnected?

Research methods include:

- fMRI (functional magnetic resonance imaging) – allows observation of brain activity during language tasks;

- EEG (electroencephalography) – records electrical brain activity in response to words and sentences;

- TMS (transcranial magnetic stimulation) – helps determine how specific brain areas influence speech;

- Behavioral experiments – assess speech responses and reaction times;

- Analysis of speech disorders (aphasia, dyslexia, alexia, etc.) – provides insight into “normal” language functioning through pathology.

One key finding has been the identification of specialized brain regions: Broca’s area (speech production), Wernicke’s area (comprehension), and temporal and frontal lobes responsible for semantics and syntax.

The Role of Neurolinguistics in the Development of Language Competence

Philological education has traditionally focused on mastering language as a system. However, understanding the cognitive and neurophysiological processes underlying speech activity allows educators to design more effective teaching strategies. Knowledge of how the brain processes language can enrich philological education. For instance, bilingualism influences the development of the prefrontal cortex, which is responsible for cognitive control [*Kroll & Bialystok,* 2013. 141].

Foreign Language Teaching

Neurolinguistics provides answers to many practical questions arising in second language instruction:

- Why is phonetic acquisition harder for adults?

- How does bilingualism affect the brain?

- What are the most effective vocabulary memorization techniques?

Studies show that multimodal stimuli (visual, auditory, motor) activate more neural networks, enhancing learning. Associative learning (linking new words with images, emotions, movement) improves long-term memory. Understanding learners' cognitive differences supports educational individualization. It is crucial to note that learning through multimodal stimuli improves retention [*Shakhnarovich,* 2018. 102].

Writing Skills Development

Developing written speech involves cognitive processes such as text planning, structure organization, and work with syntax, semantics, and orthography. Neuroscientific data offer new perspectives on common student errors, which may reflect immature neural pathways rather than mere carelessness.

Text Comprehension

The perception and interpretation of text are central to philological education. Neurolinguistics shows that understanding metaphors, irony, and figurative language activates not only language-related areas but also emotional and motor brain zones, supporting the concept of embodied cognition.

Neurolinguistics in Educational Practice: From Research to Methodology

Integrating neuroscience knowledge into pedagogy can revitalize philological education. Key areas where this is especially relevant include:

Individualized Learning

Neuroscience confirms that learning styles (auditory, visual, kinesthetic) have neural foundations. Recognizing these differences enables educators to build personalized learning trajectories, particularly in inclusive education settings.

Addressing Reading and Writing Difficulties

Understanding the neurophysiological causes of dyslexia and dysgraphia allows for effective correction strategies, involving collaboration among speech therapists, neuropsychologists, and philologists.

Use of Digital Technologies

Neurolinguistics informs the development of adaptive educational apps that monitor cognitive load, learning pace, and memory zones. These tools help automate feedback and increase student motivation.

The Interdisciplinary Future of Philology

Neurolinguistics offers approaches that make philology more practical and scientifically grounded. This supports the development of students’ metacognitive strategies [*Hasson & Small,* 2008. 87].

Philology, while preserving its humanistic essence, gains new momentum through collaboration with brain and cognitive sciences. Such integrative approaches shape a new educational paradigm:

- Based on scientific data (evidence-based teaching);
- Aiming for conscious learning (metacognitive learning);
- Focused on students’ neuropsychological well-being (neuroeducation).

To prepare a new generation of philologists, educational programs must include introductions to neurolinguistics, cognitive psychology, and neuro-pedagogy. This will not only broaden students’ horizons but also allow them to perceive language as a living cognitive system.

Neurolinguistics provides philologists with new tools for analyzing language, literature, and speech activity. It bridges traditional humanistic approaches with precise neuroscientific data, paving the way for more effective, research-based education. Integrating neurolinguistics into philological knowledge fosters interdisciplinary thinking — a quality increasingly valued in today’s academic and educational communities.

This integration opens new horizons in both research and education, shaping a new type of philologist: an interdisciplinary specialist equipped with both linguistic and cognitive tools for language analysis.

References:

1. Friederici A.D. Language in Our Brain: The Origins of a Uniquely Human Capacity. — MIT Press, 2017.
2. Kroll J.F., Bialystok E. Understanding the consequences of bilingualism for language processing and cognition. // Journal of Cognitive Psychology, 2013, Vol. 25, pp. 140–146.
3. Shakhnarovich, A.M. Language and the Brain: A Cognitive Approach. Moscow: LKI, 2018.
4. Stemmer B. and Whitaker H. Handbook of the Neuroscience of Language, 2008, pp. 57-87
5. Zherebilo, T.V. Neurolinguistics and Cognitive Linguistics. Moscow: Languages of Slavic Cultures, 2019.

ӘОЖ 811.111

МАХМҰД ҚАШҚАРИ «ДИУАНИ ЛҰҒАТ АТ-ТҮРІК» ТІЛІНІҢ МОРФОЛОГИЯЛЫҚ ЖӘНЕ ЛЕКСИКАЛЫҚ СИПАТЫ

Қ. Жұбанов атындағы Ақтөбе өңірлік университеті
Қазақстан, Ақтөбе
Бисенгали Іңкәр
Кенебаева Ұлпан
Жақанқызы Меруерт
bisengaliinkar@mail.ru
Ғылыми жетекші ф.ғ.к., қауымдастырылған профессор
Мұратбек Бағила Құрманбекқызы

Аннотация. Бұл мақалада түркі халықтарының орта ғасырлық жазба мұраларының рухани, мәдени және тарихи маңызы жан-жақты талданады. Түркі жазба ескерткіштері тарихи, әдеби, діни, құқықтық және тілдік тұрғыдан аса құнды дереккөз ретінде қарастырылады. Орхон-Енисей жазбаларынан бастап, ислам дәуіріндегі әдеби және ғылыми еңбектерге дейінгі жазба мұралар мазмұны мен жанры бойынша сараланып, олардың түркі халықтарының дүниетанымы мен тілдік құрылымындағы орны айқындалады. Әсіресе Махмұд Қашқаридің «Диуани лұғат ат-түрк» атты еңбегі тілдік тұрғыда талданып, көне түркі мен қазіргі қазақ және түрік тілдеріндегі сабақтастық көрсетіледі. Сондай-ақ араб тілінің әдеби тіл ретінде қызмет етуі, оның түркі тілдеріне әсері мен әдеби тіл қалыптасуындағы рөлі ғылыми негізде қарастырылады.

Аннотация. В данной статье всесторонне анализируется духовное, культурное и историческое значение средневековых письменных памятников тюркских народов. Тюркские письменные источники рассматриваются как ценные сведения с исторической, литературной, религиозной, правовой и языковой точек зрения. От орхоно-енисейских надписей до литературных и научных трудов исламской эпохи — памятники классифицируются по содержанию и жанру, определяется их роль в мировоззрении и языковом строе тюркских народов. Особое внимание уделено лингвистическому анализу труда Махмуда Кашгари «Диуани лугат ат-тюрк», в котором прослеживается преемственность между древнетюркским языком и современными казахским и турецким языками. Также рассматривается роль арабского языка как литературного языка и его влияние на формирование тюркской литературной речи.

Abstract. This article provides a comprehensive analysis of the spiritual, cultural, and historical significance of medieval written monuments of Turkic peoples. These written sources are examined as valuable references from historical,

literary, religious, legal, and linguistic perspectives. From the Orkhon-Enisei inscriptions to the literary and scientific works of the Islamic period, the monuments are classified by content and genre, and their role in the worldview and linguistic structure of Turkic nations is emphasized. Special attention is given to the linguistic analysis of Mahmud al-Kashgari’s “Diwan Lughat al-Turk”, which illustrates the continuity between Old Turkic and modern Kazakh and Turkish languages. The article also examines the role of the Arabic language as a literary medium and its impact on the development of literary Turkic language.

Кілт сөздер: *Орта ғасыр, жазба мұра, көне түркі, руникалық жазу, араб графикасы, әдеби тіл, «Диуани лұғат ат-түрк», тілдік сабақтастық, лексикалық қор, араб мәдениеті, фразеология, сұраулық шылау, тарихи ескерткіш.*

Түркі халықтарының орта ғасырлық жазба мұралары – бұл олардың рухани, мәдени және тарихи дамуын бейнелейтін құнды дереккөздер. Бұл мұралар тарихи, әдеби, құқықтық, діни және тілдік тұрғыдан өте маңызды болып табылады. Ортағасырлық жазба ескерткіштері – түркі халықтарының мәдениеті, тарихы, тілі мен әдебиетінің даму кезеңдерін сипаттайтын маңызды дереккөздер. Бұл ескерткіштер әртүрлі әліпбиде жазылған (руника, ұйғыр, араб графикасы) және мазмұны бойынша тарихи, әдеби, діни, құқықтық және лексикографиялық еңбектер болып бөлінеді. Ортағасырлық жазба ескерткіштердің тарихи мәніне тоқталсақ, орта ғасырлар – түркі халықтарының қалыптасуы мен дамуының маңызды кезеңі. Бұл дәуірде түркі мемлекеттері (Батыс және Шығыс Түрік қағанаты, Ұйғыр қағанаты, Қарахан мемлекеті, Алтын Орда т.б.) құрылып, олар өздерінің жазба дәстүрін қалыптастырды. Аталған жазба мұраларды мазмұны мен жанры бойынша бірнеше топқа бөлуге болады:

- Ерте ортағасырлық жазба ескерткіштер (VI-IX ғғ.) Бұл кезеңге көне түркі руникалық жазба ескерткіштері жатады, оған мысал ретінде Орхон-Енисей ескерткіштерін (VI-VIII ғғ.) жатқызамыз. Күлтегін, Білге қаған, Тоныкөк жазбалары – түркі халықтарының алғашқы тарихи жазба деректері. Бұл ескерткіштерде мемлекет басқару, жауынгерлік рух, түркілердің саяси бірлігі баяндалған. «Күлтегін» және «Тоныкөк» жырлары өткен дәуірден нақты мәлімет беруге арналған тарихи деректер.

- Ислам дәуіріндегі жазба мұралар (X-XV ғғ.) Бұл кезеңде түркі халықтары араб жазуын қабылдап, ғылыми және әдеби еңбектер дами бастаған болатын.

- Махмұд Қашқаридың «Диуани лұғат ат-түрік» (1072-1074). Бұл еңбек барша түркі тілдерінің алғашқы сөздігі болатын сонымен қатар,

грамматикалық және лексикалық ерекшеліктері сипатталған алғашқы еңбек болатын.

- Жүсіп Баласағұнның «Құтадғу білік» (1069-1070) Бұл еңбек- Түркі әлеміндегі алғашқы саяси-дидактикалық поэма. «Құтты білік» дастанында мемлекет басқару, билік, адамгершілік мәселелері қарастырылады.

Ахмет Яссауидің «Диуани хикмет» (XII ғ.) Бұл еңбең алғашқы сопылық философия мен исламдық дүниетаным көрініс тапқан діни-дидактикалық өлеңдер жинағы. Ахмет Яссауи – сопылық әдебиеттің ірі өкілі, әйгілі ақын, есімі күллі ислам әлеміне мәшһүр болған ғұлама-ойшыл. «Диуани хикмет» («Парасаттылық» немесе «Даналық кітабы») есімді өлеңдер жинағы — түркі халықтарының XII ғасырдан сақталған әдеби ескерткіш. «Даналық кітабын» зерттеушілер М.Ф.Кепрүлү–заде, А.К.Боровков, А.Н.Самойлович, Е.Э.Бертельс, Э.Р.Рустамов, Н.М.Маллаев, Г.Ф Благова т.б. «Диуани хикметтің» қазақ тіліне тәржімалануы жолында: М.Жармұхамедұлы, С.Дәуітұлы, М.Шафиғи, Р.Сыздық, Б.Сағындықұлы сынды ғалымдар зор үлестерін қосты. «Диуани хикмет» — исламның діншариғат жолдарын дидактикалық мазмұнда поэзия тілімен бейнелеген жыр жинағы. «Даналық кітабы» адамзатты адамгершілікке, туыстастыққа, бауырмалдылыққа, ізгілікке, имандылыққа шақырып, мұсылман дінінің сүннет, қағидалары мен шарттарын, Алланы танудың жолдарын баяндайды. «Диуани хикметтің» сопылық ағымның философиялық ой-пікірлері: әділдік жолына түсу, ақиқатты іздеп табу, адамның рухани өмірінің таза болуы.

Ортағасырлық түрколог Махмұд Қашқаридің **“Диуани лұғат-ит-түрк”** (Түркі тілдері сөздігі) еңбегі түркі дүниесіне ортақ мұра және ежелгі тіл жәдігері. Бұл еңбек араб тілінде жазылған түркі сөзінің аудармалары. Бұл еңбектен қазақ және түркі тіліне ортақ лексиканы көреміз. **Махмуд Қашқаридың “Диуани лұғат ит-түрк” (Түркі тілдерінің сөздігі)** атты еңбегі 11 ғасырдың (1072-1074 жылдары жазылған) ескерткіші. Бұл сөздік тұңғыш рет 1915-1917 жылдары Түркияда үш том болып жарық көрген болатын. Кейінірек оны түрік ғалымы Бесім Аталай араб тілінен түрік тіліне аударып, 1939, 1940, 1941 жылдары үш кітап етіп жариялады. Бұл сөздіктің қазақ тіліне аударылуына қазақ зерттеушілері де, лингвистері де ат салысып келеді.[1;48-б]

Алдымен, бұл еңбек не себепті араб тілінде жазылды деген сауалға тоқталайық. Тарих қойнауына көз жүгіртетін болсақ 751 жылы Атлах шайқасы түркі әлеміне үлкен өзгеріс әкелді. Біріккен араб пен түрік әскері Тан империясымен соғысының нәтижесі жеңіспен аяқталды. Соғыстың тағы бір нәтижесі күллі түркі жеріне араб мәдениеті мен білім-ғылымы таралды. Әсіресе, 8-10 ғасырларда Ұлы Жібек жолы бойына арабша сауат ашатын

медресселер салынып, көне түркі жазуының орнын араб жазуы басты. Араб тілі түркі халықтарында *әдеби тіл* қызметін атқарып, түркі әлеміне құнды еңбектер араб тілінде жарық көрді.Дегенмен, біздің түркі халықтарының тілі жойылып кетпеді,тек араб қаріпіне көшті.

Талдау:

- « اِنْكَاكْ » іңек деп оқылатын бұл сөздің мағынасы сиыр деген ұғымды білдіреді. Түрік тілінде «сиыр» сөзі дәл осылай, яғни «inek» деп қолданылады. Бір қарағанда араб тілінде жазылғанымен бұның түрік сөзі екендігін аңғаруымызға болады.Себебі араб тілінде сиыр сөзі «البقرة» «бақара» деп қолданылады.

- «أُنِتِّى» сөзінің оқылуы унутты. Қазақ тілінде «ұмытты» деген ұғымды білдіретін бұл сөз түрік тілінде «unuttu» деп қолданылып, дәл осы мағынаны білдіреді. Араб тілінде «نسيت» «ұмытты» сөзінің жазылуы әрі айтылуы басқа фонетикалық құрылым екенін негізінде құрылған [2,252].

- «بُشْى» бос мағынасында жазылған бұл сөздің екі тілде де беретін мағынасы «бос» (қазақ тілінде) «boş» (түрік тілінде), ал негізгі араб тілінде « فارغة» «фариға» [2,386].

- « قشْى» қыш деп оқылатын бұл сөз қазақ және түрік тілінде де бір мағынада «қыс» мезгілін білдіру үшін жұмсалады [2,389].

- «اَرْ» ер мағынасында қолданылатын сөз. Қазақ тілінде де түрік тілінде де «ер, ер адам» мағынасында лексикалық қорымызда бар.

- « اَوَتْ» EVET ЕБЕТ: «дұрыс, мақұл, әлбетте» деген мағынадағы сөз.

Бұл сөз қазақ тілінде қолданыста жоқ, ал түрік тілінде осы күйінде қолданылады.

Диуани лұғат-ит-түрк” еңбегіндегі «Алып ер Тұңғаны жоқтау» жырының тілдік қолданысындағы шылаулар:

- Алп ер Тұңға өлдүмү

Әсіз аздұн қалдымү

Өзләк өдшін алдымү

Әмді йүрәк йүртілүр [3,63]

Аударма:

Алып ер Тұңға өлгені ме,

Әзіз дүние қалғаны ма.

Тағдыр өшін алғаны ма

Енді жүрек жыртылды

«өлдүмү»; «қалдымү»; «алдымү» - бұл жерде тіркесіп тұрған «мү» қазіргі қазақ тіліміздегі «ма, ме, ба, бе, па, пе» сұраулық мағынадағы шылаулар. Қазақ тілінде дыбыстық өзгеріске ұшырған.

Түрік тіліндегі нұсқа бойынша, «мү» - «mu, mi, mü, mı» сұраулық шылау. Бұл жерде фонетикалық құрылым сақталған.

Осы мысалдар арқылы Махмұд Қашқаридің “Диуани лұғат-ит-түрк” (Түркі тілдері сөздігі) еңбегі араб тілінде жазылды дегенімізбен, сөздік ішінде сөз қолданыстары көне түркі тілін құрайды. Сонымен қатар, қазіргі лексикалық қорымызда қолданыста жүрген бұл сөздер 11-ғасырлық еңбекке тән екендігін көре аламыз.

Араб тілі түркі тілдер семьясына жатпағанымен, қазақ және түркі тіліне ұқсайтын лексикалық қоры да жоқ емес. Сөздік ретінде жинастыратын болсақ, мына бір мысалдарды келтіруімізге болады:

Араб – түрік – қазақ

دفتر – defter – дәптер

كتاب – kitap – кітап

دكان – dükkan – дүкен

دنيا – dünya – дүние

فيل – fil – піл

خبر – haber – хабар

حقيقة – hakikat – ақиқат

بركة – bereket – берекет

هواء – hava – ауа.

Түрік халықтарында араб тілі әдеби тіл қызметін атқарды десек, осы ретте әдеби тілдің қалыптасу жолдарын ашып көрсетейік. Әдеби тілдің негізгі үш қалыптасу жолы бар.

1. Бөтен халықтың тілінен қалыптасады;

2. Ең күшті диалектіден қалыптасады;

Бұл түрколог ғалым Сәрсен Аманжоловтың тұжырымдамасы бойынша

3. Жалпыхалықтық тілдің негізінде қалыптасады;

Демек, араб тілінің түркі тіліне әсерін әдеби тілдің қалыптасуының бірінші жолы ретінде қарастырамыз. Түркі тілдерінің лексикалық қорын араб тілі ғылым-білім, дін негізінде құрады. Бұл жоғары дамыған мәдениет пен өркениет, әрі тарихи оқиғалардың көрінісі.

Махмұд Қашқари сөздігінде келтірілген әдеби үлгілердің тіл және жанр жағынан қазіргі тілдерге ұластығын талдап көрсету-бүгінгі түркологтердің алдағы міндеті болып табылмақ. Өлеңдегі тармақ соңындағы сөздердің ұйқасуы қазақ өлең тезникасында да орын алатынын, сондай-ақ, ***йағы отын өчурген*** (дұшпан отын өшірген) , ***есүз ажун қалдыму*** (дүние иесіз қалды ма), **йүзі сарғайыб** (жүзі сарғайып) сияқты тұрақты тіркестердің беретін мағынасы жасалуы жағынан қазақ тіліндегі фразеологизимдерге келіп ұласатындығын көрсетуге болады.[1;49-б]

Қорытынды

Қорыта айтқанда, түркі халықтарының орта ғасырлық жазба мұралары – ұлттың рухани, мәдени және тілдік болмысының негізін айқындайтын баға жетпес мұралар. Бұл жазба ескерткіштер арқылы түркі халықтарының мемлекет басқару жүйесі, діни дүниетанымы, әдеби ойлау жүйесі мен тілдік құрылымы жөнінде құнды мәлімет алуға болады. Махмұд Қашқари еңбегі арқылы көне түркі тілі мен қазіргі түркі тілдерінің сабақтастығы айқындалса, араб тілі мен мәдениетінің ықпалы түркі тілдерінің әдеби тіл ретінде дамуына елеулі үлес қосқанын байқаймыз. Бұл жазба ескерткіштерді жан-жақты зерттеу – бүгінгі түркология ғылымының маңызды бағыттарының бірі болып табылады.

Пайдаланылған әдебиеттер:

1. 1.Сыздық Р. Қазақ әдеби тілінің тарихы. Алматы. “Арыс” баспасы. 2004жыл, -288б.
2. Қашқари М. “Диуани лұғат-ит-түрк” (1- том. Аударған Егеубай А.Қ). Алматы,2017 ж.
3. "Қазақ әдеби тілінің тарихы" хрестоматия. Қалиев Ғ , Боранбаева. Г (Алматы, 2000 ж.)
4. Сыздықова. Р «Қазақ тіліндегі ескіліктер мен жаңалықтар» (Алматы, 2009 ж.)
5. Мұратбек Б.Қ, Әмірғали Ж.О. «Қазақ-түрік тілдері ортақ лексикасының салыстырмалы сөздігі». Ақтөбе, 2022ж.
6. М.Балақаев "Қазақ тілі мәдениетінің мәселелері" (Алматы,1965 ж.)
7. С.Аманжолов "Қазақ әдеби тілі" (Алматы, 1949 ж.)

THE MAIN DIFFERENCES BETWEEN A STORY AND A FEATURE.

Kessikbayeva Assem Kabibebekqyzy,
PhD candidate
Yessenova Kalbike Omirbayevna,
Doctor of Philological Sciences, Professor
hazretosman@mail.ru
Abai Kazakh National Pedagogical University

Abstract. The main feature of an essay is its focus on accuracy and truthfulness, as well as a lively, artistic depiction of reality. An essay combines elements of reportage and artistic narration, which allows the author not only to convey information, but also to create an emotional perception of an event or object. Unlike a short story, which is a work of fiction with a fictional plot, an essay is based on real facts. A short story, as a rule, includes fiction, the construction of fictional characters and events, while in an essay the author focuses on reality, striving to reflect it in the most reliable form. An essay can be freer in style, often using elements of narration, dialogue, description, but always remains tied to reality. Thus, an essay and a short story, despite similarities in narrative form, perform different functions in literature and journalism, and it is important to understand these differences when writing or analyzing works.

Annotatsiya. Inshoning asosiy jihati uning aniqlik va haqqoniylikka qaratilganligi, voqelikning yorqin, badiiy tasviridadir. Insho reportaj va badiiy hikoya elementlarini birlashtiradi, bu muallifga nafaqat ma'lumotni etkazish, balki voqea yoki ob'ektni hissiy idrok etish imkonini beradi. O‘ylab topilgan syujetli badiiy asar bo‘lgan hikoyadan farqli o‘laroq, insho real faktlarga asoslanadi. Hikoya, qoida tariqasida, badiiy adabiyotni, badiiy qahramonlar va voqealarni qurishni o'z ichiga oladi, inshoda muallif haqiqatga e'tibor qaratadi va uni eng ishonchli shaklda aks ettirishga intiladi. Insho ko'pincha hikoya, dialog, tavsif elementlaridan foydalangan holda uslubda erkinroq bo'lishi mumkin, lekin har doim haqiqatga bog'liq bo'lib qoladi. Shunday qilib, insho va qissa, hikoya shaklidagi o‘xshashliklariga qaramay, adabiyot va publitsistikada turli vazifalarni bajaradi va asar yozish yoki tahlil qilishda bu farqlarni tushunish muhimdir.

Аннотация. Основной особенностью очерка является его фокус на точности и правдивости, а также на живом, художественном изображении реальности. Очерк сочетает элементы репортажа и художественного повествования, что позволяет автору не только донести информацию, но и создать эмоциональное восприятие события или объекта. В отличие от рассказа, который является художественным произведением с вымышленным сюжетом, очерк базируется на реальных фактах. Рассказ, как правило, включает в себя вымысел, построение вымышленных персонажей и событий,

тогда как в очерке автор ориентируется на действительность, стремясь отразить её в максимально достоверной форме. Очерк может быть более свободным по стилю, часто использует элементы повествования, диалогов, описаний, но всегда остаётся привязанным к реальности. Таким образом, очерк и рассказ, несмотря на сходства в форме повествования, выполняют разные функции в литературе и журналистике, и важно понимать эти различия при написании или анализе произведений.

Key words: *journalism, documentary as a feature of a feature article, feature article-story, feature article genre and its features.*

Introduction. In the world of writing, various forms of content serve different purposes, from informing and entertaining to expressing personal opinions and ideas. Among these forms, **features**, **essays**, and **stories** are some of the most common, yet they each possess unique characteristics. A **feature** is a journalistic style of writing that goes beyond the basic facts, offering in-depth analysis, personal angles, and detailed exploration of a subject. In contrast, an **essay** is more reflective and often focused on presenting the writer’s own viewpoint or argument on a specific topic, while a **story** – especially a **news story** – is typically a concise and objective account of a recent event, presenting facts in a clear and direct manner.

Understanding the distinctions between these forms of writing is crucial for both writers and readers. Each form has its own set of rules, purposes, and stylistic approaches that cater to different types of audience engagement. In this discussion, we will explore what a feature is, how it differs from an essay and a story, and why recognizing these differences is essential for effective writing and communication.

The founder of the theory and practice of the essay in Soviet literature, M. Gorky, defined the essay as "a work of art, but distinguished by its attachment to journalism, about the closeness of essays to stories in some essayists", "the essay has a full-fledged artistic value, at the same time it is a genre that meets the demands of the time", "the essay, on the one hand, is adjacent to journalism, on the other - to the story, a special genre that is unable to replace either a story or journalism". In the 1950-60s, this question was raised again among prose writers, literary scholars and critics: "if we consider the essay to be journalism, can journalism be called a work of art, how to distinguish between an essay and a story?" [1]. There were controversial interpretations around such questions. If the majority reduced the meaning of the essay to journalism, then some considered it a work of art, bringing it closer to a story. In his article “An essay is journalism,” published on April 25, 1958 in the Literaturnaya Gazeta, Yu. Surovtsev expressed the opinion: “An essay is either just journalism, or artistic journalism, but it is still journalism.” And those who consider an essay to be a “work of fiction,” in order to support their thoughts,

cite as examples artistic essays by the great Russian writers Korolenko G., Uspensky, Saltykov-Shchedrin, Gorky, Serafimovich, and others [1].

Definition of a Feature and a Story

Feature: A feature is a detailed, in-depth article that explores a subject, event, or individual in a way that provides context, background, and often a more personal or nuanced perspective. It is often longer and more comprehensive than a standard news story. Features focus on storytelling and are often found in lifestyle, culture, or human-interest sections of newspapers or magazines. The common feature between a short story and short stories is the following: both an essay and a short story, like a short story, are written in the form of literary artistic language, and they widely use portrait characteristics, landscape descriptions, and psychologism. The structure scheme is like that of a short story: it has an exposition, a plot, a climax, and a denouement. If in a short story the line of constructing an image lasts from the beginning of the work to its end, in an essay it is the same, although even the construction of an image is put first here [2].

Key characteristics of a feature:

Length: Typically, longer than news stories, often several hundred to a few thousand words.

Tone: More descriptive and narrative; it often delves into the emotions, motivations, or histories of the people or events it covers.

Structure: Feature stories often follow a narrative arc, using techniques like anecdotes, detailed descriptions, and interviews to build a vivid, engaging picture.

Purpose: To provide insight, evoke emotions, or offer a deeper understanding of a subject.

Story: A story, in the journalistic sense, refers to the more concise and immediate reporting of an event or piece of information. It is typically a news article that focuses on reporting the facts with clarity and precision, answering the "who, what, when, where, why, and how."

Key characteristics of a news story:

Length: Generally shorter than features, often just a few hundred words.

Tone: Objective and straightforward, aiming to inform rather than entertain or provide deep analysis.

Structure: Follows the inverted pyramid structure, where the most important information (the "lead") comes at the beginning, followed by supporting details.

Purpose: To quickly convey breaking news or provide essential facts on a subject.

Importance of Understanding the Differences for Writers and Journalists.

For writers and journalists, understanding the differences between a feature and a story is essential for several reasons:

Approach and Research:

Feature: Requires extensive research, interviews, and a deeper understanding of the subject. Journalists must be prepared to explore angles that are not immediately obvious and provide rich, detailed content.

Story: Requires rapid gathering of facts and the ability to synthesize information quickly. A journalist focusing on a story needs to be clear and concise.

Writing Style:

Feature: Writers need to embrace storytelling techniques, making use of descriptive language, vivid imagery, and emotional resonance to captivate the reader.

Story: The writing style needs to be clear, factual, and straightforward, focusing on delivering the news without unnecessary embellishment.

Deadlines and Timing:

Feature: Features may have more flexible deadlines as they involve more in-depth reporting and editing.

Story: News stories often have strict deadlines, especially if they relate to breaking news, requiring journalists to be quick and precise in their reporting.

Audience Engagement:

Feature: Features engage readers by offering rich content that invites them to think, reflect, or emotionally connect with the story. They are often used to attract readers' attention on a deeper level.

Story: News stories engage readers by providing essential updates quickly and efficiently, ensuring they are informed about current events.

Purpose and Impact:

Feature: The purpose of a feature is often to explore, explain, or humanize an issue. It can provide a lasting impact on the reader, encouraging reflection and discussion.

Story: The purpose of a news story is to inform, ensuring that the audience knows what happened as soon as possible. Its impact is often more immediate but may not carry the same depth as a feature.

Several well-known writers and journalists have written features from abroad, bringing unique perspectives and in-depth reporting on international topics.

These writers have often covered foreign events, cultures, and issues for various publications. These writers and journalists have contributed significantly to feature writing from abroad, using their international experiences to provide detailed, thoughtful, and often emotional reporting on global issues. Their work continues to shape the way readers understand cultures, conflicts, and human experiences from different parts of the world. Here are some notable writers and journalists known for their international feature writing:

1. Paul Theroux is a renowned travel writer and novelist known for his detailed features on places abroad. His book *The Great Railway Bazaar* is a classic of travel writing, where he explores train journeys through Europe, the Middle East, and Asia. Many of his works involve immersive reporting from various international locations, offering rich insights into local cultures.

2. Richard Engel, the chief foreign correspondent for NBC News, has written several features about conflicts, human rights issues, and political upheaval in the Middle East and beyond. His reporting on the Arab Spring and his coverage from war zones have often been featured in long-form articles and broadcast reports.

3. Though best known for her philosophical writings, Hannah Arendt also wrote feature-style articles that explored political situations in Europe and the broader world. Her reports on the Eichmann trial and her coverage of totalitarian regimes provide deep dives into the political realities of her time.

4. Ryszard Kapuściński, a Polish journalist, is famous for his feature writing on African and Latin American countries, often capturing the complexities of political and social situations in these regions. Books like *The Emperor* (about Ethiopia) and *The Shadow of the Sun* (about Africa) are considered masterpieces of feature writing, blending journalism with literary flair.

5. An acclaimed war correspondent and journalist, Martha Gellhorn wrote feature articles from conflict zones, including World War II and the Vietnam War. Her writings for *Collier's Weekly* and later other publications detailed the human cost of war and political crises, offering readers vivid, on-the-ground reporting from abroad.

6. Known for his book *The Perfect Storm*, **Sebastian** Junger has written long-form features about international conflicts, particularly in the Middle East. His book *War* is a collection of his experiences embedded with U.S. soldiers in Afghanistan, offering an in-depth, humanistic view of war.

7. Pico Iyer is a travel writer and essayist who has written many features about his travels abroad, often exploring themes of identity, culture, and the intersection of the East and West. His work, like *The Art of Stillness* and *The Open Road*, offers deep reflections on places and cultures he encounters.

8. Marie Colvin, a war correspondent for *The Sunday Times*, wrote numerous feature pieces from conflict zones like Iraq, Sri Lanka, and Syria. Her deeply human-centered approach to war reporting earned her widespread acclaim. Colvin’s pieces often emphasized the impact of war on civilians, and she was known for her courage in reporting from dangerous locations.

9. Though he’s primarily known for books, Tracy Kidder’s work in feature writing has been impactful. His reporting in *The Soul of a New Machine* (which explores the development of a new computer) and his feature writing for *The New York Review of Books* and other publications often takes a deep, investigative approach to topics abroad, especially in relation to technology and innovation.

10. Ari Shavit, an Israeli journalist, wrote extensively about the Middle East, offering nuanced features on the Israeli-Palestinian conflict, Israeli politics, and regional dynamics. His work often focuses on personal and societal narratives in the context of larger geopolitical issues, bringing international stories to readers in a compelling, accessible way.

A. Taabaldiev in the article "On the Peculiarities of Journalism" defines the following differences: "An essay does not dwell on all the problems of the day, but dwells on the most acute and new of them. It finds various manifestations, the first swallow of life, questions, problems, introduces new living heroes of the day into literature and submits them to the judgment of society. An essay with its narration of the new is called a combat genre - a scout. And a story in its own way is able to show all this after their dissemination, after everything becomes typical, characteristic. In this regard, the author's attitude to various new manifestations, which is the object of the essay, is shown openly and directly. And the story, describing the characteristic aspects of life, the author's attitude, his thoughts, mainly conveys them through well-known images" [3]. However, it seems to us that the discussion of synthesis arises because all the listed features of the essay are characteristic of other genres. If we consider the essay from the point of view of a distinctive feature, then the process of formation of new literary forms may indicate a different process than the process of disintegration of genres. Here we suggest turning for help to I. Kuzmichev, the author of the theory of synthesis of generic and differentiation of intra-generic, or genre forms [4]. Among other things, I. Kuzmichev's concept appeals to us because it does not deny the classical Aristotelian theory, but proves its validity and relevance. I. Kuzmichev suggests leaving behind the essay as a distinctive generic feature the real fact of reality, which organizes the remaining epic elements: plot, composition, methods of typification, the role of the narrator, means of subject depiction. This is how we understand the words of I. Kuzmichev that in the essay “circumstances are not ‘created’ in the course of the interaction of characters, but are taken directly from life” [5; p. 286].

For journalists and writers, understanding the distinction between features and stories allows them to adapt their approach to the needs of the assignment, whether it’s delivering quick news updates or crafting a thoughtful, immersive narrative. Each type of writing has its unique role in informing and engaging the audience, and knowing when and how to use each form ensures effective communication and storytelling.

Conclusion. For journalists and writers, understanding the distinction between features and stories allows them to adapt their approach to the needs of the assignment, whether it’s delivering quick news updates or crafting a thoughtful, immersive narrative. Each type of writing has its unique role in informing and engaging the audience, and knowing when and how to use each form ensures effective communication and storytelling.

In conclusion, while **features**, **essays**, and **stories** may all share the goal of engaging and informing the reader, they each have distinct characteristics that set them apart. A **feature** is an in-depth journalistic piece that delves into a topic or event, offering context, background, and often a human element to make the subject more relatable and engaging. It combines storytelling with factual reporting, making it longer and more narrative-driven than a news story.

An **essay**, on the other hand, is a more personal and reflective form of writing. It focuses on presenting an argument or exploring an idea from the writer's perspective, often using analysis, reasoning, and personal insights. Essays tend to be more subjective and philosophical, with less emphasis on external research compared to features.

Finally, a **story**, particularly a **news story**, is typically a shorter, factual account of an event that prioritizes clarity, objectivity, and the delivery of essential information quickly. Its structure is more rigid and follows the inverted pyramid model, ensuring that the most important details are presented first.

Understanding these differences is essential for writers and journalists, as it allows them to tailor their writing style, structure, and approach to suit the specific purpose and audience of each form. Whether informing, reflecting, or telling a story, each form of writing plays a unique role in shaping how we engage with information.

References:

1. Zhusueva S. Socio-moral problems in essays Ch. Aitmatova. Material International. scientific practice. conf. "Aitmatovsky readings-2016". Bishkek, 2016. S. 321-322, 255.

2. Osmonaliev K. Genre evolution of the Kyrgyz essay. F. "Kyrgyzstan", 1980. S. 7-9.

3. Tabaldiev A. Monument Oktyabryu. Frunze. "Kyrgyzstan", 1985. S. 221.

4. Podrobnee ob etom: Kuz'michev I.K. Literaturnye perekrestki: Tipologiya zhanrov, ikh istoricheskaya sud'ba. Gor'kiy. 1983.

5. Kuz'michev I.K. Vvedenie v obshchee literaturovedenie KhKhI veka. Lektsii. N.Novgorod. 2001.

ӘОЖ 81.373:001.4

ЭЛЕКТРОНДЫ СӨЗДІКТЕР: ҚАЗАҚ ЛЕКСИКОГРАФИЯСЫНЫҢ ЖАҢА МҮМКІНДІКТЕРІ

Дәулетші Дильназ Рахатқызы,
dilnazrahatkizi0505@gmail.com
Қ.Жұбанов атындағы Ақтөбе өңірлік университеті
Ғылыми жетекшісі: филология ғылымдарының кандидаты, доцент **Б.С.Карагулова**

Аңдатпа: Бұл мақалада компьютерлік лексикографияның теориялық негіздері мен қазақ тіліндегі электронды сөздіктердің даму бағыттары қарастырылады. Зерттеу барысында отандық және шетелдік электронды сөздік ресурстары салыстырылып, олардың құрылымдық, функционалдық және әдістемелік ерекшеліктері талданады. Сонымен қатар, қазақ тіліндегі заманауи электрондық платформалар – sozdik.kz, qazcorpus.kz, termincom.kz сияқты ресурстардың лексикографиядағы рөлі мен әлеуеті сипатталады. Мақалада қазіргі электронды сөздіктердің артықшылықтары мен кемшіліктері сараланып, болашақтағы даму бағыттары мен жетілдіру жолдары бойынша нақты ұсыныстар беріледі.

Abstract*:* This article explores the theoretical foundations of computational lexicography and the development trends of electronic dictionaries in the Kazakh language. The study compares domestic and foreign electronic dictionary resources, analyzing their structural, functional, and methodological features. Additionally, the article discusses the role and potential of modern Kazakh electronic platforms, such as sozdik.kz, qazcorpus.kz, and termincom.kz, in lexicography. The advantages and disadvantages of contemporary electronic dictionaries are examined, and specific recommendations for future development and improvement are provided.

Аннотация*:* В статье рассматриваются теоретические основы компьютерной лексикографии и направления развития электронных словарей казахского языка. В ходе исследования проводится сравнительный анализ отечественных и зарубежных электронных словарных ресурсов, их структурных, функциональных и методологических особенностей. Также рассматривается роль и потенциал современных казахских электронных платформ, таких как sozdik.kz, qazcorpus.kz, termincom.kz, в лексикографии. Анализируются преимущества и недостатки современных электронных словарей, даются конкретные рекомендации по направлениям их дальнейшего развития и улучшения.

Мақаланың негізгі мақсаты *– компьютерлік лексикографияның теориялық негіздерін зерттеп, қазақ тіліндегі электронды сөздіктердің даму бағыттарын анықтау.*

Мақаланың өзектілігі - компьютерлік лексикографияның дамуы ақпараттық технологиялар мен тіл ғылымы арасындағы байланысқа жаңа көзқарас ұсынады. Қазақ тіліндегі электронды сөздіктердің сапасы мен қолжетімділігі тілдің сандық дәуірдегі өзара байланысы мен дамуын анықтайды. Сонымен қатар, электронды платформалардың лексикографиядағы рөлі мен әлеуетін зерттеу қазақ тілін цифрлық кеңістікке енгізуге ықпал етеді. Мақала тілдің қазіргі жағдайы мен даму бағыттарын, сондай-ақ, болашақтағы жаңа технологиялар *мен ресурстарды жетілдіру жолдарын айқындауға мүмкіндік береді.*

Тірек сөздер: *компьютерлік лексикография, электронды сөздіктер, қазақ тілі, лексикографиялық ресурстар, ақпараттық технологиялар, заманауи платформалар, функционалдық ерекшеліктер, лексикографияның дамуы.*

Қоғам мен ғылым мен техникалық прогресс дамыған сайын қоғамда ақпаратпен алмасу, жаңа ақпарат іздеу оңай бола бастады. Технологиялардың, соның ішінде ақпараттық технологиялардың адам өміріне берер пайдасы зор. Олар адамның әртүрлі қызмет саласындағы мүмкіндіктері мен қызметтерін кеңейтуге, жеңілдетуге бағытталған. Соның ішінде, ақпаратты өңдеудеудің автоматтануы қазақ тіл біліміне елеулі үлес қосты. Ақпараттық автоматтандыру қазақ лексикографиясында компьютерлік бағдарламаларды тиімді түрде пайдалануға мүмкіндік бар екенін дәлелдейді.Осы орайда қазақ тіл білімін дамытуға септігін тигізіп отырған сала - компьютерлік лексикография саласы.

Алдымен, лексикография дегеніміз не? Oxford Dictionary сөздігінде бұл терминге “сөздіктерді жазудың теориясы мен практикасы” деген анықтама берілсе [1], Қазақ әдеби тілінің он бес томдық сөздігінде: “сөздік кітап болып, басылып шығуы да, тарауы да мүмкін. Мұндай сөздіктердің жинақты аты лексикография деп те аталады” делінген[2].

Ал компьютерлік лексикография болса - лексикография ғылымының заманауи техникалық мүмкіндіктерге бейімделіп, электронды форматтағы сөздіктер жасау, жаңарту және қолдану үрдістерін қамтитын бағыт.

А. Жұбановтың еңбектерінде лексикография және тілдің ақпараттық өңделуі жайлы пікірлері кездеседі:

Автоматты сөздіктің міндеті ғылыми-техникалық мәтіндердің барлық түрлерімен әрекет жасау болғандықтан, оның құрамына барлық терминдер

түрлері енулері қажет. Мысалы, оларға жататындар: стандарттанған және ұсынылған, сол сияқты ұсынылмаған, тұрақтанған немесе жаңадан қолданысқа енген терминдер. Сонымен қатар А.Жұбанов автоматты сөздіктің құрылымын анықтау жолында мынадай негізгі үш мәселенің шешімі табылғаны жөн дейді:

1) сөздік мақаланың басқы бірлігінің тұрпатын анықтау;

2) тілдік бірлікті сипаттауға қатысты ақпарат жиынтығын анықтау;

3) автоматты сөздік базасындағы сөздік мақаланы құрастырудың ұтымды тәсілін таңдау[3]. Сол сияқты Г.Б. Мәденова, С.Б.Бектемірова, Ж.Ж. Күзембекова электрондық сөздіктерді – қажетті лексикалық бірліктерді жедел аударуға, іздеуге және аударма бағытын икемдеуге мүмкіндік беретін, құрылымы мен қызметтік ерекшеліктеріне қарай әртүрлі мақсатта пайдаланылатын заманауи компьютерлік лексикографиялық құрал ретінде сипаттайды[4].

Қазақстанда соңғы онжылдықта қазақ тілін цифрландыру және компьютерлік лексикография бағытында нақты ілгерілеушіліктер байқалды. ҚР Ұлттық Ғылым академиясы жанындағы Тіл білімі институты тарапынан «Қазақ сөздігі» жобасы жасалып, 15 томдық электрондық база құрылды. Қазақ тіліндегі электронды сөздіктер мен ресурстар қазақ тілінің дамуында, оның қолданылуын жеңілдетуде және тіл үйренушілер мен зерттеушілер үшін өте маңызды құралдар болып табылады. Әрбір электронды сөздік өз ерекшеліктерімен және мақсатымен ерекшеленеді, бұл пайдаланушыларға қажетті ақпаратты тез әрі тиімді табуға мүмкіндік береді. Бүгінде халық қолданысы аясында кездесетін электронды сөздіктер қорлары:

❖Emle.kz – қазақ тілінің орфографиялық электрондық базасы. Бұл сайт орфографиялық ережелерді үйрену және тексеру үшін тиімді құрал болып табылады. Пайдаланушылар бұл ресурсты сөздердің дұрыс жазылуын тексеру, онлайн диктанттар жазу және орфографиялық ережелермен танысу үшін қолдана алады. Базада 100 000-нан астам сөз қамтылған, бұл тіл үйренушілер мен қазақ тіліндегі жазу мәдениетін жетілдіруге ұмтылғандарға пайдалы ресурс.

❖ Sozdik.kz – қазақша-орысша және орысша-қазақша онлайн сөздік. Бұл сөздік екі тіл арасында сөздер мен сөз тіркестерінің аудармасын ұсынады. Әр бағыт бойынша 60 000-нан астам сөздік мақалалар бар, сондай-ақ қазақ тіліндегі 16 000-нан астам сөздің айтылуы да ұсынылған. Пайдаланушыларға тілдің дыбысталуын нақты тыңдау мүмкіндігі беріледі, бұл әсіресе қазақ тілін үйренушілер үшін маңызды.

❖Sozdikqor.kz – әртүрлі салалар бойынша қазақ тіліндегі сөздер мен олардың аудармаларын қамтитын платформа. Бұл ресурс синонимайзер, сөздерді жүктеу мүмкіндіктері және басқа да қызметтер ұсынады. Пайдаланушыларға тілдік қорын байыту және нақты бір сала бойынша аударма жасауда көмек көрсетеді.

❖Termincom.kz – ғылыми, білім, экономика, қоғам және әлеумет салаларындағы терминдер мен ұғымдарды жинақтайтын республикалық терминологиялық электрондық қор. Бұл ресурс әртүрлі мамандық өкілдеріне қажетті терминологиялық ақпаратты ұсынумен қатар, терминдердің дұрыс қолданылуын қамтамасыз етеді. Сайт базасында барлығы 389 000 термин қарастырылған.

❖Lugat.kz – жалпы сөздік қоры 534мыңнан асатын қазақ тіліне арналған кешенді сөздік. Бұл ресурста қазақша-орысша, қазақша-түрікше, қазақша-ағылшынша, қазақша-арабша сөздіктер мен аймақтық сөздік, орфографиялық, синонимдік, аймақтық сөздіктер бар. Платформа қазақ тілінің байлығын толыққанды зерттеуге және тіл үйренушілерге барынша кеңес беруге бағытталған.

❖Atau.kz – қазақ тіліндегі ономастикалық атаулар мен терминдерді жүйелеуге арналған платформа. Бұл ресурс ерекше атаулар мен терминдермен жұмыс істейтін мамандарға, сондай-ақ ғылыми зерттеулерде немесе аударма жұмыстарына қажетті ақпаратпен қамтамасыз етеді.

Шетелдік және отандық электрондық сөздіктер: салыстырмалы талдау жасап көрелік.

Компьютерлік лексикография саласы халықаралық деңгейде де, отандық ғылымда да қарқынды дамып келе жатқан бағыттардың бірі. Бұл салада шетелдік тәжірибелердің үлгісі зор, себебі шет елдерде электрондық сөздіктерді құрастыру мен қолдану технологиясы ертерек қалыптасып, жүйелі зерттеулер мен практикалық шешімдер жүзеге асқан. Осы орайда шетелдік және отандық электрондық сөздіктердің ерекшеліктерін салыстыра қарастыру компьютерлік лексикографияның даму тенденцияларын анықтауға мүмкіндік береді. Шетелдік электронды сөздіктер – мысалы, Oxford English Dictionary (OED), Cambridge Dictionary, Merriam-Webster, Collins Dictionary сынды ресурстар – терең құрылымды, көпдеңгейлі лексикографиялық жүйемен ерекшеленеді. Бұл сөздіктер тек сөздің мағынасы мен аудармасын ғана беріп қоймай, оның этимологиясын, қолдану жиілігін, тіркесімділігін, грамматикалық ерекшеліктерін, стилистикалық реңкін және нақты мәтіндердегі қолданыс үлгілерін де қамтиды. Сонымен қатар, мұндай сөздіктер жиі түрде корпус лингвистикасы негізінде құрастырылып, нақты

деректерге сүйенеді. Қолданушы интерфейсі жоғары деңгейде дамыған, іздеу функциясы күрделі алгоритмдер арқылы жүзеге асады, ал кейбірі жасанды интеллект элементтерін де пайдаланады. Отандық электрондық сөздіктер, мысалы, sozdik.kz, termincom.kz, lugat.kz сынды платформалар, соңғы жылдары айтарлықтай жетістіктерге жетіп келе жатса да, әзірше кейбір техникалық және құрылымдық шектеулерге ие. Алайда отандық лексикографияда тілді сақтау, дамыту және кеңінен қолдану мақсатындағы ұлттық мүдде алдыңғы орында тұр. Қазақ тіліндегі электрондық сөздіктер көбіне екітілді (қазақ-орыс, қазақ-ағылшын) аударма бағытына негізделген және семантикалық өріс пен қолданым аясын көрсету жағынан шетелдік баламалардан қарағанда шектеулі. Сонымен қатар, тілдік деректерді автоматты өңдеу, толық мәтінді корпустармен байланыс жасау секілді мүмкіндіктер әлі де даму сатысында.

№1 Кесте Шетелдік және отандық электрондық сөздіктердің салыстырмалы сипаттамасы:

Критерий	**Шетелдік электрондық сөздіктер**	**Отандық электрондық сөздіктер**	**Мысал сөздіктер**	**Ресурстар (сілтемелер)**
Құрылымдық деңгейі	Көпдеңгейлі, мағыналық, грамматикалық, стилистикалық сипаттамалар қамтылған.	Қысқа анықтама, аударма және кейде айтылым беріледі.	Oxford English Dictionary, Merriam-Webster / Sozdik.kz, Emle.kz	www.oed.com, www.merriam-webster.com / www.sozdik.kz, www.emle.kz
Корпуспен байланысы	Үлкен тілдік корпустар негізінде жасалған.	Корпустық жұмыс енді қолға алынуда (qazcorpus.kz).	British National Corpus, COCA / Qazcorpus.kz	www.english-corpora.org / www.qazcorpus.kz
Мәліметтер қорының көлемі	Миллиондаған сөздік бірліктер, кең дерекқор.	60 000 – 500 000 аралығында, ресурстарға байланысты.	Cambridge Dictionary / Lugat.kz, Sozdikqor.kz	dictionary.cambridge.org / www.lugat.kz, www.sozdikqor.kz
Функционалды мүмкіндіктері	Айтылым, тіркесім, контекст, транскрипция, аудиомысалдар.	Аударма, айтылым, іздеу, диктант құралдары бар.	Longman Dictionary, Collins / Sozdik.kz, Emle.kz	www.ldoceonline.com, www.collinsdictionary.com / www.sozdik.kz
Интерфейс және қолдануыңғайлылығы	Заманауи UX/UI, мобильді нұсқалар, пайдаланушыға бейімделген жүйе.	Кейбіреулерінде заманауи интерфейс бар, бірақ толық оңтайландырылмаған.	Macmillan Dictionary / Termincom.kz, Atau.kz	www.macmillandictionary.com / www.termincom.kz, www.atau.kz

Лексикографиялық әдіснама	Корпустық, құрылымдық, семантикалық тәсілдер үйлесімі.	Дәстүрлі лексикографияға негізделген, жаңа әдістер енді еніп жатыр.	Oxford, Collins / Lugat.kz, Sozdikqor.kz	www.collinsdictionary.com / www.lugat.kz
Мақсатты аудитория	Тіл үйренушілер, зерттеушілер, аудармашылар, ғалымдар.	Қазақ тілін үйренушілер, оқушылар, мамандар, мемлекеттік мекемелер.	Merriam-Webster, Cambridge / Sozdik.kz, Emle.kz, Termincom.kz	www.merriam-webster.com / www.termincom.kz
Мемлекеттік қолдау	Көбінесе коммерциялық немесе академиялық жоба ретінде жүзеге асады.	ҚР БҒМ, Тіл комитеті, Тіл білімі институтының ресми қолдауымен жасалған.	– / Qazcorpus.kz, Atau.kz, Termincom.kz	www.qazcorpus.kz, www.atau.kz, www.termincom.kz
Мақсаты	Тілдің жаһандық қолданылуын кеңейту, лингвистикалық зерттеулерді қолдау.	Ұлттық тіл саясатын жүзеге асыру, тіл үйрету, терминологияны жүйелеу.	Oxford, COCA / Sozdik.kz, Lugat.kz, Atau.kz	www.oed.com, www.english-corpora.org / www.lugat.kz, www.atau.kz

Жалпылай алғанда, шетелдік сөздіктердің артықшылықтары – мәліметтер қорының ауқымдылығы, интерактивтілік деңгейінің жоғарылығы, жүйелі корпусқа сүйенген құрылымдылығы, айтылым мен тыңдалым мүмкіндіктерінің кеңдігі. Ал отандық сөздіктердің ерекшелігі – ұлттық тілді сақтау мен насихаттауға бағытталуы, мемлекеттік қолдау аясында жүзеге асуы, қазақ тілін үйренушілерге бейімделгені. Дегенмен, қазіргі таңда отандық электрондық лексикографияда да жаңа серпін байқалады. Жоғарыда атап өткен платформалар тілді құрылымдық деңгейде зерттеу және мәтіндер базасын жүйелеу жолында маңызды қадамдар жасауда. Мұның өзі болашақта шетелдік тәжірибелермен тең келетін, тіпті өзіндік ерекшелігімен дараланатын заманауи қазақ электронды сөздіктерінің пайда болуына негіз болады.

Қазіргі таңда компьютерлік лексикография – ақпараттық технологиялармен тығыз байланыста дамып келе жатқан тіл білімінің өзекті саласына айналды. Қазақ тіліндегі электронды сөздіктер тілдің заманауи қолданыс өрісін кеңейтіп, лексикографиялық ресурстардың қолжетімділігін арттыруда үлкен рөл атқаруда. Шетелдік тәжірибемен салыстырғанда, отандық электронды сөздіктердің әлеуеті жоғары болғанымен, оларды

жетілдіру мен сапалық тұрғыдан дамыту қажеттілігі туындап отыр. Осыған байланысты төмендегідей нақты ұсыныстарды атай кетуге болады:

1. Корпустық лингвистика негізінде сөздік базасын жетілдіру: Қазақ тілінің толық мәтіндер корпусы кеңейтіліп, электронды сөздіктермен интеграцияланған жағдайда, сөздердің қолданыс жиілігі, контекстік үлгілері мен стильдік ерекшеліктері ғылыми негізде ұсынылатын болады.

2. Мультимедиялық элементтерді енгізу: Электронды сөздіктерге сөздердің айтылым үлгілері, иллюстрациялық суреттер мен грамматикалық диаграммалар қосылуы олардың практикалық құндылығын арттырады. Бұл әсіресе қазақ тілін үйренушілер үшін тиімді болмақ.

3. Мобильді қосымшалар нұсқаларын жетілдіру: sozdik.kz, emle.kz секілді платформаларға арналған мобильді қосымшаларға оффлайн режим, дауыспен іздеу және оқу тарихын сақтау сияқты заманауи функциялар енгізілуі қажет.

4. Салалық бағыттағы электронды сөздіктерді көбейту: Медицина, құқық, техника, педагогика, өнер салаларында терминологиялық сөздіктерді электронды форматта әзірлеу мамандарға да, студенттерге де үлкен көмек болмақ.

Жалпы алғанда компьютерлік лексикография – тілдік мұраны сақтау мен жаңғыртудың тиімді жолы. Бұл салада жасалып жатқан әрбір жаңа қадам қазақ тілінің ғылым тіліне, жаһандық ақпараттық кеңістікке бейімделуіне және болашақ ұрпаққа қолжетімді әрі сапалы тілдік ресурстар ұсынуға мүмкіндік береді.

Пайдаланылған әдебиеттер:

1. *https://www.oxfordlearnersdictionaries.com/definition/english/lexicography*
2. *Қазақ әдеби тілінің сөздігі. Он бес томдық. 10-том. /Құраст.: М.Малбақов, Н.Оңғарбаева, А.Үдербаев және т.б. - Алматы, 2011. -752 б.*
3. *Жұбанов А.Қ. 3 Компьютерлік лингвистикаға кіріспе: Оқу құралы. Алматы: Қазақ университеті, 2007.*
4. *Лингвистикалық зерттеулердегі компьютерлік техно- логиялар: оқу құралы / Г.Б. Мәдиева, С.Б. Бектемірова, Ж.Ж. Кү зем бе ко ва. - Стер. бас. - Алматы: Қазақ универ- ситеті, 2020. - 119 б.*

ОӘЖ 378.1

SMART БІЛІМ БЕРУ ТЕХНОЛОГИЯЛАРЫНЫҢ НЕГІЗГІ ҚАҒИДАЛАРЫ

Айтбенбетова А.К., Молдабаева З.М,
Имангазина А.А., Култанбаева Н.К.,
aitbenbetova74@mail.ru
Қазақстан Республикасы
Қ.Жұбанов атындағы Ақтөбе өңірлік университеті

Аннотация: Мақалада SMART - оқытудың негізгі идеясы мен адамды өмірге және жұмысқа дайындық жасауға мүмкіндік беретін білім қоғамы туралы әдістемелік тұрғыдан қарастырылады. SMART оқытуда білімді жетілдіру керек екендігін өзіңізге таңдауға мүмкіндік беретін жан-жақты талданады.

Кілт сөздер: *SMART технологиясы, ақылды оқыту, SMART қала, SMART фон, SMART құрылғы, SMART қондырғы, SMART орта.*

Осы күнге дейін мұғалім оқушы үшін білім алудың жалғыз көзі болып келді және ол қосымша мағлұматтарды кітапханадағы кітаптардан немесе тек нақты бөлінген күндер мен сағаттарда теледидардан сирек көрсететін ғылыми-танымдық бағдарламалардан қосымша ақпарат ретінде алды. Қазіргі уақытта білім тек мұғалімнен оқушыға ғана емес, сондай-ақ оқушылар арасында да әртүрлі тәсілдермен беріліп жүр, яғни ол білімнің жаңа деңгейін құруға мүмкіндік беріп отыр. Оқытуда жаңа білім беру технологиялары белсенді түрде қолданыла бастады, ал мұғалімдер білімдерін тек сыныптарға ғана емес, сонымен қатар өз әріптестерімен де бөлісуге қабілетті деңгейге жетті. Өзара оқыту қазіргі уақытта мұғалімдердің педагогикалық дағдыларын жетілдіру жүйесіндегі жаңа кезең болып табылады және бұл біз үйренген тәлімгерлік емес, бұл өзара тиімді ынтымақтастық.

Интернет оқушы үшін негізгі білім көзі болатын уақыт та алыс емес. Технологиялар жеке бағдарланатын және жаңа білімді қалыптастыруға бағытталатын болады. Оқу процесі кез-келген бағытта білім объектілерінің өзгерістерін қамтиды және қазіргі заманғы мұғалім мұндай өзгерістерге дайын болуы керек және ол өзінің тәжірибесін әріптестеріне оқыту үшін кәсіби дағдылары мықты болуы керек. Осы мақсатта Смарт-оқыту - әртүрлі облыстар мен аудандардағы жарыстар, семинарлар, біліктілікті арттыру курстарына және т.б. қатысу үшін қозғалыс проблемаларын шешуді ұсынады. «Электронды университет» халықаралық консорциумының президенті В.П. Тихомировтың айтуы бойынша интернетті және жаңа білім жасайтын арнайы

дайындалған адамдарды талап ететін Smart - бұл қоғам дамуының жаңа парадигмасы [3]. Мұндай оқыту кәсіптік өсу траекториясын өз бетімен дамытуға, қалалық және ауыл мектептері мұғалімдерінің білім деңгейін теңестіруге, халықаралық білім кеңістігіне шығуға мүмкіндік береді.

SMART технологиясы ұғымы, яғни ақылды оқыту пайда болды және бүгінде Smart қала, Smart фон, Smart құрылғы, Smart қондырғы, Smart орта деген тіркестерді жиі естіп жүрміз, яғни, E-learning орнына SMART-learning келуі, ақпараттық қоғам орнына «smart қоғам»-ның қалыптасуы біздің елімізде де күтілуде. Қазіргі таңда Smart технологиясын өндірушілер – АҚШ, Канада, Оңтүстік Корея елдері [5,6]. Қазір елордамызда да Smart мектептер қалыптасып келеді.

Smart білім беру тұжырымдамасы – үлкен көлемдегі әр түрлі білім ресурстарының (аудио, видео, графика) тыңдаушыға қажетінше оңай әрі жылдам, ыңғайлы түрде жеткізілуі, сонымен қатар Smart білім беру оңай басқарылатын және әрдайым сыртқы білім ресурстарымен толығып отыруы [2].

Smart education немесе ақылды білім беру – бұл ыңғайлы түрде интерактивті ортада жүзеге асырылған әлемдік деңгейдегі қолжетімді контент арқылы білім беру болып табылады. Smart education түсінігінің негізі ол кең ауқымды қол жетімді білім ресурстары [3].

Smart education барлық білім беру үдерістерін және осы үдерісте қолданылатын барлық әдістер мен технологияларды түгелдей жаңаруын жүзеге асырады және де ақылды тақта, ақылды экран, кез-келген жерден интернетке қосылу сияқты жаңа технологиялардың туындауына себеп болады. Осы пайда болған жаңа технологиялар контентті жасақтап, жүйелеп, пайдаланушыға жеткізеді. Соның арқасында білім беруді жүзеге асыру тек сыныпта ғана емес, кез-келген жерде: үйде немесе мұражайда, кафе сияқты қоғамдық жерлерде де мүмкінке ие болады. Яғни білім беру үдерісінің кез келген қатысушысы (оқушы немесе мұғалім) Интернет желісі арқылы ортақ контентті белсенді пайдалана алады. Бұл өз кезегінде білім беру үдерісінің уақытқа немесе орынға (сынып, аудитория) тәуелділігін жояды.

«SMART» сөзі ағылшын тілінен аударғанда «Ақылды» дегенді білдіреді. SMART критерийлеріне талдау жасар болсақ, «Ақылды мақсаттарға» қойылатын негізгі талаптар төмендегідей:

S - Self-directed (өзін-өзі зерттеуге бағытталған)

M - Motivated (белсенді танымдық белсенділікті ынталандырады)

A - Adaptive (білім беру пәніне бейімделген)

R - Resource Free (білім беру ресурстарына еркін қол жеткізе алады)

T - Technology Embedded (технологиямен қамтамасыз етілген).

Smart-оқыту мақсаты: XXI ғасырдағы адам үшін қажетті құндылықтар, білім мен дағдыларға негізделген электрондық ортадағы мұғалімнің кәсіби және ақпараттық құзыреттіліктерін дамытуға жәрдемдесу [4].

Smart-оқытудың негізгі идеясы – ол адамды өмірге және жұмысқа дайындық жасауға мүмкіндік беретін білім қоғамы. Smart-оқыту сізге қандай салада білімді жетілдіру керек екендігін өзіңізге таңдауға мүмкіндік береді.

Дегенмен, мұғалімдер бүгінгі таңда Smart-оқытуға қаншалықты дайын және ауылдық жерлерде қаншалықты қол жетімді деген сұрақ туындайды:?

Мұғалімдерді оқыту орталықтары тренерлері осы оқыту технологиясын меңгеруге қалай көмектесе алады? Smart-оқыту білімнің жаңартылған мазмұнына мұғалімдер үшін қаншалықты сәйкес келеді?

2020 жылы Қазақстан Республикасында білім беруді дамытудың 2011-2020 жылдарға арналған мемлекеттік бағдарламасы жаңа ұлттық көрмені ұсынатын болады: 2020 жылы Қазақстан - білімді ел, ақылды экономика және жоғары білікті кадрлар бағдарламасына көшпек. «Ақылды» smart-технологиялар «ақылды» үйлерде, кеңселерде, өнеркәсіптік қондырғыларда және т.б. қолданылатын болады. Республикамыздың Елордасында мемлекеттік білім басқару салаларында ақпараттық ақылды технологиялар бойынша белсенді жұмыстар жүргізілетін болады. Ол мұғалімдер мен тәрбиешілер үшін, сондай-ақ бүкіл мемлекет, білім беру мен басқару саласындағы жаңа білім беру мүмкіндіктері мен экономика саласындағы білім беру үшін ең өзекті мәселелердің бірі болмақ.

Қазіргі уақытта жаңа ақпаратты байланысқа және аудиовизуалды технологияларға көбірек қызықтыру ұсынылып отыр. Зияткерлік технология - оқытудың жеңілдетілген әдісі болып табылады. Ол сабақта немесе аудиторияда шағын сынып оқушыларымен жұмыс істеуге, сондай-ақ әртүрлі анықтамалық және электрондық кітаптарға қол жеткізуге мүмкіндік беретін қарапайым процесс. Зияткерлік технологияларды пайдалану дәрісті неғұрлым серпінді етеді, мотивацияны ынталандырады, сондай-ақ мұғалімге, қоғамға сәйкес сапалы білім алуға мүмкіндік береді. Интерактивті оқыту технологияларын пайдалану интерактивті тақталарды, проекторларды, компьютерлерді және бағдарламалық қамтамасыз етуді үнемдеуді қамтиды. Бұл кешенді пайдалану дәстүрлі мультимедиалық мазмұнды реттеуге мүмкіндік береді.

Оқушыларды оқытуда Smart - технологияларды қолдану мүмкін және қол етімді. Бірақ кез-келген жаңа технология қолданылған кезде ғана тиімді.

Smart Education білімге деген көзқарасты желі ресурстарын тиімді пайдалану үшін маңызды жұмыспен және практикалық қадамдармен ауыстырыуға тиіс.

Оқыту процесін жаңарту үшін ақпараттық технологиялардың жаңа мүмкіндіктерін енгізу оқушылар мен мұғалімдер арасында жаңа білімге қызығушылықты арттырып қана қоймай, олар жаһандық қоғамдастықта сәтті қарым-қатынаста болуға мүмкіндік береді және сонымен бірге оқытуды әлдеқайда тиімді етуге және қазіргі заманғы экономика талабына сай өмір бойы жұмыс дағдыларын жақсарту қырларын ашады.

Пайдаланған әдебиеттер:

1. "Smart - қоғам жолындағы әлем" халықаралық білім беру форумының материалдары.Алматы, 2022;.
2. Бузаубакова К.Д., К.Д.,Амирова А.С., Маковецкая А.А. Цифрлы педагогика: Оқулық. –Тараз: «ИП «Бейсенбекова Ә.Ж.», 2022. –314 б.
3. Тихомиров В. П. Мир на пути Smart education. Новые возможности для развития // Открытое образование. –2011. –№ 3. –С. 22-28.
4. Матасова О. STEAM-образование внедряет сельская школа в Акмолинской области, 18 ноября 2018 [Online].Available: https://www.inform.kz/ru/steam-obrazovanie-vnedryaet-sel-skaya-shkola-v-akmolinskoy-oblasti_ a3459352 [Accessed 11 февраль 2019].
5. Рабинович П. Д., Заведенский К. Е., Кушнир М. Э., Храмов Ю. Е., Мелик-Парсаданов А. Р. Цифровая трансформация образования: от изменения средств к развитию деятельности // Информатика и образование. –2020. –№ 5 (314). – С. 4-14.

COMPUTER LINGUISTICS AND ITS IMPORTANCE TODAY

Askarova Umidakhon,
Lecturer of the National University of Uzbekistan
samina090822@gmail.com
NUU

Abstract. If we say that the 21st century is the age of technology, then computer technologies, like technologies in all spheres, are developing even more. As a result, the demand for computers naturally increases day by day. It is precisely due to this need that the science of computational linguistics arose as a result of the consistent connection of language and computers.

In this research work, we also reflected on the importance of the field of computational linguistics in the formation of general linguistic foundations, the improvement of relations between languages, bringing the Uzbek language to the level of world languages, as well as the wide opportunities provided to us through computational linguistics.

Annotatsiya. XXI asr – texnika asri der ekanmiz, barcha sohadagi texnologiyalar kabi kompyuter texnologiyalari ham yanada rivojlanib bormoqda. Buning natijasida esa, kompyuterga bo‘lgan talab kundan-kunga tabiyki oshib boradi. Aynan mana shu ehtiyoj tufayli til va kompyuterni izchil bog‘lanishi natijasida kompyuter lingvistikasi fani paydo bo‘ldi.

Biz ham ushbu tadqiqot ishimizda, Kompyuter lingvistikasi sohasining umumtil asoslarini shakllantirishda, tillar o‘rtasidagi munosabatlarni takomillashtirishda, o‘zbek tilini dunyoviy tillar darajasiga olib chiqishdagi ahamiyati hamda kompyuter lingvistikasi orqali bizga berilayotgan keng imkoniyatlar to‘g‘risida fikr yuritganmiz.

Аннотация. XXI век - век техники, компьютерные технологии, как и технологии во всех областях, продолжают развиваться. В результате спрос на компьютеры естественным образом растет день ото дня. Именно из-за этой потребности последовательное взаимодействие языка и компьютера привело к появлению компьютерной лингвистики.

В данной исследовательской работе мы также рассмотрели важность компьютерной лингвистики в формировании общеязыковых основ, совершенствовании отношений между языками, поднятии узбекского языка на уровень мировых языков, а также широкие возможности, предоставляемые нам компьютерной лингвистикой.

Keywords: *computer, computational linguistics, technology, language, text, automatic editing, translation.*

As we all know, it is no exaggeration to say that raising the Uzbek language to the level of world languages is a vital necessity in the process of globalization taking place today on a global scale. Language is a means of understanding reality, having a certain idea of reality. At the same time, language is the foundation of innovations, inventions, and discoveries in all fields.

The social functions of language are expanding year by year. As a clear result of this, we can cite the field of Computational Linguistics. Computational linguistics arose as a result of the application of computer system capabilities in the field of linguistics. Computational linguistics occupies a special place in the formation of the foundations of a common language, in the improvement of relations between languages. [Computer Linguistics, Study Guide: 4]

Currently, the process of computerization is taking place at different levels in countries around the world. Regardless of the level of development, humanity has come to understand the incomparable role of information technologies in society. Computerization of all spheres of human activity is an important task of society today and a factor of social development. Without computerizing society, it is difficult to implement social changes and ensure a decent standard of living. [Computer Linguistics, Study Guide: 13] Computer linguistics involves increasing the level of computer literacy, teaching the user to think coherently, logically, and correctly, acquiring comprehensive knowledge about the functional scope of language, teaching languages, assessing knowledge, editing and analyzing texts, translating texts that are equally necessary for everyone today from one language to another, as well as forming theoretical knowledge and, of course, developing skills for practical application in certain areas. In general, *Computer linguistics is the effective use of computers and the determination of ways to solve linguistic problems using computer tools and the determination of ways to solve them using computer tools, the formation of theoretical knowledge and certain directions facilitates the development of practical application skills for the*

In the statistics of linguistic data, in the creation of dictionaries of various content, computer linguistics plays an important role. This science has a positive impact not only on linguistics, but also on the development of other fields. In particular, automatic editing (BORIS system), the translation process can be used in the analysis and translation of literary texts. [Computer Linguistics, Study Guide: 18]

At a time when the computer has become a part of our lives, specialists and linguists are conducting scientific research on the creation of mathematical and computer models of the Uzbek language, editing texts in the Uzbek language, and developing English-Uzbek or Russian-Uzbek computer translation programs. Issues

related to linguistics, such as language teaching, text editing, and translation from one language to another, are being consistently studied.

It is no secret that the aforementioned conveniences contribute to the development of negative consequences, such as computer addiction. This opinion is also reflected in the following sentences by Kh.E. Rakhmanov, J.T. Sunatov: It can be seen on social sites that becoming a slave to this helper is also reflected in the language, which is the soul and expression of the soul of a person. Although this is not among the main tasks of computational linguistics in the research on computational linguistics that we are currently observing, there is no doubt that this problem will be posed as the main issue at the intersection of psycholinguistics and computational linguistics in the near future. [5]

The development of computer linguistics, the creation of programs for various purposes, serves to ensure the development of other areas of linguistics. Computational linguistics also helps to eliminate cases of violation of orthographic norms and non-grammatical expressions.

However, as they say, there are two sides to a coin, one might wonder if these aforementioned conveniences will lead to negative consequences for some users. Because when we ask students to write flawlessly during university classes, it's heartbreaking to hear some students say, “We don't need to write spelling or stylistically correctly, there are programs on computers that can correct text” We must remember one thing: no matter how wisely we use computers, it certainly shouldn't harm our literacy. Whether to use a computer wisely or become dependent on it also depends on the person themselves. We have often heard that the internet is likened to a knife. In my opinion, a computer can also be compared to a knife based on the aforementioned positive and negative consequences. Because, like a knife, if we use it wisely, it won't harm us. Otherwise, damage is inevitable.

In conclusion, it can be said that the creation of a computer style of the Uzbek language is of great importance in our life and development. A computer style is created in the union of computer linguists and philologists. This is precisely the positive aspect of computational linguistics: it unites philologists and computer specialists.

Despite the fact that computational linguistics is a new direction, we can confidently say that if scientific research is carried out in this direction, as well as personnel training, and in general, if this direction develops further, it will become a large-scale field in the future, and a number of directions and areas will appear on its basis. Because the computer style, in accordance with the requirements of world standards, is clear, concise, clear to everyone, free from various artistic means of expression, strictly takes into account computer requirements and is adapted for computer processing. We can come to the general conclusion that without creating

this style, we will not be able to fully transition to conducting business in our country in our own language, that is, in the Uzbek language.

REFERENCES

1. Abdurakhmonova N., Alisher I., Sayfulleyeva R. MorphUz: Morphological Analyzer for the Uzbek Language. In 2022, the 7th International Conference on Computer Science and Engineering (UBMK) (pp. 61-66). IEEE. 2022, September.

2. Abdurakhmonova N., Tuliyev U., Ismailov A., Abduvahobov G. Uzbek electronic corpus as a tool for linguistic analysis. In Computer processing of Turkic languages. TURKLANG 2022 (pp. 231-240). 2022.

3. Abdurakhmonova N., Abduvahobov G. Theoretical and Methodological Foundations of Compiling an Educational Dictionary. International Journal of the Art of Words, 4 (6). 2021.

4. Hamrayeva H. Language and style of Ozod Sharafiddinov's journalism, dissertation submitted for the degree of Candidate of Philological Sciences, abstract, Tashkent, 2012.

5. Rakhmanov Kh.E., Sunatov J.T. Scientific research conducted in the field of computer linguistics of the Uzbek language, article “Computer linguistics: problems, solutions, prospects” International scientific-practical conference, Tashkent, 2022.

6. Kholmonova Z.T. Computer Linguistics textbook, Tashkent, 2019.

FIZIOLOGIK HOLAT FE’LLARINING INGLIZCHA-O‘ZBEKCHA TARJIMA LUG‘ATLARDAGI TALQINI

Fozilov Pulat Shamilovich,
mustaqil tadqiqotchi
ToshDO‘TAU

Annotatsiya. Maqolada ingliz va o‘zbek tillaridagi fiziologik holat fe’llarining tarjima lug‘atlardagi talqini masalalari yoritilgan. Ikki tarjima lug‘atlarida mazkur fe’llarni aks ettirishdagi leksikografik tamoyillar, so‘zlarni izohlash prinsiplari, leksik-semantik tizimi, grammatik tavsiflari qiyosan o‘rganilgan.

Kalit so‘zlar. *amaliy leksikografiya, metaleksikografiya, leksikografik tamoyil, diaxron va sinxron aspekt, leksik va fonetik ko‘rsatkich, grammatik tavsif, leksik-semantik tizim, semantik ekvivalent.*

Ikki tilli leksikografik manbalar ming yillar ilgari paydo bo‘lganligiga qaramay, uning nazariy jihatlarini o‘rganadigan soha – bilingual metaleksikografiya, ya’ni nazariy tarjima leksikografiyasi yaqin o‘tmishda shakllandi. Tarjima lug‘atining asosiy vazifasi foydalanuvchining til bilish darajasini oshirish, unda mavjud bilim bazasini leksik jihatdan boyitishdir [1:147-159]. Bunday lug‘atlar birlamchi til o‘rganish manbasi bo‘lmasa ham, tilni o‘zlashtirish jarayonda asosiy vositalardan biriga aylanadi. Bugungi kunda tarjima lug‘atlari asosan ikki xil shaklda tuziladi. Birinchisi alifbo tartibi asosidagi universal lug‘at bo‘lsa, ikkinchisi bir yirik ma’no yoki grammatik turni qamraydigan ideografik lug‘atlardir. Til o‘rganuvchilar ko‘p hollarda birinchi turdagi manbalarni ma’qul ko‘rishadi. Yirik hajmdagi lug‘at tilni chuqur o‘rganuvchilar ehtiyojini qondirsa, tor qamrovli manba tilni eng zaruriy leksik bazani shakllantiradi. Inglizcha-o‘zbekcha tarjima lug‘atlari aksariyat hollarda alifbo tartibida, universal qamrovga ega bo‘ladi. Ularning strukturasi leksema, grammatik parametr, transkripsiya va tarjima ekvivalenti elementlaridan tuzilgan bo‘ladi. Ba’zan tarjimalar misollar bilan dalillanishi mumkin. Fe’l tarjimasiga bag‘ishlangan lug‘at maqolalarda misollar fe’l+predlog ma’nolarini ochib berish uchun kiritiladi. Shuningdek, so‘zlarning yasama shakllariga ham e’tibor qaratiladi. Masalan, eng yirik tarjima lug‘atlaridan hisoblangan Sh.Bo‘tayev, A.Irisqulovlar hammuallifligidagi manba leksemalarni quyidagicha leksikografik talqin qiladi [2;217]:

breath [bre(-)] n nafas; **to draw** ~ nafas olmoq;

to draw the first ~ tug‘ilmoq; dunyoga kelmoq.

breathe [bri:d] v nafas chiqarmoq; nafas olmoq.

to ~ again – yengillik sezmoq.

Breath va *to breathe* leksemasi eng asosiy fiziologik holat, *nafas olishni* ifodalaydi. Shuning uchun bu holat fe’lining predlog bilan uyg‘unlashuvidan hosil bo‘lgan ma’no ham e’tiborga olinadi. Shunday qilib, o‘quvchi bu so‘zlar *nafas olmoq, tug‘ilmoq, nafas chiqarmoq* hamda *yengillik sezmoq* holatlarini anglatishini biladi. So‘zning turli birikuvdagi barcha ma’nolari aynan fiziologik holatni ifodalashi e’tiborga loyiq.

Xuddi shu lug‘atda boshqa fiziologik holat fe’llari predloglar bilan turg‘un birikma hosil qilmagani sababli faqat sinonimli tarjimalar bilan berilgan [2; 66]:

exhale [eks´heɪl] *v* 1) nafas chiqarmoq; 2) bug‘lanmoq, bug‘ga aylanmoq.

inhale [ɪn´heɪl] *v* nafas olmoq, ichiga tortmoq.

Exhale so‘zining aynan o‘zidek bir so‘z qolipidagi ekvivalenti yo‘q, shuning uchun birikma shaklida tarjima qilingan, *inhale* ham shu sababdan tarjimada birikmaga aylangan. *Exhale* so‘zining mexanik semasi bug‘lanish bilan bog‘liq. Shu o‘rinda ikki ma’no voqelanishini ko‘rsatadigan misol qo‘shish maqsadga muvofiq bo‘lar edi. Fiziologik holat fe’llarining bir ulushi egalik semasi atrofida jamlanadi. Ularning izohida shu borlik, egalik semalari yaqqol aks etishi kerak. Tarjimada esa ma’nodoshlarning farqlariga urg‘u berish maqsadga muvofiq bo‘ladi. Masalan, *to own, to possess, to have* va *to occupy* ma’no jihatdan bir-biriga yaqin. Ular tarjima lug‘atda quyidagicha talqin qilingan [2;128].

have [hæv] *v* (**had; had**) ega bo‘lmoq; **to ~ a walk** sayr qilmoq; **to ~ a dinner** tushlik qilmoq; **to ~ a pleasant time** vaqtni yaxshi o‘tkazmoq; **to ~ a headache** boshi og‘rimoq.

own [əun] **1.** *v* egallamoq, ega bo‘lmoq; **2.** *a* o‘ziniki.

possess [pə´zes] *v* 1) egalik qilmoq, ega bo‘lmoq, qo‘lga kiritmoq, zabt etmoq; 2) bosmoq (*biror hissiyot, tuyg‘u haqida*); **you are surely ~ed** siz aqldan ozibsiz.

To have bu ma’nodoshlar ichida nutqda eng faol qo‘llanadigan biridir. Zero, uning ma’no ko‘lami faqat egalik bilan cheklanmaydi. Bu ma’nolar esa nutqiy konteksta to‘liq voqelanadi. Shu kontekstlarga ko‘ra *to have* fe’li egalik ma’nosidan tashqari *tanovvul qilmoq, vaqt o‘tkizmoq* kabi sof nutqiy semalar kasb etadi. Bu so‘zning ma’nodoshi esa *to own* – ma’nosi bir sema doirasida cheklangan. Shuning

uchun yana kontekstual birikma misollarni ilova qilish shart emas. *To possess* esa sof lug‘aviy polisemiyani namoyon qiladi. Uning *aqldan ozmoq, egalik qilmoq, hissiyotni bosmoq* ma’nolari bir-biridan keskin farq qiladi. Lekin shu farqlar misollar bilan yaqqol dalillanishi lozim.

Tarjima lug‘atlarini tuzishda mualliflar qator muammolarga duch kelishadi. Birinchi muammo asosan ikki har xil tizimdagi tillar orasida bo‘lib, ularda bir-biriga semantik nomutanosib so‘zlar ko‘p uchraydi. Ularni tarjima qilish imkonsiz, ya’ni ma’nolarni to‘liq saqlab qolib bo‘lmaydi. Tilshunoslikda *lingvomadaniy realiya* deb ataluvchi betakror leksik guruh faqat izoh shaklida o‘girilishi mumkin. Biroq realiyalar faqat nomlar sirasida bo‘ladi, harakat yoki holat ichida realiya uchramaydi. Shunday ekan, bu vajni fiziologik holat fe’llari talqinidagi muammo sifatida sanamaymiz.

freeze [fri:z] *v.* muzlab qolmoq; sovuq yemoq, sovqotmoq, lo'ngmoq; muzlatmoq, yaxlatmoq, muzda qotirmoq, to'ndirmoq. **freezer** *n.* muzlatgich, muzxona.

Lug‘at talqinidagi yana bir muammo polisemik so‘zlardir. Fe’llarning aksariyati ko‘p ma’nolarga ega. Demak, ular talqini barcha asosiy ma’nolar tarjimasi bilan boyitish lozim. Bunday semalar kontekstual bo‘lgani uchun ularni misollar bilan dalillash talab etiladi. Masalan, tarjima lug‘atlarida ko‘p ma’noli fiziologik holat kontekstual jihatdan polisemik bo‘lganda shunchaki ma’nolarni ajratmasdan bir guruhda sanab o‘tish holati mavjud. Bu esa lug‘atning amaliy foydasini susaytiradi, o‘quvchini chalg‘itadi [3;88]:

Navbatdagi muammo tillar orasidagi madaniy farqqa borib taqaladi. Qardosh tillar kesimida bunday masalani uchratmaymiz. Biroq boshqa tillar, xususan, ingliz tili bilan tafovutlarimiz ko‘p. Bu esa, o‘z navbatida, lug‘atchilar uchun talqin mushkulotini keltirib chiqadi. Fiziologik holat *inson tanasida kechuvchi tabiiy jarayonni, tana tabiiy shamoyili va holini* bildiradi. Insonlarning bu jihati esa millat yoki madaniyatga qarab farqlanmaydi. Demak, ayni leksik-semantik guruh talqinida madaniy muammolarni chiqarib tashlash mumkin.

Maxsus atamalar ham muammoli vaziyat. Atamalar talqini barcha lug‘atlar uchun standart bo‘lishi lozim. Nostandart yondashuv aynan atama borasida kuzatilsa, bu lug‘atning originalligini emas, ilmiy kompetensiyasi sust ekanini anglatadi. Bundan tashqari, til doim taraqqiy etuvchi jonli organizm bo‘lganligi

sabab, leksik sath ham doim o‘zgarishda bo‘ladi. So‘zlar vaqt davomida yangi ma’nolar kasb etishi, mavjud ma’nolari eskirishi tabiiy. Bu kabi muammoni hal qilish uchun lug‘at tuzuvchilari har ikki tilni mukammal egallagan, milliy-madaniy kontekstini to‘liq anglagan bo‘lishi lozim. Muayyan o‘quvchilar auditoriyasiga lug‘atni moslash ham dolzarb vazifalardan hisoblanadi.

Dolzarb muammolardan yana biri tarjimada sinonimlarga murojaat etishdan kelib chiqadi. Ko‘pincha talqinda bir necha sinonimlarni sanab o‘tib hech qanday misollar bilan mustahkamlamaslik holatlari mavjud. Shunda o‘quvchi sinonimlardan qaysi biri asos leksemaga yaqinligini bilolmaydi, har birining o‘zaro farqi bor, nutqda qo‘llash o‘ziga xosligi mavjud. Shunday ekan, sinonimlarni berishda aniq tamoyillarni shakillantirish muhim. So‘zlar ma’no yaqinligiga ko‘ra tartiblanishi, har biri qisqa birikma shaklida bo‘lsa ham misol bilan dalillash zarurdir.

Til ijtimoiy hodisa ekanligi uning iqtisodiy, siyosiy, tarixiy jarayonlar bilan chambarchas bog‘liq ekanida ko‘zga tashlanadi. Global dunyoda xalqaro aloqalarning uzviylashib, jadallashib borayotgani esa tarjima lug‘atlariga ehtiyoj kuchaytirmoqda. Ularni onlayn platformaga o‘tkazish talab etilmoqda. Sun’iy intellektga tilni sintez qilish, onlayn muharrir, tarjimon dasturlari uchun ham keng qamrovli mukammal tarjima lug‘atlari asos qilib olinadi.

Leksikografik manba so‘zlarning grammatik, semantik, uslubiy xususiyatlarini aniq ko‘rsatishi, ularning nutqiy parametrlarini o‘z ichiga olishi shart. Jarayon so‘zlarni tizimlashtirishdan boshlanadi. Ma’no va grammatik ko‘rsatkuchlar dolzarbligiga ko‘ra tartiblanadi. Misollar so‘zning uslubiy xarakterini ham ko‘rsatishi kerak. Murakkab, o‘xshashi yo‘q so‘zlar esa izohli tarjima qilinishi ham mumkin. Tarjima lug‘atlari mavzulashtirilgan yoki grammatik jihatga urg‘u beradigan turlarga ajraladi. Fiziologik holat fe’llarini talqin qilishda mana shu talablarga rioya etib mukammal mavzulashtirilgan lug‘at yaratish mumkin. Bunday lug‘at tarkibi quyidagicha bo‘lishi mumkin:

1. Leksema
2. Talaffuz me’yori
3. Grammatik ko‘rsatkichlar
4. Tarjima (zarurat bo‘lsa izohli tarjima)
5. Birikmalar
6. Gap shaklidagi misollar
7. Ma’nodoshlar
8. Uslubiy va chastota ko‘rsatlichlari.

COUGH [kɒf] *v.* yo‘talmoq: *The smoke made me cough. I* ***coughed*** *all night long.* Neutral style – B2.

EXIST [ɪgˈzɪst] *v.* **1.** mavjud bo‘lmoq: *I don't think ghosts **exist**. Poverty still **exists** in this country.* SYNONYM: to be extant. **2.** Arang yashamoq: *Some species **exist** in this small area of forest and nowhere else on earth. to be alive Few people can **exist** without water for more than a week.* SYNONYM: to survive. Neutral style – B2.

SNEEZE [sniːz] *v.* aksirmoq*: My* nose *is* tickling*, I* think *I'm going to **sneeze**. I* keep ***sneezing*** *- I* hope *I'm not getting a* cold*. This* pepper *is making me **sneeze**.* Neutral style – B2. Bu lug‘at shakli har bir ma’noni misol bilan dalillagani sabab sinonimlar qatori ko‘rinishidagi tarjimani mukammallashtirish, boyitish yo‘llarini ko‘rsatadi. Uslub ko‘rsatkichi qaysi doirada qo‘llash haqida tavsiya bersa, B1-B2 va hokazo chastota belgilari til bilishning qaysi darajasida bu so‘zni o‘rganish talab etilishini ko‘rsatadi.

Adabiyotlar ro‘yxati:

1. Zgusta, L. Translational Equivalence in Bilingual Lexicography Bhukośyam, Dictionaries, vol. 9 (1987): 1-47 (also, abridged, in Hartmann, ed. 1984: 147-159).

2. Irisqulov A., Bo‘tayev Sh. Inglizcha-o‘zbekcha, o‘zbekcha-inglizcha lug‘at. – Toshkent: Ilm-ziyo, 2008.

3. Nazarov S., Nazarova F., Poltayeva D., Adizova N. English-uzbek dictionary. Inglizcha-o‘zbekcha lug‘at – Toshkent: Niso Poligraf, 2013.

VII SHO‘BA KOMPYUTER LEKSIKOGRAFIYASI

UDK: 811.511.141.004

O‘ZBEK TILI DIALEKTLARI IZOHLI LUG‘ATINI YARATISH MODEL VA ALGORITMI

Eshmurodova Madina Dilshod qizi,
Magistrant
madinadilshodovna313@gmail.com
Muhammad al-Xorazmiy nomidagi Toshkent Axborot Texnologiyalari Universiteti
Xudoykulova Nargiza Ravshanovna,
Magistrant
n.khudaykulova@newuu.uz
Muhammad Al-Xorazmiy nomidagi Toshkent Axborot Texnologiyalari Universiteti

Annotatsiya. Ushbu maqolada o‘zbek tilining turli dialektlaridagi so‘zlarni tizimli ravishda tahlil qilish va izohli lug‘at yaratish uchun zamonaviy algoritmlar va metodlarni ishlab chiqish masalasi ko‘riladi. Bunda dialektal farqlarni aniqlash va tushunish uchun lingvistik (fonetik, morfologik, sintaktik) va texnologik yondashuvlar (semantik tahlillar) qo‘llanilish ahamiyati ko‘rib o‘tiladi. Bunda o‘zbek tili dialektlarining o‘ziga xos xususiyatlarini saqlagan holda ularni adabiy tilga moslashtirish usullari, shuningdek, dialektlararo farqlarni avtomatik tarzda aniqlash va tasniflash uchun mashina o‘rganish va tabiiy tilni qayta ishlash (NLP) texnologiyalarini samarali qo‘llash masalalari o‘rganiladi.
Natijada, yaratilingan izohli lug‘at nafaqat ilmiy tahlil va tadqiqotlar uchun, balki til o‘rganish va dialektal farqlarni o‘rganishga qiziquvchilar uchun amaliy resurs sifatida ham foydalidir.

Абстракт. В данной статье рассматривается разработка современных алгоритмов и методов для систематического анализа слов из различных диалектов узбекского языка и создания толкового словаря. При этом подчёркивается важность применения как лингвистических (фонетических, морфологических, синтаксических), так и технологических подходов (семантический анализ) для выявления и понимания диалектных различий. Также изучаются методы адаптации диалектов к литературному языку с сохранением их особенностей, а также эффективное применение технологий машинного обучения и обработки естественного языка (NLP) для автоматического выявления и классификации диалектных различий.

В результате созданный толковый словарь будет полезен не только для научного анализа и исследований, но и как практический ресурс для изучающих язык и интересующихся диалектными различиями.

Abstract. This article discusses the development of modern algorithms and methods for the systematic analysis of words from different dialects of the Uzbek language and the creation of an explanatory dictionary. The importance of using both linguistic (phonetic, morphological, syntactic) and technological (semantic analysis) approaches to identify and understand dialectal differences is emphasised. In addition, methods of adapting dialects to the literary language while preserving their characteristics are analysed, as well as the effective use of machine learning and natural language processing (NLP) technologies for the automatic detection and classification of dialectal differences.
The resulting explanatory dictionary will be of use to the scientific community as well as to language learners and those interested in dialectal differences.

Kalit so‘zlar: *O‘zbek tili, dialektlar, sheva so‘zlar, izohli lug‘at, dialektologiya, tabiiy tilni qayta ishlash (NLP), mashina o‘rganish (ML), semantik tahlil, fonetik tahlil, morfologik tahlil, adabiy til, kompyuter lingvistikasi, raqamli lug‘at, algoritmlar, tizimli tasniflash, madaniy meros, til boyligi.*

O‘zbek tili, ayniqsa uning turli hududlar doirasidagi rang-barang dialektlari xalqimizning madaniy va tarixiy merosining ajralmas qismi hisoblanadi. Ularni izohli lug‘atda jamlash, saqlash va tarqatish orqali tilning turli shakllarini va lug‘at boyligini aniqlash, mazmunini tushunish imkonini beradi va o‘zbek tilini o‘rganayotganlar uchun dialektal farqlarni tushunishda qo‘llanma bo‘lib xizmat qiladi. Dialektlarni tahlil qilish va izohli lug‘at yaratishda zamonaviy algoritmlar va modellarni qo‘llash o‘zbek tilini raqamlashtirish va kompyuterlinguistikani rivojlantirishga imkon yaratadi. Bu, o‘z navbatida, matnni qayta ishlash, nutqni tanish, avtomatik tarjima kabi sohalarda samarali tizimlarni yaratishga yordam beradi. O‘zbek dialektal (sheva) so‘zlarni tahlil qilish, ularni izohlash va o‘xshashliklar va farqlarni aniqlashda sun’iy intellekt, mashinali o‘qitish (ML) va tabiiy tilini qayta ishlash (NLP) texnologiyalarini qo‘llash o‘zbek tili izohli lug‘atini dialektlari izohli lug‘ati uchun model va algoritm ishlab chiqish orqali bu jarayonni yanada samarali qilish mumkin. Bunda, dialektlarni izohli lug‘atga kiritish uchun avtomatik alifbo moslashuvi, semantik tahlil va morfologik strukturani aniqlashga qaratilgan algoritmlar ishlab chiqilishi kerak.

O‘zbek tili dialektlarini o‘rganish va ularning farqlarini aniqlash lingvistika sohasida muhim ahamiyat kasb etadi, dialektologiya, leksikologiya va morfologiya kabi tilshunoslik yo‘nalishlarida yangi dialektal birliklarni aniqlash, ularning

semantik xususiyatlarini tahlil qilish uchun izohli lug‘atlar va algoritmik modellar asosiy vosita bo‘lib xizmat qiladi.

Dialektal farqlar tahlili nafaqat o‘quvchilarga va balki o‘zbek tilini o‘rganayotgan xorijliklar uchun mo‘ljallangan ta’lim materiallari yaratishda katta ahamiyatga ega. Bu materiallar dialektlar o‘rtasidagi leksik, fonetik va morfologik tafovutlarni yoritib, til o‘rgatish jarayonini yanada samarali bo‘lishiga yordam beradi.

Dialektal tizimlar zamonaviy texnologiyalar bilan integratsiyalashganda u kompyuter lingvistikasi doirasida juda keng qo‘llanilishi mumkin, tabiiy tilni qayta ishlash (NLP) va mashina o‘rganish (Machine Learning) texnologiyalaridan foydalangan holda, turli dialektlar orasidagi farqlarni avtomatik aniqlash, tahlil qilish va tarjima qilish imkoniyati mavjud. Bu jarayon, ayniqsa, avtomatik tarjima va nutqni tanish tizimlarida muhim ahamiyat kasb etadi. Bir dialektda yozilgan matnni boshqa inson uchun tushunarli bo‘lgan dialektga yoki rasmiy adabiy tilda ifodalash uchun maxsus algoritmlar ishlab chiqilishi mumkin.

Nutqni tanish tizimlari uchun o‘zbek tilining turli hududiy shakllarini aniqlay oladigan modellar yaratish, tilning tabiatini chuqurroq o‘rganish va uni texnologik qurilmalarda to‘g‘ri tanish imkonini beradi.

Shuningdek, dialektlar o‘zbek xalqining boy madaniy merosini o‘zida mujassam etadi. Turli hududlardagi dialektal birliklar va ularning izohli hujjatli shakllari tarixiy-lingvistik tadqiqotlar uchun muhim manba bo‘lishi mumkin. Shu jihatdan, bu model madaniy merosni saqlash, hujjatlashtirish va kelajak avlodlarga yetkazish vositasi sifatida ham xizmat qiladi.

O‘zbek tili juda ko‘p dialektal shakllarga (shevalarga) ega. Har bir dialekt quyida ko‘rsatilgan o‘ziga xos fonetik, leksik, grammatik va semantik xususiyatlarga ega:

Dialekt /Sheva	Hudud	Xususiyati
Shimoliy (Samarqand)	Samarqand atrofida rivojlangan	O‘ziga xos fonetik va leksik o‘zgarishlar
Janubiy (Farg‘ona)	Farg‘ona vodiysi va Toshkent v.da mavjud	Ba'zi turkcha va forscha so‘zlar ko‘proq ishlatiladi
G‘arbiy (Buxoro)	Buxoro viloyatida tarqalgan	O‘ziga xos grammatik va fonetik xususiyatlar
Shimoliy-g‘arbiy (Xorazm)	Xorazm viloyatida rivojlangan	O‘ziga xos fonetik xususiyatlar, ba'zi so‘zlar boshqacha talaffuz qilinadi
Janubi-sharqiy (Surxondaryo)	Surxondaryo viloyatida mavjud	Janubiy turkiy til qoidalari ta'sirida fonetik va leksik o'zgarishlar bor
Qoraqalpoq	Qoraqalpog‘iston hududida mavjud	O‘ziga xos grammatik va fonetik farqlar
Toshkent (Markaziy)	Toshkent atrofida rivojlangan	O‘zbek tilining markaziy varianti

Shuning uchun, dialektlar o‘rtasidagi farqlarni aniqlash va tasniflash birinchi asosiy muammo bo‘lib, bu jarayonni ilmiy va texnik jihatdan to‘g‘ri tashkil etish masalasi mavjud. Buning uchgun, o‘zbek tili dialektlarining tarkibini aniq tasniflash va ularni turli xususiyatlar (fonetik, leksik, sintaktik, semantik) bo‘yicha tahlil qilish va so‘zlarning sinonimlari, antonimlari va semantik farqlarini ajratib ko‘rsatish talab etiladi. Dialektal so‘zlarni yagona izohli lug‘atda to‘plash va ularga tushuntirish berish, ularning ma’nolarini va ishlatilish kontekstini ta’riflash - bu yana bir katta muammo.Bunday kata hajmdagi ma’lumotlarni kitob holatida to‘plash hozirgi kun foydalanuvchilari uchun qiyinchilik tug‘dirishi mumkin, chunki bunda kerakli manbaani samarali topa olish va undan ma’lumot izlab topish ko‘p vaqt talab qiladi. Shuning uchun, lug‘at yaratish jarayonini avtomatlashtirish muhim ahamiyat kasb etadi, buning uchun mashinali o‘qitish (ML) va tabiiy tilni qayta ishlash (NLP) texnologiyalarining integratsiyasi zarur.

Yaratilish texnologiyalari:

Ma’lumot to‘plash va tahlil qilish:

- O‘zbek tilining dialektlarini aniqlash va to‘plash uchun tahlil qilish manbaalarini tanlash kerak. Bu manbaalar og‘zaki va yozma matnlar,

lug‘atlar, adabiy va ilmiy asarlar, xalqaro va mahalliy lug‘atlar bo‘lishi mumkin.

- Turli hududlardan (shaharlar, qishloqlar) olingan matnlar yoki yozma asarlarni tahlil qilish. Bu matnlar orqali o‘ziga xos so‘zlar, iboralar va grammatik shakllarni yig‘ish mumkin.
- Dialektal so‘zlar va ularning izohlarini jamlash uchun maxsus ma’lumotlar bazasi yaratish zarur. Bu bazaga dialektal so‘zlarning turli variantlarini va ularning semantik ma’nolarini kiritish mumkin.

Dialektlarni tasniflash va tahlil qilish:

- Har bir dialektning fonetik, leksik, sintaktik, semantik xususiyatlarini tahlil qilish. Masalan, bir dialektda “ko‘rmoq” so‘zi boshqasida “ko‘rish” deb aytiladi, bu dialektal farqlarni to‘g‘ri tasniflash kerak bo‘ladi.
- So‘zlar va iboralarning ma’nolarini ajratish, sinonimlar va antonimlarni aniqlash.

Tabiiy tilni qayta ishlash (NLP) va mashina o‘rganish (ML):

- Dialektal so‘zlarni avtomatik tanib olish, ularning semantik xususiyatlarini tahlil qilish va izohlarini aniqlash uchun AI modellarini yaratish. Bu modellar dialektal farqlarni o‘rganish va tasniflashda yordam beradi.
- O‘zbek tili uchun maxsus NLP modellarini yaratish, masalan, so‘zlarning morfologiyasini tahlil qilish, semantik va sintaktik xususiyatlarini aniqlash.

Lug‘at tizimini yaratish va visualizatsiya:

- Yaratilgan lug‘atni foydalanuvchi uchun qulay tarzda taqdim etish, interaktiv tizimlar (web ilovalar, mobil ilovalar) yordamida lug‘atni izlash va foydalanishni osonlashtirish.
- Foydalanuvchilar uchun izohli lug‘atning tarjima qilish yoki dialektal so‘zlarni boshqa dialektlarga o‘zgartirish imkoniyatini yaratish.

O‘zbek tili dialektlari va ularning izohli lug‘atini yaratish mavzusida xorijda ham, O‘zbekiston hududida ham bir qancha ilmiy tadqiqotlar olib borilgan. Ushbu tadqiqotlar asosan tilshunoslik, dialektologiya, leksikologiya, va kompyuter lingvistikasiga oid bo‘lib, har ikki mintaqada tilni chuqur o‘rganish va dialektal farqlarni aniqlashga e’tibor qaratilgan.

Ushbu yo‘nlaishda ilk tadqiqot namunalarini so‘z mulkkining sultoni Alisher Navoiyning asarlarida ham kuzatishimiz mumkin. **“Mahbub ul-qulub”** asarida

Navoiy o‘zining tilshunoslik va grammatikaga doir qarashlarini bayon etadi. “Mahbub ul-qulub” asarida Navoiy o‘zbek tilining o‘ziga xos xususiyatlarini, uning go‘zalligini va boyligini ta'kidlaydi. Bu asar tilning strukturasini tahlil qilishda muhim ahamiyatga ega. 20-asrning o‘rtalaridan boshlab O‘zbekistonning turli hududlarida (Farg‘ona vodiysi, Qashqadaryo, Surxondaryo, Buxoro, Samarqand va boshqalar) tilshunoslar o‘zbek tilining turli dialektlarini tahlil qilib, ularning fonetik, grammatik va leksik xususiyatlarini o‘rgandilar.

- **Qori Niyozov** va **Abdullaev Shukrullo** kabi tilshunoslar o‘zbek tilining turli dialektlarini tahlil qilgan va ushbu dialektlar orasidagi farqlarni o‘rgangan.
- **Erkinov Mirzaehsan** (O‘zbekistonda dialektologiya sohasida yetakchi tadqiqotchilardan biri) o‘z tadqiqotlarida o‘zbek tilining har bir dialekti uchun alohida lug‘atlar yaratish, so‘zlarning fonetik va morfologik o‘zgarishlarini aniqlashni maqsad qilgan. U dialektal so‘zlarning tarixi, semantik xususiyatlari, va ishlatilish kontekstlarini o‘rgangan.
- **G‘afur G‘ulom va Erkin Vohidov** kabi o‘zbek shoir va yozuvchilari dialektal leksikani o‘z asarlarida aks ettirib, tilning turli shakllarini ommaga tanishtirganlar. Shuningdek, o‘zbek tilining turli dialektlaridagi leksik o‘zgarishlarni o‘rganish bo‘yicha ilmiy izlanishlar mavjud.

O‘zbek tilining dialektlarini o‘rganishda tashqi omillarning roli ham muhim. Masalan, **Toshkent va Samarqand** kabi shaharlar o‘zgarishlarga moyil bo‘lgan, chunki bu hududlarda turli etnik guruhlar va ko‘chmanchilar birgalikda yashagan. Bu hududlarda o‘zbek tilining turli shakllari, jumladan, so‘zlarning fonetik xususiyatlari o‘zgarib ketgan. Bu mavzu ham ko‘plab ilmiy tadqiqotlarda yoritilgan.

- Sovet hukumati davrida ham, muayyan tadqiqotlar olib borilgan bo‘lib ulardan, **Yekaterina G. Prokofyeva** (SSSR davrida) o‘zbek tilining dialektlarini va turk til oilasiga mansub boshqa tillar o‘rtasidagi aloqalarni o‘rganib chiqqan. U dialektal so‘zlarni tahlil qilish, ularning tarixi va semantik xususiyatlarini izohlashda zamonaviy metodlarni qo‘llagan.
- **Maksim Il’ich Kormanov** va **Anatoliy S. Pocheptsov** kabi tadqiqotchilar kompyuter lingvistikasiga asoslangan dialektal izlanishlarni olib borgan. Ular o‘zbek tilining turli dialektlarining morfologiyasini va sintaksisini tahlil qilish uchun matematik modellar va algoritmlardan foydalanganlar.
- **Sun’iy intellekt va tabiiy tilni qayta ishlash (NLP)** sohasida yirik chet el tadqiqotchilari, masalan, **Berkeley Universiteti** va **Max Planck Instituti** lingvistik tadqiqotlar olib boradi. Bu tadqiqotlar doirasida o‘zbek tilining dialektal xususiyatlarini tahlil qilish va so‘zlarni avtomatik ravishda tanib olish uchun mashina o‘rganish va algoritmlardan foydalanish ko‘zda tutilgan.
- **Oxford Universiteti** va **Cambridge Universiteti** kabi nufuzli ilmiy muassasalarda, o‘zbek tili va uning dialektlari haqida xalqaro tadqiqotlar olib

borilgan. Bular asosan o‘zbek tilining boshqa turkiy tillar bilan aloqalarini va dialektlararo o‘zgarishlarni o‘rganishga qaratilgan.

- **Amerika va Yevropa ilmiy markazlarida** o‘zbek tilining turli dialektlari o‘rtasidagi farqlarni o‘rganish, xalqaro kommunikatsiyada dialektal o‘zgarishlarni hisobga olish uchun metodlar ishlab chiqish ishlari olib borilmoqda. Bu dialektlararo tafovutlarni aniqlash va tarqatishda madaniy almashinuvlar, masalan, **tillararo tarjima tizimlari** va **interkultural muloqot** uchun ahamiyatli bo‘lishi mumkin.

O‘zbekiston hududida va chet elda turli dialektologik tadqiqotlar olib borilgan va hozirda ham bu soha rivojlanmoqda. Chet elda kompyuter lingvistika va sun’iy intellekt yordamida dialektal farqlarni tahlil qilishda ilg‘or texnologiyalar qo‘llanilsa, O‘zbekistonda esa hozirgi kunda asosan an’anaviy tilshunoslik metodlari va zamonaviy tadqiqotlar birgalikda olib borilmoqda. Bu tadqiqotlar tilning boyligini, uning dialektal farqlarini o‘rganishda katta ahamiyatga kasb etadi.

O‘zbek tili dialektlari izohli lug‘atini yaratish va dialektologiya sohasida olib borilgan tadqiqotlar yutuqlari va kamchiliklari mavjud. Quyida ular tahlil qilinadi:

Yutuqlar:

O‘zbekistonda amalga oshirilgan dialektologik tadqiqotlar, o‘zbek tilining turli hududlaridagi dialektal farqlarni aniqlash va tasniflashda katta yutuqlarga erishdi. Masalan, **Alisher Navoiyning asarlari** orqali tilning tarixiy rivojlanishini o‘rganish, **Erkinov Mirzaehsan** kabi mutaxassislar tomonidan dialektlarning fonetik va grammatik xususiyatlari tahlil qilindi. **Sovet davrida** va undan keyin olib borilgan ishlarda, o‘zbek tilining turli dialektlari, ularning fonetik, leksik va morfologik xususiyatlari to‘liq o‘rganildi. O‘zbekistonlik tadqiqotchilar tomonidan **dialektal lug‘atlar** yaratish va **so‘zlarning semantik tahlili** amalga oshirilgan. Bu ilmiy ishlar tilni yaxshiroq tushunishga va o‘rganishga yordam beradi, ayniqsa, tilni o‘rganayotganlar yoki tarjimonlar uchun foydali. **Kompyuterlingvistika** va **tabiiy tilni qayta ishlash (NLP)** texnologiyalaridan foydalanish, masalan, dialektlarni avtomatik tanib olish va tahlil qilish imkoniyatini yaratgan. Bu, lug‘at va dialektal tahlillarni avtomatlashtirishda katta yutuqdir.

Chet eldagi ilmiy markazlar, masalan, **Oxford Universiteti** va **Max Planck Institut**ida olib borilgan tadqiqotlar dialektlararo o‘zgarishlarni va turk tillari o‘rtasidagi aloqalarni o‘rganish orqali O‘zbekistondagi dialektologik tadqiqotlar bilan taqqoslash imkonini yaratdi. Bu tadqiqotlar xalqaro ilmiy hamkorlikni va o‘zbek tilini global miqyosda o‘rganishni rivojlantirdi.

Mashina o‘rganish (machine learning) va **sun’iy intellekt** texnologiyalaridan foydalanish dialektal so‘zlarni avtomatik tanib olish va ularni tahlil qilishda katta yutuq keltirdi. Bu sohada chet elda amalga oshirilgan tadqiqotlar so‘zlarning morfologiyasini va semantik aloqalarini o‘rganish uchun yangi imkoniyatlarni yaratdi.

Kamchiliklar:

- Hozirgi tadqiqotlar, ayniqsa, O‘zbekistondagi hududiy dialektlar to‘liq o‘rganilmagan. O‘zbekistonda ko‘plab kichik qishloq va viloyatlarda yashovchi aholi orasidagi dialektal farqlar hali to‘liq aniqlanmagan.
- **Ba’zi dialektal so‘zlar** yoki **yangi so‘zlar** izohli lug‘atlarga kiritilmagan, chunki bu dialektal shakllar hali ilmiy jurnallarda yoki tadqiqotlarda yetarlicha yoritilmagan.
- **Mashina o‘rganish** va **tabiiy tilni qayta ishlash (NLP)** texnologiyalarining tillar o‘rtasidagi dialektal farqlarni to‘liq aniqlashdagi cheklovlari mavjud. O‘zbek tilidagi dialektal farqlar murakkab va kontekstual bo‘lib, avtomatik tizimlar hali ham ularni to‘liq tushunish va izohlashda qiyinchiliklarga duch keladi.
- **Mashina o‘rganish modellari** dialektal farqlarni aniqlashda yuqori aniqlikni ta'minlashda qiyinchiliklarga duch kelishi mumkin, chunki dialektlarning o‘ziga xos grammatik va fonetik xususiyatlari mavjud.
- **O‘zbek tilining dialektlari** haqidagi mavjud tadqiqotlar ko‘pincha an'anaviy metodlarga asoslangan. Ba'zi holatlarda ilmiy ishlar va lug‘atlar ko‘proq klassik lingvistik metodlarga asoslangan, ammo zamonaviy texnologiyalar, masalan, **yuzaga keladigan yangi dialektal so‘zlar** va **onlayn lug‘atlarni** yaratish jarayonida hali ham yetarlicha foydalanilmayapti.
- **Dialektal til materiallarini raqamlashtirish** masalasi O‘zbekistonda to‘liq amalga oshirilmagan. Ko‘plab dialektal matnlar va yozuvlar hozirda raqamlashtirilmagan va tizimli tarzda arxivlanmagan.
- **Xalqaro hamkorlik**da, ayniqsa, kompyuter lingvistikasi va sun'iy intellekt sohasida O‘zbekiston va chet el ilmiy markazlari o‘rtasidagi o‘zaro aloqalar hali ham yetarli darajada rivojlanmagan. Bu esa, masalan, dialektal so‘zlarni aniqlash va tahlil qilishda yangi yondashuvlar va metodlarni ishlab chiqish imkoniyatlarini cheklaydi.
- O‘zbek tili dialektlari vaqt o‘tishi bilan o‘zgarib bormoqda, shuning uchun eski dialektlarni o‘rganish va hujjatlashtirish qiyinlashmoqda. Ba'zi dialektlar yo‘qolib bormoqda, bu esa ilmiy tadqiqotlar va lug‘atlar yaratishda qiyinchiliklarga olib keladi.

O‘zbek tili dialektlari izohli lug‘atini yaratish va dialektologiya sohasida olib borilgan tadqiqotlar o‘zining muhim yutuqlariga ega, lekin ular hali ham to‘liq va mukammal emas. Kompyuterlingvistika, sun'iy intellekt va zamonaviy ilmiy metodlarni qo‘llash, dialektal farqlarni aniqlash va tahlil qilishda katta imkoniyatlar yaratgan. Biroq, dialektal farqlarni to‘liq qamrab olish, raqamlashtirish, va ilmiy hamkorlikni kuchaytirish bo‘yicha hali ko‘plab ishlar qilish kerak.

“O‘zbek tili dialektlari izohli lug‘atini yaratish modeli va algoritmi” mavzusining metodologiyasi ilmiy tadqiqotlar, texnologiyalar va tilshunoslik metodlariga asoslanadi. Bu metodologiya bir nechta bosqichlardan tashkil topgan bo‘lib, dialektal farqlarni tahlil qilish, lug‘at yaratish va ma’lumotlarni tizimli

ravishda taqdim etishni o‘z ichiga oladi. Quyida ushbu metodologiyaning asosiy bosqichlari keltirilgan:
Dialektlarni o‘rganish va tahlil qilish: Dialektlar o‘rtasidagi farqlarni aniqlash uchun tilshunoslik metodlari, masalan, fonetik, morfologik, syntaktik va semantik tahlillar ishlatiladi. Har bir dialektning o‘ziga xos xususiyatlarini aniqlashda quyidagi metodlar qo‘llaniladi:

- Fonetik tahlil: Dialektal farqlarni aniqlash uchun s**o‘**zlarning talaffuzi, intonatsiyasi va aksentlari **o‘**rganiladi. Bu dialektlarning fonetik xususiyatlarini tahlil qilishda yordam beradi.
- Dialektal s**o‘**zlarning shakllanishi va **o‘**zgarishini **o‘**rganish. Masalan, bir dialektda s**o‘**zlarning tugallanma shakllari boshqacha b**o‘**lishi mumkin.
- Dialektlar **o‘**rtasida sintaktik tuzilishdagi farqlarni aniqlash, masalan, jumlalarni qurish va ularda ishlatiladigan predikatlar va q**o‘**shimchalarning farqlanishi.
- Dialektal s**o‘**zlarning ma'nolarini **o‘**rganish. Ba’zi s**o‘**zlar bir dialektda boshqa ma'nolarga ega b**o‘**lishi mumkin. Bu semantik tahlilning muhim qismi b**o‘**ladi.

Dialektal leksikalarni yig‘ish: Dialektal lug‘atlarni yaratish uchun, har bir dialektga oid so‘zlar, iboralar, sinonimlar, antonimlar va so‘z birikmalarini yig‘ish zarur. Bu jarayon quyidagilarni o‘z ichiga oladi:

- Og‘zaki manbalarni yig‘ish: Mahalliy aholi, so‘zlashuvdagi dialektal farqlarni to‘plash. O‘zbek tilining turli hududlarida yashovchi odamlar bilan suhbatlar o‘tkazish va ulardan og‘zaki so‘zlarni yig‘ish.
- Yozma manbalarni yig‘ish: Mahalliy matnlar, adabiy asarlar, ilmiy ishlanmalar va boshqa yozma manbalar orqali dialektal leksikani to‘plash.
- Ma'lumotlarni tuzish va tizimlashtirish:Dialektal so‘zlarni tizimli ravishda tuzish uchun ma'lumotlar bazasi yaratiladi.

Bu bazaga har bir so‘zning:

- Fonetik, grammatik va semantik tavsifi;
- Izohlar va tarixiy konteksti kiritiladi;
- Sinonimlar, antonimlar va so‘z birikmalarining ro‘yxati kiradi.

Raqamli resurslar va raqamlashtirish: Dialektal so‘zlarni yig‘ish va tasniflashda raqamli texnologiyalarni qo‘llash zarur. Raqamlashtirish orqali dialektal so‘zlarni izlash va tahlil qilish imkoniyati oshadi. Ma'lumotlar bazasini yaratishda quyidagi metodlar qo‘llaniladi:

- XML, JSON va SQL kabi ma'lumotlar formatlari.
- Web ilovalar va mobil ilovalar orqali dialektal lug‘atga onlayn kirish imkoniyatini yaratish.

Mashina o‘rganish va tabiiy tilni qayta ishlash (NLP) texnologiyalaridan foydalanish

So‘zlarni tanib olish va tahlil qilish: Dialektal so‘zlarni avtomatik ravishda aniqlash uchun NLP algoritmlarini qo‘llash zarur. Bu metod quyidagilardan iborat bo‘lishi mumkin:

- So‘zlarni morfologik tahlil qilish: Mashina o‘rganish yordamida dialektal so‘zlarni morfologik tahlil qilish, so‘zning ildizini va qo‘shimchalarini aniqlash.
- So‘zlarning vektorial ifodasi (word embeddings): Dialektal so‘zlarning ma'nolarini va semantik aloqalarini aniqlash uchun sun'iy neyron tarmoqlari yoki word2vec kabi modellarni ishlatish.
- Semantik tahlil: So‘zlarning semantik ma'nolarini aniqlash va dialektal so‘zlarni izohli tarzda tasvirlash. Bu jarayon uchun, BERT yoki GPT kabi til modellaridan foydalanish mumkin.

Dialektlarni tasniflash va ajratish: Mashina o‘rganish modellaridan (masalan, k-means clustering, random forests, support vector machines) foydalangan holda, dialektlar o‘rtasidagi farqlarni aniqlash va avtomatik ravishda tasniflash mumkin. Bu dialektal farqlarni o‘rganish va ularni tizimli tarzda guruhlash imkonini beradi.

Leksik va sintaktik o‘zgarishlarni aniqlash: Dialektlarning leksik va sintaktik o‘zgarishlarini aniqlashda, NLP texnologiyalaridan foydalanish, jumladan, dependency parsing, named entity recognition (NER) kabi usullarni qo‘llash mumkin.

Izohli lug‘at yaratish: Dialektal so‘zlarning izohlarini yaratish uchun, har bir so‘zning ishlatilish kontekstini va tarixini taqdim etish kerak. Bu jarayonda quyidagilar amalga oshiriladi:

- Dialektal so‘zlar va iboralar uchun izohlar: Har bir so‘zning ma'nosi, ishlatilish konteksti va sinonimlari kiritiladi.
- Hikoyalar va namunalar: So‘zlar ishlatilgan matnlar yoki xalq og‘zaki ijodi namunalarini kiritish.

Interaktiv tizimlar va vizualizatsiya: Izohli lug‘atni foydalanuvchilar uchun qulay tarzda taqdim etish uchun interaktiv tizimlar yaratish. Bu quyidagilarni o‘z ichiga oladi:

- Web ilovalar: Foydalanuvchilarga dialektal so‘zlarni izlash va o‘rganish imkonini beradi.
- Mobil ilovalar: Dialektal so‘zlarni mobil qurilmada o‘rganish va tahlil qilish.
- Ma’lumotlar vizualizatsiyasi: So‘zlar va dialektlar o‘rtasidagi aloqalarni ko‘rsatish uchun grafikalar va diagrammalar yaratish.

Dialektlararo aloqalarni tahlil qilish: Dialektlar o‘rtasidagi aloqalarni aniqlash va tahlil qilishda, semantic networks (semantik tarmoqlar) va ontology-based approaches (ontologiyalar asosida yondashuvlar) qo‘llanilishi mumkin. Bu so‘zlar o‘rtasidagi semantik bog‘lanishlarni va dialektal farqlarni yanada chuqurroq tushunishga yordam beradi.

Ushbu mavzu doirasidagi metodologiya tilshunoslik, kompyuter lingvistika, mashina o‘rganish va tabiiy tilni qayta ishlash texnologiyalaridan foydalangan holda, dialektal so‘zlarni tizimli ravishda yig‘ish, tahlil qilish va izohli lug‘at yaratish jarayonlarini o‘z ichiga oladi. Yaratilgan lug‘at interaktiv tizimlar yordamida foydalanuvchilarga taqdim etiladi, shu bilan birga dialektal farqlarni o‘rganish va tahlil qilish uchun zamonaviy texnologiyalarni qo‘llash ilmiy tadqiqotlarda yangi imkoniyatlarni yaratadi.

Foydalanilgan adabiyotlar:

1. Iskandarov, N. (2001). O‘zbek tilining dialektologiyasi. – Tashkent: O‘zbekiston Milliy Ensiklopediyasi.
2. Raximov, B. (1997). O‘zbek tilining dialektik rivojlanishi va uning muammolari. – Tashkent: Fan.
3. Sultonov, X. (2006). O‘zbek tili va uning dialektik turlari. – Tashkent: O‘zbekistan davlat nashriyoti.
4. Usmonov, N. (2003). O‘zbek tili grammatikasi va dialektologiyasi. – Tashkent: Akademnashr.
5. Xolmatov, Sh. (2009). O‘zbek tilining dialektik va leksik xususiyatlari. – Tashkent: O‘qituvchi.
6. Abduqodirov, S. (2012). O‘zbek tili dialektlarining fonetik xususiyatlari. – Tashkent: Fan va texnologiya.
7. “O‘zbek tili va adabiyoti” jurnali
8. “Filologiya masalalari” jurnali
9. “O‘zbekiston tilshunosligi” jurnali

10. O‘zbek tilshunoslarining ilmiy jurnali - (https://www.tilshunoslik.uz)
11. S. Datta Investigating English-Language Dialect-Adjusted Models. A Thesis Submitted to the Faculty in partial fulfillment of the requirements for the degree of Bachelor of Arts in Computer Science by Samiha Datta Advised by Professor Soroush Vosough. Dartmouth College Dartmouth Digital Commons. June 2023. p95
12. A.Melikyan Technology for creating digital explanatory dictionaries. E3S Web of Conferences 210, 02003 (2020)Don State Technical University, 344003, Gagarin Sq., 1, Rostov-on-Don, Russia 2020. p11
13. www.language.uz
14. www.tilshunoslik.uz

VIII SHO‘BA. SEMANTIK TIZIMLAR VA LINGVISTIK ONTOLOGIYALAR

SENTIMENT TAHLILIDA SO‘Z TURKUMLARINING ROLI

Elov Botir Boltayevich,
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Xusainova Zilola Yuldashevna,
filologiya fanlari falsafa doktori (PhD)
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Annotatsiya. Sentiment tahlilda so‘z turkumlarining roli matnning emotsional ulushini aniqlashda hal qiluvchi omil hisoblanadi. O‘zbek tilida bu jarayon tilning agglyutinativligi, morfologik murakkabligi va kodli almashish (rus/ingliz so‘zlari) kabi xususiyatlar tufayli yanada murakkablashadi. Olib borilgan tadqiqotlar shuni ko‘rsatadiki, sifatlar (“ajoyib”, “yomon”) va ravishlar (“juda”, “ortiqcha”) to‘g‘ridan-to‘g‘ri sentimentni ifodalashda ustunlik qiladi, fe’llar (“yoqtirmoq”, “yomon ko‘rmoq”) esa harakatning emotsional yo‘nalishini belgilaydi. Otlar (“zavq”, “qayg‘u”) ko‘pincha abstrakt tushunchalar orqali bilvosita hissiyotni anglatadi. Inkor konstruksiyalari (“emas”, “yo‘q”) va bog‘lovchilar (“lekin”, “shuningdek”) kabi elementlar sentimentni teskariga o‘zgartirishi yoki nuanslarni kuchaytirishi mumkin. Ushbu maqolada sifat, fe’l, ravish va ot kabi asosiy so‘z turkumlarining matnning hissiy mazmunini aks ettirishdagi ahamiyati hamda ularning semantik xususiyatlari yoritildi. Tadqiqotda leksikon asosidagi metodlar va mashinali o‘qitish yondoshuvlari yordamida so‘z turkumlarining aniqligi va to‘liq aks ettirilishini ta’minlash zarurligini isbotlandi. Shuningdek, inkor va salbiy konstruksiyalar, bog‘lovchilar va kontekstual elementlar yordamida matnda ifodalangan hissiyotlarning murakkab o‘zaro bog‘liqligi tahlil qilindi. Ushbu yondashuv tilning murakkab semantik va morfologik xususiyatlarini hisobga olgan holda, sentiment ifodalarining nozik tafsilotlarini aks ettirish imkoniyatini beradi. Eksperimental natijalar, so‘z turkumlari asosida ajratilgan xususiyatlarning sentiment tahlilda yuqori samaradorlikka erishish uchun asosiy omil ekanligini tasdiqlaydi. Maqola sentiment tahlilida so‘z turkumlarining roli va ularning integratsiyalashuvi asosida innovatsion metodologik yechimlarni taklif etib, kelgusidagi ilmiy izlanishlar uchun mustahkam poydevor yaratadi.

Аннотация. В анализе настроений роль словосочетаний является решающим фактором в определении эмоционального содержания текста. В

узбекском языке этот процесс еще больше осложняется такими особенностями языка, как агглютинативность, морфологическая сложность и переключение кодов (русские/английские слова). Исследования показывают, что прилагательные («отлично», «плохо») и наречия («очень», «чрезмерно») доминируют в выражении прямого чувства, тогда как глаголы («нравится», «не нравится») определяют эмоциональную направленность действия. Существительные («удовольствие», «грусть») часто выражают эмоции косвенно, через абстрактные понятия. Такие элементы, как отрицательные конструкции («не», «не») и союзы («но», «также»), могут изменить смысл или усилить нюанс. В статье подчеркивается важность ключевых категорий слов, таких как прилагательные, глаголы, наречия и существительные, в отражении эмоционального содержания текста, а также их семантических свойств. Исследование продемонстрировало необходимость обеспечения точности и полноты классов слов с использованием методов, основанных на лексике, и подходов машинного обучения. Сложная взаимосвязь эмоций, выраженных в тексте, также анализировалась с использованием отрицательных и отрицательных конструкций, союзов и контекстуальных элементов. Такой подход позволяет отразить тонкие детали выражения эмоций, учитывая сложные семантические и морфологические особенности языка. Экспериментальные результаты подтверждают, что признаки, извлеченные на основе категорий слов, являются ключевым фактором в достижении высокой эффективности анализа настроений. В статье предлагаются инновационные методологические решения, основанные на роли групп слов и их интеграции в анализ настроений, что создает прочную основу для будущих научных исследований.

Kalit so‘zlar: *Sentiment tahlil, so‘z turkumlari, sifat (adjective), fe’l(verb), ravish (adverb), ot(noun), inkor konstruksiyalari, emotsional intensivlik, kontekstual elementlar, semantik tahlil, NLP.*

Kirish

Tabiiy tilni qayta ishlash sohasida sentiment tahlil algoritmlari kundalik hayot, biznes va ijtimoiy media analitikasi uchun dolzarb vazifa hisoblanadi. Ushbu jarayonda so‘z turkumlari – *otlar, fe’llar, sifatlar, ravishlar* va boshqa lingvistik kategoriyalar – matnning semantik va hissiy mazmunini aniqlashda muhim rol o‘ynaydi. Adabiyotlarda, ayniqsa, leksikon asosidagi sentiment tahlilda **sifatlar va ravishlarning** hissiy qiymatini aks ettirishdagi ahamiyati keng yoritilgan [1]. Shuningdek, **fe’llar va otlar** kabi asosiy so‘z turkumlari matnning asosiy ma’lumotini yetkazishda kalit so‘z turkum sifatida rol o‘ynab, ularning kontekstual aloqalari tahlilda sezilarli ahamiyatga ega ekanligi ta’kidlanadi. Olib borilgan adqiqotlarda so‘z turkumlarini ajratish va tahlil qilish orqali sentimentni aniqlashda yanada boy va to‘liq semantik ma’lumot olish mumkinligi ko‘rsatilgan. Ushbu

yondashuv orqali, matnning strukturaviy va leksik xususiyatlari inobatga olinib, hissiyotlarning aniqligi oshiriladi. Bunday yondashuv, ayniqsa, murakkab sintaktik va semantik tuzilishga ega matnlarda sentiment tahlilining samaradorligini sezilarli darajada oshiradi. Shuningdek, so‘z turkumlarining rolini tahlil qilish, sentimentni aniqlash algoritmlarining umumlashuv qobiliyatini mustahkamlash va aniqlikni oshirishga yordam beradi. Leksik, morfologik va semantik jihatdan so‘z turkumlarini ajratib olish metodlari, modelning trening jarayonida muhim xususiyatlar sifatida qo‘llaniladi. Ushbu metodologik yondashuv, sentiment tahlil algoritmlariga nafaqat to‘g‘ri ma’lumot uzatish, balki ma’lumotlar orasidagi noaniq aloqalarni ham aniqlash imkonini beradi. Shuningdek, so‘z turkumlarining statistik tahlili va semantik modellashtirish, modelning optimallashtirish bosqichida ham qo‘llaniladi.

Tadqiqotlar natijalari shuni ko‘rsatadiki, sifat va ravish kabi so‘z turkumlarining sentiment ifodasidagi roli, **matnning hissiy ahamiyatini aniqlashda yuqori darajada seziladi**. Bundan tashqari, so‘z turkumlarini ajratish orqali olinadigan xususiyatlar, mashinali o‘qitish algoritmlarida o‘zaro bog‘lanishlarni yanada puxta aks ettirishga yordam beradi. Ushbu ilmiy yondashuv, tilning murakkab semantik va kontekstual jihatlarini inobatga olib, sentiment tahlilning yuqori aniqlikdagi yechimlarini yaratishda muhim poydevor hisoblanadi. Shunday qilib, maqolamiz so‘z turkumlarining sentiment tahlilida o‘ynagan rolini, modelning umumiy samaradorligiga qo‘shgan hissasini va kelgusidagi tadqiqotlar uchun istiqbolli metodologik yechimlarni taqdim etishga qaratilgan. Natijada, ushbu tadqiqot nafaqat mavjud adabiyotlar asosida, balki amaliy tajriba va statistik tahlillar orqali ham so‘z turkumlarining sentiment tahlilidagi ahamiyatini ilmiy asosda isbotlaydi.

Tabiiy tilni qayta ishlash sohasida sentiment tahlilining samaradorligini oshirishda so‘z turkumlarining roli keng o‘rganilgan va adabiyotlarda e’tirof etilgan. B.Pang va L.Lee [1] tomonidan olib borilgan tadqiqotlarda sifatlar va ravishlarning matnning hissiy mazmunini aks ettirishdagi o‘ziga xos roli ta’kidlangan bo‘lsa, ularning to‘g‘ri identifikatsiyasi tahlil natijalarining aniqligini oshirishga xizmat qilmoqda. Shuningdek, Socher va boshqa tadqiqotchilar tadqiqotlari fe’l va otlar kabi asosiy so‘z turkumlarining kontekstual ma’lumotlarni yetkazishda muhim ahamiyatga ega ekanligini ko‘rsatdi [2]. Ushbu natijalar shuni anglatadiki, har bir so‘z turkumi o‘zining semantik xususiyatlariga ega bo‘lib, ular sentiment tahlilning nozikligini va to‘liq aks ettirilishini ta’minlashda muhim omil hisoblanadi. Morfologik tahlil va leksik resurslarga asoslangan metodlar yordamida so‘z turkumlarini aniqlash, sentiment tahlilda so‘zlarning hissiy qiymatini aniq belgilash imkonini beradi[3]. Bunga qo‘shimcha ravishda, chuqur o‘rganish metodlari, masalan, Kim (2014) tomonidan ishlab chiqilgan konvolyutsion neyron tarmoqlari, so‘z turkumlarini inobatga olgan holda matnning semantik va kontekstual

xususiyatlarini samarali o‘rganishga yordam beradi[4]. Tadqiqotchilar, leksikon asosidagi yondoshuvlar va mashinali o‘qitish algoritmlarini integratsiyalash orqali, so‘z turkumlarining hissiy ifodalarini yanada mukammal tahlil qilish imkoniyatini oshirishni maqsad qilgan. Ushbu yondoshuvlar yordamida, sifatlar va ravishlar kabi so‘z turkumlari orqali sentiment tahlilning aniqligi sezilarli darajada oshirilishi, otlar va fe’llar esa matnning kontekstual tuzilishini aks ettirishda qo‘llanilishi ta’kidlangan. Adabiyotlarda keltirilgan natijalar, so‘z turkumlariga asoslangan xususiyat ajratish metodlarining sentiment tahlilning aniqligini oshirishda samarali ekanligini dalillar bilan isbotlagan. Shu bilan birga, so‘z turkumlarining statistik va semantik tahlili, modelning umumlashuv qobiliyatini mustahkamlashda muhim rol o‘ynashi ham aniqlangan. Natijada, sentiment tahlilda so‘z turkumlarining roli chuqur o‘rganilib, bu yondoshuv kelgusidagi ilmiy izlanishlar va amaliy ilovalar uchun istiqbolli metodologik poydevor sifatida xizmat qilishi isbotlandi. Ushbu tadqiqot natijalari, so‘z turkumlarining sentiment ifodasidagi ahamiyatini ilmiy asosda yoritib, tilning murakkab semantik va kontekstual xususiyatlarini aks ettirishda innovatsion yechimlarni ilgari surishga xizmat qiladi.

O‘zbek tilida so‘z turkumlarining sentiment tahlilga ta’sirini o‘rganish bo‘yicha ilmiy ishlar nisbatan kam bo‘lsa-da, so‘nggi yillarda bir qator muhim tadqiqotlar amalga oshirildi. Elov va boshqa tadqiqotchilar (2024) negatsiyani ifodalovchi so‘zlar (“hech”, “yo‘q”) va ularning turkumlarini (ot, ravish) tahlil qilib, salbiy kontekstlarni aniqlash algoritmini taklif qildi [5]. Z.Xusainova (2023) O‘zbek tilining agglyutinativligi tufayli so‘z turkumlarining morfologik o‘zgarishi sentimentni aniqlashdagi murakkabliklarni tahlil qildi [6]. B. Elov, A.Abdullayevlar (2024) ravishlar (“juda”, “g‘oyat”) va modal so‘zlar (“ehtimol”, “balki”) hissiy intensivlikni modellashtirishda 20% aniqlik oshishiga sabab bo‘lishini isbotladi [7]. Shuningdek ular LSTM tarmoqlari yordamida so‘z turkumlarining uzoq masofali bog‘lanishlarini (masalan, sifat + ot) o‘rganib, modelning F1-skorini 18% oshirdi. SVM yordamida otlar (“zavq”, “qayg‘u”) va ularning semantik maydonlarini klassifikatsiya qilish samaradorligini solishtirdi. Xalqaro tadqiqotlar shuni ko‘rsatadiki, ingliz tilida sifatlar sentimentni aniqlashda 70% ishonchlilikga ega[8], o‘zbek tilida esa bu ko‘rsatkich morfologik murakkablik tufayli pastroq (55-60%). Cambria (2024) semantik-psixolingvistik yondashuvlar orqali so‘z turkumlarining pragmatik ma’nolarini hisobga olishni tavsiya qildi [9]. Abdurakhmonova va boshq. (2024) “emotsional lug‘at” yaratishda so‘z turkumlarini alohida kategoriyalashtirish zarurligini ta’kidladi [10]. Tadqiqotlar shuni ko‘rsatadiki, O‘zbek tilida so‘z turkumlarini hisobga olgan holda ishlangan modellar aniqligini oshirish uchun morfologik va kontekstual tahlillarni integratsiyalash talab etiladi.

Yuqorida olib borilgan tadqiqotlarni umumlashtirgan holda, o‘zbek tilida so‘z turkumlarini sentiment tahlil jarayonidagi ahamiyatini keltiramiz. Sentiment tahlilida matn mazmunidagi hissiyotlarni aniqlash uchun turli so‘z turkumlarining

rolini hisobga olish muhim. O‘zbek tilidagi matnlarni sentiment tahlil qilishda quyidagi so‘z turkumlariga e’tibor qaratish kerak:

1. **Sifat (adjective):** Sifatlar matnning hissiy mazmunini aks ettirishda eng muhim rol o‘ynaydi, chunki ular obyektning yoki holatning sifatini aniq belgilaydi va ularning ijobiy yoki salbiy xarakterini ifodalaydi. Sifatlar ko‘pincha biror narsa yoki holatning sifatini, ya’ni uning qanday ekanligini ifoda etadi. **Sifatlar** matndagi asosiy hissiyotlarni bevosita ifoda etadi. Sifatlar orqali foydalanuvchilarning fikrlarida aks etgan hissiyot darajasi va intensivligi aniqlanadi, bu esa sentiment tahlil algoritmlarining semantik vaznini shakllantirishda muhim omil hisoblanadi. Masalan, “*yaxshi”*, *“ajoyib”*, “*yomon”*, *“dahshatli”*, *“qo‘rqinchli”* kabi sifatlar yordamida matnda qanday hissiyot mavjudligi aniqlanadi.

2. **Fe’l (verb):** Fe’llar, matndagi harakatlar va jarayonlarni ifodalash orqali hissiyotning dinamik jihatlarini yoritadi. Masalan, *“yoqtirdi”, “qabul qilmaydi”* kabi fe’llar gapning asosiy ma’nosini va uning hissiy yo‘nalishini aniqlashda qo‘llaniladi. Fe’l tarkibidagi semantik ma’lumotlar, modelga sentimentning qaysi yo‘nalishda ifodalanganligini aniqlash imkonini beradi.

3. **Ravish (adverb):** Ravishlar sifatlar yoki fe’llarga nisbatan qo‘shimcha ma’lumot beradi va hissiyotlar darajasini kuchaytirishi yoki yumshatishi mumkin. **Ravishlar** sifatlar va fe’llarning kuchini, intensivligini oshirish yoki yumshatishda yordam beradi. Masalan, “*juda”, “ancha”, “tezda”* kabi ravishlar gapdagi hissiyotlarning nozikligini oshiradi, bu esa tahlil modelining aniq va puxta sentiment ifodasini shakllantirishiga yordam beradi.

4. **Ot (noun):** Otlar matnning asosiy predmetlarini va kontekstini belgilovchi elementlardir. Ular orqali, gapning mavzusi va unda aks etgan sentiment obyektlari aniqlanadi (masalan, *“film”, “xizmat”, “kitob”*). Otlar tahlilda kontekstual munosabatlarni belgilashda, sentimentning qaysi obyektga nisbatan ifodalanganligini tushunishda muhim rol o‘ynaydi.

5. **Inkor va salbiy konstruksiyalar:** Inkor va salbiy konstruksiyalar matnda ifodalangan salbiy hissiyotlarni aniqlash va ijobiy fikrlarni inkor etishda hal qiluvchi omil hisoblanadi. “*Emas”, “yo‘q”, “inkor qilinadi”* kabi so‘zlar gap tarkibidagi salbiy yoki neytral sentimentni aniq belgilaydi va modelga noaniq ma’lumotlarni kamaytirishda yordam beradi.

6. **Bog‘lovchilar va kontekstual elementlar:** Bog‘lovchilar (masalan, *“ammo”, “lekin”, “va”, “shuningdek”*) va kontekstual elementlar gaplar o‘rtasidagi semantik aloqalarni shakllantirib, matnning umumiy kontekstini belgilaydi.

Ushbu elementlar yordamida, model gapning strukturasini yaxshiroq tahlil qilib, ijobiy, salbiy yoki neytral hissiyotlarning o‘zaro bog‘liqligini aniqlash

imkoniyatiga ega bo‘ladi. Ular gapning semantik tuzilishini aniqlashda yordam beradi.

Shunday qilib, sifat, fe’l, ravish, ot, inkor va salbiy konstruksiyalar, shuningdek, bog‘lovchilar va kontekstual elementlar sentiment tahlilida matnning semantik mazmunini to‘liq aks ettirishda, nozik hissiyotlarni aniqlashda va aniq tasniflashda o‘zaro hamkorlikda ishlaydi. Ushbu so‘z turkumlarining to‘g‘ri ajratilishi va tahlili, sentiment modelining aniqligi hamda ishonchliligini sezilarli darajada oshirishga yordam beradi.

№	So‘z tukumi	Izoh
1.	**Sifat**	his-tuyg‘uni bevosita bildiradi
2.	**Fe’l**	harakat yoki holat orqali his-tuyg‘uni ifodalaydi
3.	**Ot**	hissiyot va tushunchalar
4.	**Ravish**	kuchaytirish, zaiflashtirish yoki aniqlik berish

Shuningdek, kontekst va morfologik tuzilmani hisobga olgan holda, maxsus sentiment lug‘atlaridan yoki mashinali o‘qitish metodlaridan foydalanish, so‘zlar orasidagi munosabatlarni aniqlashda qo‘shimcha ma`lumotlarni beradi. Shu bilan birga, kontekstual semantik aloqalarni ham hisobga olish zarur, chunki bitta so‘z turli kontekstlarda turlicha hissiyotlarni bildirishi mumkin. Yuqoridagi so‘z turkumlari yordamida matnni bo‘laklarga ajratish, so‘zlarni tegishli sentiment kategoriyalariga joylashtirish va natijada matnning umumiy hissiy xususiyatini aniqlash mumkin.

1-jadval. Sifatlar va ularning sentiment bahosi

№	Sifat	Sentiment bahosi	Namuna
1.	ajoyib	5	Bu film ajoyib edi.
2.	yaxshi	3	Xizmat yaxshi bo‘ldi.
3.	yomon	-3	Bu joy yomon ekan.
4.	dahshatli	-5	Ko‘rgan manzara dahshatli edi.
5.	mukammal	5	Uning chiqishi mukammal edi.
6.	zo‘r	4	Konsert zo‘r bo‘ldi.
7.	zerikarli	-3	Uchrashuv zerikarli edi.

2-jadval. Fe’llar va ularning sentiment bahosi

№	Fe’l	Sentiment bahosi	Namuna
1.	yoqtirmoq	+4	Men bu musiqani **yoqtiraman**.
2.	yomon ko‘rmoq	-4	U bunday odatni **yomon ko‘radi**.
3.	xursand bo‘lmoq	+5	Men bu xabarni eshitib **xursand bo‘ldim**.
4.	xafa bo‘lmoq	-3	U natijadan **xafa bo‘ldi**.
5.	sevinmoq	+5	U bu sovg‘ani olib **sevinib** ketdi.

6.	qo‘rqmoq	-4	Tunda yolg‘iz qolishdan **qo‘rqaman**.
7.	hayratlanmoq	+3	Yangi kashfiyot uni **hayratlantirdi**.
8.	hafsalasi pir bo‘lmoq	-5	Yomon natija uning **hafsalasini pir qildi**.

3-jadval: Otlar va ularning sentiment bahosi

№	Ot	Sentiment bahosi	Namuna
1.	quvonch	+5	Yangi g‘alaba **quvonch** baxsh etdi.
2.	g‘am	-3	Yolg‘izlik uni **g‘amga botirdi**.
3.	baxt	+4	Haqiqiy **baxt** oiladan boshlanadi.
4.	qayg‘u	-5	Yo‘qotish uning hayotida katta **qayg‘u** keltirdi.
5.	omad	+4	Unga **omad** kulib boqdi.
6.	tashvish	-3	U ertangi imtihondan **tashvishda** edi.
7.	g‘azab	-4	Uning yuzida **g‘azab** ko‘rinib turardi.
8.	ilhom	+5	Yangi she’r yozishga **ilhomi oldi**.

4-jadval: Ravishlar va ularning sentiment bahosi

№	Ravish	Sentiment bahosi	Namuna
1.	juda	+3	Men bu joyni **juda** yaxshi ko‘raman.
2.	mutlaqo	+5	Bu kitob **mutlaqo** ajoyib!
3.	hech	-4	Bu ish **hech** yoqmadi.
4.	biroz	+1	U kecha **biroz** xursand edi.
5.	ancha	+3	Uning kayfiyati **ancha** ko‘tarildi.
6.	nihoyatda	+5	U **nihoyatda** baxtiyor edi.
7.	bejirim	+4	Ushbu bezaklar **bejirim** ishlangan.
8.	juda ham	+3	Bu restoran **juda ham** chiroyli ekan.

Xulosa

Sentiment tahlilning asosiy vazifasi matndagi hissiy tonni (ijobiy, salbiy, neytral) aniqlashdan iborat bo‘lib, bunda so‘zlar semantik va pragmatik jihatdan asosiy indikator hisoblanadi. So‘z turkumlari (ot, sifat, fe’l, ravish va boshqalar) sentimentni ifodalashda turlicha rol o‘ynashi lingvistik tadqiqotlarda ko‘p marta ta’kidlangan. O‘zbek tilida bu masala yanada murakkabdir, chunki tilning agglyutinativ tuzilishi so‘z shakllarining ko‘pligiga, morfologik bog‘liqliklarga va kontekstual nuanslarga olib keladi. Masalan, sifatlar (“ajoyib”, “yomon”) to‘g‘ridan-to‘g‘ri emotsional ulushlangan bo‘lsa, otlar (“zavq”, “qayg‘u”) ko‘pincha mavhum tushunchalarni ifodalaydi. Fe’llar esa (“yoqtirmoq”, “yomon ko‘rmoq”) nutqdagi harakatning hissiy yo‘nalishini ko‘rsatadi. O‘zbek tilida ravishlar (“juda”, “ortiqcha”) emotsional intensivlikni kuchaytirishda muhim rol o‘ynaydi. Biroq, so‘z

turkumlarining sentimentga ta’sirini tahlil qilishda ikkita asosiy muammo mavjud: birinchidan, bir xil so‘z turkumining turli kontekstlarda qarama-qarshi ma’nolar kasb etishi (masalan, “do‘st” – ijobiy, lekin “soxta do‘st” – salbiy); ikkinchidan, kodli almashish (rus/ingliz tillaridan so‘zlar) tufayli semantik noaniqlik. O‘zbek tilidagi so‘z turkumlarini morfologik jihatdan normalizatsiya qilish (lemmatizatsiya) orqali ularni sentiment lug‘atlariga integratsiyalash imkoniyatlarini ko‘rsatdi. Ushbu tadqiqotda o‘zbek tilida so‘z turkumlarining sentiment tahlilga ta’sirini tizimli o‘rganilib, ularni modellashtirishda samarali yondashuvlar taklif qilindi. Natijalar nafaqat lug‘atli tizimlarni boyitish, balki neyron tarmoqlarning kontekstni tushunish qobiliyatini oshirishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Pang B. and Lee L. (2008). Opinion mining and sentiment analysis. Foundations and Trends® in information retrieval, 2(1-2), 1-135.
2. Socher R., Perelygin A., Wu J., Chuang J., Manning C. D., A. Y. and Potts C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).
3. Esuli A. and Sebastiani F. (2006, May). Sentiwordnet: A publicly available lexical resource for opinion mining. In LREC (Vol. 6, pp. 417-422).
4. Kim S. H. and Kim D. (2014). Investor sentiment from internet message postings and the predictability of stock returns. Journal of Economic Behavior & Organization, 107, 708-729.
5. Elov B. and Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).
6. Xusainova Z. (2023). O‘zbek tili milliy korpusi qidiruv tizimini optimallashtirishda lemmatizatsiyadan foydalanish. Uzbekistan: Language and Culture, 2(2).
7. Elov B., Alayev R. and Abdullayev A. (2024). Tabiiy tilning statistik modellari. digital transformation and artificial intelligence, 2(6), 178-189.
8. Tang D., Wei F., Qin B., Yang N., Liu T. and Zhou M. (2015). Sentiment embeddings with applications to sentiment analysis. IEEE transactions on knowledge and data Engineering, 28(2), 496-509.
9. Cambria E. (2024). Semantics Processing. In Understanding Natural Language Understanding (pp. 113-228). Cham: Springer Nature Switzerland.
10. Mengliev D., Abdurakhmonova N., Barakhnin, V. Vasliddinova, K. Rahimov H. and Djalolova, K. (2024, June). Enhancing sentiment analysis in Uzbek language texts through weighted lexical features. In 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM) (pp. 2450-2453). IEEE.

O‘ZBEK TILIDAGI MATNLARNI SO‘Z TURKUMLARI ORQALI SENTIMENT TAHLIL QILISHDA GAP MODELLARINI ISHLAB CHIQISH

Elov Botir Boltayevich,
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Xusainova Zilola Yuldashevna,
filologiya fanlari falsafa doktori (PhD)
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Abdullayev Abdulla Quranbayevich,
NTM ta’limni kredit tizimini boshqarish bo‘lim boshlig‘i
abdulla_abdullayev9270@mail.ru
Urganch innovatsion university

Annotatsiya. Zamonaviy axborot texnologiyalarining jadal rivojlanishi natijasida tabiiy tilni qayta ishlash (NLP) sohasiga bo‘lgan talab ortib bormoqda. Ayniqsa, sentiment tahlil texnologiyalari foydalanuvchilar fikrlarini avtomatik ravishda tahlil qilish, ijobiy, salbiy yoki neytral munosabatlarni aniqlash imkonini beradi. Bugungi kunda sentiment tahlil texnologiyalari ijtimoiy tarmoqlar, yangiliklar, sharhlar va foydalanuvchi fikrlariga asoslangan ko‘plab sohalarda keng qo‘llanilmoqda. Ingliz, xitoy, ispan va boshqa yirik tillar uchun bunday tizimlar yaxshi rivojlangan bo‘lsa-da, o‘zbek tilida sentiment tahlilga yo‘naltirilgan tadqiqotlar hali boshlang‘ich bosqichda qolmoqda. Shu bois, o‘zbek tili uchun xos leksik va grammatik xususiyatlarni hisobga olgan holda samarali sentiment tahlil modellarini ishlab chiqish dolzarb ilmiy va amaliy muammodir. O‘zbek tilida so‘z turkumlari, ya’ni sifat, fe’l, ravish, ot va boshqa grammatik birliklar, gapdagi semantik ulushni belgilovchi asosiy elementlar hisoblanadi. Ushbu turkumlar matndagi hissiy ifodaning aniqligini ta’minlashda muhim o‘rin egallaydi. So‘z turkumlariga asoslangan sentiment tahlil modellarini yaratish esa tilning ichki tuzilmasini chuqur tahlil qilishni talab etadi. Ayniqsa, gap darajasida so‘z turkumlari bilan boyitilgan modellarni ishlab chiqish orqali matn mazmuni aniqroq talqin qilinadi. Shu munosabat bilan, ushbu maqolada o‘zbek tilidagi matnlarni so‘z turkumlari asosida sentiment jihatdan tahlil qilishga yo‘naltirilgan gap modellarini ishlab chiqish bo‘yicha metodologik yondashuvlar bayon etiladi. Maqolada, avvalo, mavjud sentiment tahlil yondashuvlari tahlil qilinadi, so‘ng o‘zbek tili morfologiyasiga moslashtirilgan gap modeli ishlab chiqiladi. Tadqiqotda sifat, fe’l

va ravishlarning hissiy ohangni belgilashdagi roli, shuningdek, ot va kontekstual birliklarning qo‘llanishi chuqur o‘rganiladi. Bundan tashqari, inkor konstruksiyalari va bog‘lovchi vositalarning sentimentni o‘zgartirishdagi funksional ta’siri ham tahlil qilinadi. Ushbu yondashuvlar asosida yaratilgan gap modellari sentiment tahlilning aniqligini oshirishda xizmat qilishi ko‘zda tutiladi. Natijada, maqola o‘zbek tilida sentiment tahlilning ilmiy asoslarini kengaytirish hamda real amaliyotda qo‘llashga mos modellarni ishlab chiqishga hissa qo‘shadi.

Kalit so‘zlar: *Sentiment tahlil, so‘z turkumlari, sifat (adjective), fe’l(verb), ravish (adverb), ot(noun), inkor konstruksiyalari, emotsional intensivlik, kontekstual elementlar, semantik tahlil, NLP.*

Kirish

Tabiiy tilni qayta ishlash sohasida sentiment tahlil bugungi kunda ijtimoiy tarmoqlar, sharhlar va boshqa axborot manbalaridan olinayotgan matnlarning hissiy mazmunini aniqlashda dolzarb vazifa sifatida namoyon bo‘lmoqda. O‘zbek tili kontekstida tilning o‘ziga xos leksik va morfologik xususiyatlari sababli, sentiment tahlilning aniqligi va ishonchliligini ta’minlash uchun maxsus yondashuvlar talab etiladi. Sentiment tahlilning dolzarbligi zamonaviy NLP (Tabiiy Tilni Qayta Ishlash) sohasida mijozlar sharhidan tortib ijtimoiy tahlilgacha turli ilovalarda o‘z aksini topmoqda. Ushbu maqolada, o‘zbek tilidagi matnlarni so‘z turkumlari orqali tahlil qilishda gap modellarini ishlab chiqishning nazariy va amaliy jihatlari o‘rganiladi. So‘z turkumlarining (sifat, fe’l, ravish, ot) sentimentga ta’sirini tizimli o‘rganish ushbu muammolarni bartaraf etishda muhim rol o‘ynaydi.

Ushbu maqolada so‘z turkumlarining semantik va sintaktik xususiyatlarini inobatga olgan holda, gap modellarining samaradorligini oshirishga qaratilgan. Avvalo, matnning hissiy mazmunini aks ettirishda sifat, fe’l, ravish va ot kabi asosiy so‘z turkumlarining roli aniqlik bilan o‘rganilib, ularning gap tarkibidagi semantik og‘irligi tahlil qilindi. Ushbu so‘z turkumlari gap modellarining tuzilishini belgilab, sentiment tahlilning nozik tafsilotlarini ochib berishda hal qiluvchi omil sifatida e’tirof etiladi.

Tadqiqotchilar tomonidan ilgari olib borilgan ishlar, masalan, Socher va boshq. (2013) tomonidan taklif etilgan chuqur o‘rganish yondashuvlari, gap modellarining murakkab semantik va kontekstual aloqalarni aniqlashdagi samaradorligini isbotlagan. Maqolada, strukturlanmagan **ma’lumotlarni to‘plash, tozalash** va **xususiyat ajratish** bosqichlari orqali so‘z turkumlariga asoslangan xususiyatlar yig‘ilib, gap modellarining asosiy komponentlari shakllantiriladi. Modellarni o‘rgatish bosqichida mashinali o‘qitish algoritmlari yordamida gap modellarining parametrlari optimallashtirilib, ularning sentiment tahlilidagi aniqligi kross-validatsiya metodlari orqali sinovdan o‘tkaziladi. Natijalar gap modellarining

so‘z turkumlari asosida sentiment tahlilda yuqori aniqlik va ishonchlilikka erishishda muhim rol o‘ynashini ko‘rsatadi.

Shuningdek, o‘zbek tilining leksik va morfologik xususiyatlarini chuqurroq o‘rganish gap modellarini yanada mukammallashtirishga imkon beradi. Ushbu yondashuv tilning murakkab semantik va kontekstual jihatlarini inobatga olgan holda, sentiment tahlil algoritmlarining samaradorligini oshirishda innovatsion metodologik yechimlarni taqdim etadi.

Tadqiqotlar shuni ko‘rsatadiki, an’anaviy usullar (masalan, Bag-of-Words) o‘zbek tilidagi so‘z shakllarining ko‘pligi va kontekstual bog‘liqliklarini hisobga olmaydi. Shu sababli, gap darajasidagi modellarni ishlab chiqish zaruriyati paydo bo‘ladi, chunki gap tarkibidagi so‘z turkumlari o‘zaro munosabati emotsional ma’noni aniqroq aks ettiradi. Masalan, “Bu film ajoyib emas” gaplarida inkor (“emas”) va sifat (“ajoyib”) o‘zaro ta’siri salbiy sentimentni shakllantiradi. ToshDO‘TAU olimlari (2021) tomonidan yaratilgan “UzbSentiment” korpusi asosida o‘tkazilgan tahlillar shuni ko‘rsatdiki, gap darajasidagi modellar so‘z darajasidagi tahlilga nisbatan 30-35% yuqori aniqlikni ta’minlaydi. Biroq, o‘zbek tilida gap tuzilmalarini modellashtirish uchun morfologik va sintaktik tahlilni integratsiyalash talab etiladi. Maqolada taklif etilayotgan metodologiya – so‘z turkumlarini asos qilib, ularning gap ichidagi o‘zaro ta’sirini graf neyron tarmoqlari yordamida modellashtirish – o‘zbek NLP sohasida yangi yechimlarni taklif qiladi. Tadqiqotning amaliy ahamiyati shundaki, u ijtimoiy tarmoqlardagi fikr-mulohazalarni avtomatik tahlil qilish, mijozlarning sharhlarini klassifikatsiyalash kabi amaliy vazifalarga yechim bo‘ladi.

Gap modellarini ishlab chiqish

O‘zbek tilidagi matnlarni so‘z turkumlari orqali sentiment tahlil qilishda eng muhim so‘z turkumlari: sifat, ravish, fe’l hisoblanadi. Shuningdek, sentiment tahlilga qo‘shimcha ta’sir qiluvchi so‘zlar: modal so‘zlar, otlar, inkor shakllari bo‘lib, ushbu leksik birliklar asosida gap modellarini ishlab chiqish mumkin. O‘zbek tilida gap modellarini ishlab chiqishda so‘z turkumlarining quyidagicha kombinatsiyalari ishlab chiqildi:

1-jadval. Sentiment gap modellari

№	Gap modeli	Misol	Sentiment
1	**Ot +Sifat**	Ushbu **film ajoyib**.	5
2	**Ravish+Fe’l**	Men xizmatdan **juda mamnunman**.	4
3	**Sifat+Inkor**	Ushbu joy **qulay emas**.	-2
4	**Kuchaytiruvchi + Sifat**	Bu yer **mutlaqo go‘zal**.	5
5	**Zaiflashtiruvchi + Sifat**	Bu yer **shunchaki yaxshi**.	2

6	**Fe’l+Inkor**	Men bu mahsulotni **yoqtirmayman**.	-3
7	**Fe’l + Ot**	Ular bizga **yordam berdi**.	3
8	**Ot + Fe’l + Inkor**	Ular bizga **yordam bermadi.**	-3
9	**Ravish + Fe’l + Inkor**	Bu menga **juda yoqmadi.**	-4
10	**Kuchaytiruvchi (R) + Fe’l**	Men bu natijaga **nihoyatda xursandman.**	5
11	**Zaiflashtiruvchi (R) + Fe’l**	Men bu natijadan **bir oz xursandman.**	2
12	**Sifat +Fe’l + Inkor**	Bu ovqat **mazali chiqmadi.**	1
13	**Sifat + Fe’l**	O‘qituvchi juda **mehribon ko‘rinadi.**	4
14	**Sifat + Inkor + Fe’l**	O‘qituvchi **mehribon emasdek ko‘rinadi.**	-2
15	**Ot + Kuchaytiruvchi + Fe’l**	Bu **natija juda katta motivatsiya berdi.**	5
16	**Ot+ Zaiflashtiruvchi + Fe’l**	Bu **natija ozgina kuch berdi.**	2
17	**Ravish + Sifat + Ot**	Bu yangilik **nihoyatda yaxshi xabar.**	5
18	**Ravish + Sifat + Ot + Inkor**	Bu yangilik **umuman yaxshi xabar emas.**	-4
19	**Ot + Kuchaytiruvchi (R) + Fe’l**	Meni ularning **g‘alabasi nihoyatda quvontirdi.**	5
20	**Ot + Zaiflashtiruvchi (R) + Fe’l**	Nozimaning **g‘alabasidan shunchaki quvondi.**	2
21	**Inkor + Fe’l + Ot + Kuchaytiruvchi**	Mag‘lubiyat **menga umuman ta’sir qilmadi.**	1
22	**Olmosh + Zaiflashtiruvchi (R) + Fe’l + Inkor**	Mag‘lubiyat **menga deyarli ta’sir qilmadi.**	0
23	**Kuchaytiruvchi (R) + Ot**	Bu yil menga **nihoyatda quvonch** olib keldi.	5
24	**Zaiflashtiruvchi (R) + Ot**	U **biroz tushkunlikda** edi.	1
25	**Olmosh + Kuchaytiruvchi (R) + Sifat + Fe’l**	Bu xabar **barchani juda xursand qildi.**	5
26	**Olmosh + Sifat + Fe’l + Inkor**	Bu xabar **barchani xursand qilmadi.**	-2

Yuqoridagi jadvalda keltirilgan gap modellarida kuchaytirgichlar, zaiflashtirgichlar va inkor shakllari sentimentga qanday ta’sir qilishini aks ettirildi. **Fe’llar, sifatlar, otlar va ravishlarning kombinatsiyasi** orqali sentiment ifodalari

yanada aniq baholanishi mumkin. **Inkor shakllari ba’zan sentimentni neytralga yaqinlashtirishi yoki o‘zgartirishi mumkin.**

Ushbu modellarni aniqligini oshirish uchun **turli uslubdagi gaplar**: **rasmiy, norasmiy, adabiy** va **shevada ishlatiladigan gap shakllarini** qo‘shish lozim. Shuningdek, modelda **murakkab gaplar**: qo‘shma va ergash gaplarni ham qo‘shib, sentiment tahlilining aniqligini oshirish mumkin.

Sentiment tahlil qilish modeliga neytral, ironik va sarkastik so‘zlarni qo‘shish orqali model aniqligi va samaradorligini oshirish mumkin.

1. Qo‘shma gaplar modeli:
Gap qolipi: *[Asosiy gap] + bog‘lovchi + [Qo‘shimcha gap]*
Misol: *U juda aqlli*, **lekin ba’zan xatolar ham qiladi**.

Variatsiyalar:

- **va** bog‘lovchisi: *Men kitob o‘qiyapman **va** musiqa tinglayapman.*
- **lekin** bog‘lovchisi: *U yaxshi ishlaydi, **lekin** ba’zan unutib qo‘yadi.*
- **yoki** bog‘lovchisi: *Siz kelishingiz mumkin **yoki** biz telefon orqali gaplashamiz.*

2. Ergash gaplar modeli:
Gap qolipi: *[Asosiy gap] + ergash gap bog‘lovchisi + [Ergash gap]*
Misol: ***Agar** u kechiksa,* **biz kutishimizga to‘g‘ri keladi.**

Variatsiyalar:

- **shart:** ***Agar** u yordam bersa, ishimiz ancha yengillashadi.*
- **sabab:** *U tinmay ishlaydi, **chunki** o‘z ishini yaxshi ko‘radi.*
- **natija:** *Men ko‘p mashq qildim, **shuning uchun** imtihonda yaxshi natija oldim.*
- **maqsad:** *Men kech kelaman, **shunda** hamma ish bitib qo‘yiladi.*

№	Gap modeli	Misol	Sentiment	Uslub
1	**Ot+ Sifat**	Ushbu **film ajoyib** ekan.	5	Norasmiy (so‘zlashuv)
2	**Ravish +Fe’l**	Bu rejani tinglovchilar **ancha ma’qullashdi.**	4	Rasmiy (publisistik

3	**Sifat+ Inkor**	Qilingan amaliy ish foydalanuvchilar uchun **qulay emasligi** aytildi.	-2	Rasmiy (ilmiy)
4	**Kuchaytiruvchi(RR) + Sifat**	Bu yer **mutlaqo go‘zal**.	5	Norasmiy(badiiy)
5	**Zaiflashtiruvchi(RR) + Sifat**	Bu yer **shunchaki yaxshi**.	2	Rasmiy (ilmiy)
6	**Fe’l+ Inkor**	Men bu mahsulotni **yoqtirmayman**.	-3	Norasmiy(so‘zlashuv)
7	**Ot + Sifat + Inkor**	U sotib olgan **kiyim yaxshi emas** edi.	-4	Norasmiy(so‘zlashuv
8	**Ot +Fe’l**	Mazkur buyruq ijrosi ustidan **nazorat o‘rnatilsin.**	3	Rasmiy (rasmiy)
9	**Ot +Fe’l + Inkor**	Eksperiment davomida 50 nafar **ishtirokchi jalb etilmadi.**	-3	Rasmiy (ilmiy)
10	**Ravish +Fe’l + Inkor**	Bu menga **umuman yoqmadi**.	-4	Norasmiy(so‘zlashuv)
11	**Kuchaytiruvchi(RR) + Fe’l**	Men bu natijadan **nihoyatda xursandman**.	5	Norasmiy(publisistik)
12	**Zaiflashtiruvchi (RR)+ Fe’l**	Men bu natijadan **bir oz xursand bo‘ldim.**	2	Norasmiy(publisistik
13	**Ot + Fe’l**	Mening g‘alabam **odamlarni quvontirdi**.	4	Adabiy (Rasmiy)
14	**Ot + Fe’l + Inkor**	Farzandining mag‘lubiyati **otani xafa qilmadi**.	1	Norasmiy(badiiy)
15	**Murakkab gap (qo‘shma: qarama-qarshi bog‘lovchi)**	U juda aqlli, **ammo ba’zan xatolar ham qilib turadi.**	2	Norasmiy(so‘zlashuv)
16	**Murakkab gap (qo‘shma: teng bog‘lovchi)**	Arizangiz ko‘rib chiqildi **va ijobiy hal etildi.**	4	Rasmiy (rasmiy)
17	**Murakkab gap (qo‘shma: ziddiyat bog‘lovchi)**	Havo sovuq edi, **ammo biz yurishni davom ettirdik.**	-1	Norasmiy(badiiy)
18	**Murakkab gap (ergash: shart bog‘lovchi)**	Agar tomonlar rozi bo‘lmasa, **shartnoma bekor qilinishi aytildi.**	0	Rasmiy (rasmiy)

19	**Murakkab gap (ergash: sabab bog‘lovchi)**	U tinmay ishlaydi, **chunki o‘z kasbini yaxshi ko‘radi**.	4	Norasmiy(badiiy)
20	**Murakkab gap (ergash: natija bog‘lovchi)**	Men ko‘p mashq qildim, **shuning uchun imtihonda yaxshi natija oldim**.	5	Norasmiy(so‘zlashuv)
21	**Murakkab gap (ergash: maqsad bog‘lovchi)**	Men ko‘proq kitob o‘qiyapman, **shunda bilimim oshadi.**	2	Norasmiy(so‘zlashuv)
22	**Murakkab gap (qo‘shma: parallel hodisa)**	U ertalab yugurdi, **shuningdek, mashq ham qildi**.	4	Norasmiy(badiiy)
23	**Murakkab gap (ergash: imkon bog‘lovchi)**	Agar u yordam bersa, **ishimiz ancha yengillashadi**.	3	Norasmiy(publisistik)
24	**Murakkab gap (qo‘shma: qarama-qarshi fikr)**	Men bu natijani kutgandim, **ammo u umidimni oqlamadi**.	-2	Norasmiy(publisistik)
25	**Murakkab gap (ergash: ehtimol bog‘lovchi)**	Agar ariza belgilangan muddatda topshirilsa, **ko‘rib chiqilishi mumkin.**	1	Rasmiy(rasmiy)
26	**Murakkab gap (qo‘shma: izohlovchi)**	Ular kech kelishdi, **chunki reja o‘zgarib ketdi**.	-1	Norasmiy(badiiy)
27	**Murakkab gap (qo‘shma: mustahkamlovchi)**	U juda mehnatkash, **hatto kasal bo‘lsa ham ishlaydi**.	3	Norasmiy(publisistik)
28	**Murakkab gap (ergash: vaqt bog‘lovchi)**	Tadqiqot boshlanganidan beri, **natijalar muntazam tahlil qilinmoqda.**	4	Rasmiy(ilmiy)
29	**Murakkab gap (ergash: daraja bog‘lovchi)**	Men bu natijadan shunchalik hursand bo‘ldimki, **hatto yig‘lab yubordim**.	5	Norasmiy(so‘zlashuv)

Matndagi **sifat (JJ)** va **ravish (RR)** so‘z turkumlariga mansub so‘zlar birgalikda tahlil qilinsa, matndagi hissiyotlarning nafaqat mavjudligi, balki ularning qanchalik intensiv ekanligi haqida ham to‘liqroq ma’lumot olish mumkin. Shunday qilib, sentiment tahlilining natijalari yanada aniq va ishonchli bo‘ladi.

Sentiment tahlilida **ravish (RR) + fe’l (VB)** birikmalarini tahlil qilish foydali bo‘lishi mumkin.

1. **Ravish (adverb):** Ravishlar fe’llarning ma’nosini kuchaytiradi yoki o‘zgartiradi. Ular harakatning intensivligi, chastotasi, uslubi yoki vaqtini ko‘rsatishi mumkin. Masalan: *“juda”*, *“aniq”*, *“tezda”*, *“to‘liq”* kabi so‘zlar.

2. **Fe’l (verb):** Fe’llar esa harakat, holat yoki jarayonni ifoda etadi. Ular matnda asosiy voqea yoki harakatni bildiradi. Masalan: *“yoqtiradi”*, *“nafratlanadi”*, “*ko‘radi”*, *“qabul qiladi”* kabi fe’llar.

Ravish+Fe’l birikmasi ravishning qo‘shilishi orqali fe’lning ifoda etayotgan harakatiga qo‘shimcha ma’no, ayniqsa hissiyotning intensivligi yoki uslubiy xususiyatlarini beradi.

Sentiment tahlilida **ravish (RR) + fe’l (VB)** birikmalarini tahlil qilish sabablari quyidagilardan iborat:

1. **Intensivlik:** Ravishlar fe’llarga qo‘shilib, ularning *harakat* yoki *holat* ifodasini kuchaytirishi mumkin. Masalan, “*juda yoqtiradi*” birikmasida “*juda*” hissiyotning kuchini oshiradi va bu matnning ijobiy sentimentini yanada kuchaytiradi.

2. **Kontekstual nuanslarni aniqlash:** Har bir fe’l o‘zida ma’lum bir sentimentni bildirishi mumkin, lekin uning oldidagi ravish orqali bu hissiyot aniqroq ifodalanadi. Masalan, “*sevadi*” va “*chindan sevadi*” birikmalari farqli darajada ijobiy his-tuyg‘uni aks ettirishi mumkin.

3. **Kontekstni aniqlashtirish:** Fe’lning o‘zida ham, uning oldidagi ravishda ham sentiment mavjud bo‘lishi mumkin. Ravishning qo‘shilishi matndagi *hissiyotning darajasi* va *ifodalanish usulini* yanada boyitadi, shu orqali sentimentning to‘liq manzarasi paydo bo‘ladi. Ba’zi hollarda, ravishlar fe’lning sentiment **polaritesini** (ya’ni, ijobiy yoki salbiyligini) aniqlashda yordam beradi. Ravish qo‘shilishi, ayniqsa, salbiy holatlarda, masalan “*juda salbiy baholaydi*” kabi, hissiyotning salbiy tomonini kuchaytiradi.

4. **Matn tahlilining aniqligi:** Bu turdagi kombinatsiyalarni e’tiborga olish orqali sentiment tahlili algoritmlari yanada aniqroq va ishonchli qarorlar qabul qilishi mumkin. Chunki, bunday kombinatsiyalar ko‘pincha matndagi hissiyotning *subtil* va *kontekstual* jihatlarini ochib beradi.

Matnda bir nechta **ravish (RR) + fe’l (VB)** birikmalari bo‘lishi mumkin. Ularni tahlil qilish orqali har bir harakat yoki holatning hissiyot darajasini solishtirish va umumiy matnning sentiment qiymatini aniqlash mumkin. Shunday qilib, agar sentiment tahlil jarayonida mos keladigan usullar va resurslar bo‘lsa, **ravish (RR) + fe’l (VB)** birikmalarini ham alohida e’tiborga olish sentiment tahlilining *aniqligini* oshirishi mumkin.

So‘z birikmasi	Ravish	Fe’l
juda yoqtiradi	“juda” – bu his-tuyg‘uni kuchaytiradi, ya’ni foydalanuvchi narsa yoki holatni	“yoqtiradi” – ijobiy his-tuyg‘uni ifodalaydi.

	juda ijobiy qabul qilayotganini bildiradi.	
aniq rad etadi	“aniq” – bu rad etish jarayonining qat’iyligini, aniq va inkorlovchi munosabatni ko‘rsatadi.	“rad etadi” – salbiy yoki inkor qiluvchi his-tuyg‘uni ifodalaydi.
tezda unutdi	“tezda” – bu jarayonning qanchalik tez sodir bo‘lishini ifoda etadi va ba’zi kontekstlarda hissiyotning intensivligini (masalan, tezda kechgan yoki tezda unutgan his-tuyg‘u) aniqlashga yordam beradi.	“unutdi” – neytral yoki holatni o‘zgartirishni bildirishi mumkin.

Matndagi **ravish (RR) + fe’l (VB)** birikmalarni aniqlash uchun quyidagi ketma-ketlikda amallarni bajarish lozim:

1. **POS teglash (Part-of-Speech tagging):** Matndagi so‘zlar (so‘z birikmalari) va ularga mos turkumlarini aniqlash uchun POS teglash vazifasidan foydalaniladi. POS teglash orqali ravishlar va fe’llarni aniqlash, ularning kombinatsiyalarini ajratib olish mumkin.

2. **N-gram: ravish (RR) + fe’l (VB)** kombinatsiyalarini aniqlash uchun bigram
(2-gram) yondashuvi qo‘llanilishi mumkin. Bu yondashuv yordamida matn ichidagi qo‘shma so‘z birikmalari ajratib olinadi va ularga sentiment qiymatlar beriladi.

3. **Sentiment lug‘atlar va qoida asosida tahlil:** Agar siz maxsus sentiment lug‘atlaridan foydalanayotgan bo‘lsangiz, ularda ravishlar va fe’llarga tegishli sentiment qiymatlari bo‘lishi mumkin. Shu qiymatlar orqali kombinatsiyalar umumiy hissiyotga ta’sirini baholash mumkin.

4. **Mashinali o‘qitish modellari:** Agar siz mashinali o‘qitishga asoslangan yondashuvni qo‘llayotgan bo‘lsangiz, model treningi (o‘qitish) jarayonida ravish+fe’l kombinatsiyalari uchun *xususiyat (feature)* sifatida ekstraksiya qilinishi va modelga kiritilishi mumkin. Bu, masalan, *so‘z o‘zgarishi, kombinatsiya intensivligi, kontekstual bog‘lanishlarni o‘rganish* orqali amalga oshiriladi.

Xulosa

Sentiment tahlilning asosiy vazifasi matndagi hissiy tonni (ijobiy, salbiy, neytral) aniqlashdan iborat bo‘lib, bunda so‘zlar semantik va pragmatik jihatdan asosiy indikator hisoblanadi. So‘z turkumlari (ot, sifat, fe’l, ravish va boshqalar) sentimentni ifodalashda turlicha rol o‘ynashi lingvistik tadqiqotlarda ko‘p marta ta’kidlangan. O‘zbek tilida bu masala yanada murakkabdir, chunki tilning agglutinativ tuzilishi so‘z shakllarining ko‘pligiga, morfologik bog‘liqliklarga va kontekstual nuanslarga olib keladi. Masalan, sifatlar (“ajoyib”, “yomon”) to‘g‘ridan-to‘g‘ri emotsional yuklangan bo‘lsa, otlar (“zavq”, “qayg‘u”) ko‘pincha abstrakt

tushunchalarni ifodalaydi. Fe’llar esa (“yoqtirmoq”, “yomon ko‘rmoq”) nutqdagi harakatning hissiy yo‘nalishini ko‘rsatadi. O‘zbek tilida ravishlar (“juda”, “ortiqcha”) emotsional intensivlikni kuchaytirishda muhim rol o‘ynaydi. Biroq, so‘z turkumlarining sentimentga ta’sirini tahlil qilishda ikkita asosiy muammo mavjud: birinchidan, bir xil so‘z turkumining turli kontekstlarda qarama-qarshi ma’nolar kasb etishi (masalan, “do‘st” – ijobiy, lekin “soxta do‘st” – salbiy); ikkinchidan, kodli almashish (rus/ingliz tillaridan so‘zlar) tufayli semantik noaniqlik. O‘zbekcha so‘z turkumlarini morfologik jihatdan normalizatsiya qilish (lemmatizatsiya) orqali ularni sentiment lug‘atlariga integratsiyalash imkoniyatlarini ko‘rsatdi. Ushbu tadqiqotda o‘zbek tilida so‘z turkumlarining sentiment tahlilga ta’sirini tizimli o‘rganilib, ularni modellashtirishda samarali yondashuvlarni taklif qilindi. Natijalar nafaqat lug‘atli tizimlarni boyitish, balki neyron tarmoqlarning kontekstni tushunish qobiliyatini oshirishga xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in information retrieval*, *2*(1–2), 1-135.
2. Elov B. and Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).
3. Xusainova Z. (2023). O‘zbek tili milliy korpusi qidiruv tizimini optimallashtirishda lemmatizatsiyadan foydalanish. Uzbekistan: Language and Culture, 2(2).
4. Elov B., Alayev R. and Abdullayev A. (2024). Tabiiy tilning statistik modellari. digital transformation and artificial intelligence, 2(6), 178-189.
5. Tang D., Wei F., Qin B., Yang N., Liu T. and Zhou M. (2015). Sentiment embeddings with applications to sentiment analysis. IEEE transactions on knowledge and data Engineering, 28(2), 496-509.
6. Cambria E. (2024). Semantics Processing. In Understanding Natural Language Understanding (pp. 113-228). Cham: Springer Nature Switzerland.
7. Mengliev D., Abdurakhmonova N., Barakhnin, V. Vasliddinova, K. Rahimov H. and Djalolova, K. (2024, June). Enhancing sentiment analysis in Uzbek language texts through weighted lexical features. In 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM) (pp. 2450-2453). IEEE.

SENTIMENT TAHLIL UCHUN KATTA HAJMDAGI DATASETNI YARATISH BOSQICHLARI

Elov Botir Boltayevich,
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Abdullayev Abdulla Quranbayevich,
NTM ta’limni kredit tizimini boshqarish bo‘lim boshlig‘i
abdulla_abdullayev9270@mail.ru
Urganch innovatsion university

Annotatsiya. O‘zbek tilida sentiment tahlilni rivojlantirishning asosiy to‘siqlaridan biri – lingvistik xususiyatlar va yetarli hajmdagi sifatli ma’lumotlar to‘plamining yo‘qligidir. Ushbu maqolada sentiment tahlil uchun katta hajmdagi datasetni yaratish jarayoni uch bosqichli tizim asosida tahlil qilinadi. Avvalo, birinchi bosqichda yaratiladigan datasetning maqsadi, sifat mezonlari va foydalanish sohasi aniqlanib, ilmiy talablar ishlab chiqiladi. Ikkinchi bosqichda, turli axborot manbalari – ijtimoiy tarmoqlar, forumlar, bloglar, onlayn sharhlar va yangilik portallaridan xom ma’lumotlar yig‘ilib, ularning dolzarbligi va ishonchliligi tahlil qilinadi. Uchinchi bosqichda esa, yig‘ilgan ma’lumotlar tozalash, normalizatsiya, tokenizatsiya va lemmatizatsiya kabi ishlov berish operatsiyalari yordamida mos va strukturallashgan dataset shakllantiriladi. Shuningdek, annotatsiya va balanslash jarayonlari orqali har bir sentiment kategoriyasining teng va to‘liq aks etishi ta’minlanadi. Ushbu tizimli yondashuv ma’lumotlarning lingvistik, semantik va statistik jihatlarini hisobga olib, sentiment tahlil modellari uchun yuqori aniqlikdagi korpus yaratishga xizmat qiladi. Tadqiqot natijalari yaratilgan datasetning sifatli va ishonchli ekanligini, shuningdek, sentiment tahlil algoritmlarining samaradorligini oshirishga hissa qo‘shishini ko‘rsatadi. Ushbu tadqiqot sentiment tahlil uchun katta hajmdagi datasetni yaratishning ilmiy asoslangan bosqichlarini taqdim etib, sohada innovatsion yechimlarni ilgari suradi.

Kalit so‘zlar: *Sentiment tahlil, NLP, dataset yaratish, ma’lumotlarni tayyorlash, katta hajmdagi dataset, lingvistik annotatsiya, Inter-annotator kelishuv.*

Kirish

Tabiiy tilni qayta ishlash sohasida sentiment tahlilning yuqori aniqlik va umumlashuv qobiliyatiga erishishda **katta hajmdagi, yuqori sifatli** datasetning yaratilishi nihoyatda muhim ahamiyatga ega. Avvalo, turli axborot manbalaridan – ijtimoiy tarmoqlar, forumlar, onlayn sharhlar va yangilik portallaridan – xom ma’lumotlar yig‘ilib, korpusning asosiy poydevori shakllantiriladi. Keyin, yig‘ilgan

ma’lumotlar tozalash jarayoni orqali HTML teglar, maxsus belgilar, ortiqcha raqamlar va keraksiz so‘zlardan xoli holga keltiriladi, bu esa keyingi bosqichlar uchun ishonchli ma’lumot zaxirasini ta’minlaydi. Shu bilan birga, tokenizatsiya, normalizatsiya va lemmatizatsiya kabi boshlang‘ich ishlov berish operatsiyalari qo‘llanilib, matnlarning lingvistik va semantik tuzilishi soddalashtiradi. Annotatsiya bosqichida esa, har bir matn sentiment bo‘yicha ijobiy, salbiy yoki neytral deb belgilanadi, bu usul yarim avtomatik va qo‘lda amalga oshirilgan annotatsiya tizimlari yordamida bajariladi. Ma’lumotlar sifatini yanada oshirish maqsadida, dataset balanslash bosqichida turli sentiment kategoriyalarining teng taqsimlanishi statistik metodlar orqali ta’minlanadi. So‘nggi bosqichda, annotatsiyalangan va balanslangan ma’lumotlar trening, test va validatsiya to‘plamlariga ajratilib, modelni o‘rgatish uchun mustahkam asos yaratiladi. Modelni o‘rgatish jarayonida, katta hajmdagi ma’lumotlar yordamida algoritm treningi amalga oshirilib, gradient tushishi, kross-validatsiya va boshqa optimallashtirish strategiyalari orqali model parametrlarida yaxshilanishlar kiritiladi. Keyinchalik, modelni baholash bosqichida aniqlik, precision, recall va F1-score kabi metrikalar asosida eksperimental sinovlar o‘tkazilib, yaratilgan modelning amaliy qo‘llanilishidagi samaradorligi aniqlanadi. Natijada, yaratilgan dataset va o‘rgatilgan model real sharoitlarda sentiment tahlilini avtomatlashtirish imkonini yaratib, yuqori aniqlik va barqarorlikka erishishni ta’minlaydi. Shuningdek, modelni joriy qilish bosqichi orqali tizimga integratsiyalashgan yechim real vaqtda sentiment tahlilini amalga oshirishga imkon yaratadi. Ushbu tizimli yondashuv, ma’lumotlarni tayyorlash, tozalash, annotatsiya, balanslash, modelni o‘rgatish, baholash va joriy qilish bosqichlarining o‘zaro bog‘liqligini ishonchli tarzda namoyish etadi.

Yuqorida ta’kidlanganidek, zamonaviy sentiment tahlil tizimlarining samaradorligi, qaysi algoritmdan foydalanilmasin, asosan uning o‘qitiladigan ma’lumotlar **bazasining hajmi, sifati** va **representativligiga** bog‘liq. O‘zbek tilidagi matnlar uchun katta hajmdagi ma’lumotlar to‘plamini yaratish sohasi nisbatan yangi bo‘lib, tilning agglutinativ tuzilishi, kodli almashish (rus/ingliz tillaridan so‘zlar qo‘shilishi) kabi lingvistik xususiyatlar tufayli qiyinchiliklar tug‘diradi.

Xalqaro tajribalar (Hovy va Lavid, 2010[1]; Cambria, 2016[2]) shuni ko‘rsatadiki, ma’lumotlar to‘plamini shakllantirishda muvozanatli tanlov, to‘g‘ri annotatsiya va turli kontekstlarni qamrab olish algoritmlarning adolatli ishlashi uchun hal qiluvchi ahamiyatga ega. O‘zbek tiliga oid mavjud korpuslar (masalan, “UzbCorpus”) ko‘pincha cheklangan hajmga ega yoki ma’lum bir soha (ijtimoiy media, yangiliklar) bilan chegaralangan, bu esa ko‘p qirrali sentiment tahlilni qiyinlashtiradi[3]. Ma’lumotlarni yig‘ish bosqichida manbalarning xilma-xilligi (forumlar, ijtimoiy tarmoqlar, rasmiy hujjatlar) va demografik tarqalishi (yosh, mintaqaviy dialektlar) to‘plamning universalligini ta’minlashda muhim rol

o‘ynaydi. Elov va boshq. (2024) ta’kidlaganidek, o‘zbek tilida morfologik normalizatsiya (lemmatizatsiya, stemming) va tokenizatsiyani to‘g‘ri amalga oshirish matnlarni kompyuter tushunishi uchun moslashtirishning asosiy sharti hisoblanadi[4]. Tovush yozuvi, imlo xatolari va qisqartmalar kabilarni tozalash (Kuryozov va boshq., 2019) esa ma’lumotlarni vektorlashtirish jarayonidagi shovqinlarni kamaytiradi[5]. Annotatsiya bosqichida mahalliy til mutaxassislari ishtiroki (Mengliyev va boshq., 2023) emotsional nuanslarni to‘g‘ri belgilash imkonini beradi[6]. Bundan tashqari, turli annotatorlar orasidagi kelishuv darajasini (inter-annotator agreement) o‘lchash (Bobicev va boshq., 2017) ma’lumotlar sifatini baholashda muhim metrikadir[7]. Ma’lumotlarni tasodifiy bo‘linishi (train-test split), sinov to‘plamlarining diversifikatsiyasi (Hamdamov va boqsh., 2024) modelning overfittingdan himoyalanishiga yordam beradi[8]. Yana bir muammo – ma’lumotlar to‘plamini dinamik yangilab borish, chunki til tirik tizim bo‘lib, yangi so‘zlar va iboralar doimiy paydo bo‘ladi. Xalqaro standartlarga moslab ma’lumotlarni ochiq manbali ravishda taqdim etish (masalan, Kaggle, Hugging Face) ham O‘zbek tilidagi NLP hamjamiyatining rivojiga hissa qo‘shadi. Xulosa qilib, keng ko‘lamli, sifatli va diversifikatsiyalangan ma’lumotlar to‘plamini yaratish – bu nafaqat sentiment tahlil, balki boshqa NLP vazifalari uchun ham o‘zbek tilining raqamli resurslarini boyitishga qaratilgan muhim ilmiy-amaliy harakatdir.

Katta hajmdagi datasetni yaratishda quyidagi bosqichlarni amalga oshirish lozim:

1. Dataset uchun asosiy talablarni ishlab chiqish.

Ushbu bosqichda, avvalo, yaratiladigan datasetning maqsadi, qo‘llanilishi sohalari va ilmiy talablarini aniq belgilash lozim. Tadqiqot kontekstida, dataset tarkibida kutilayotgan semantik, sintaktik va lingvistik xususiyatlar, shuningdek, sentimentning aniqligi va granulyatsiyasi (masalan, ijobiy, salbiy, neytral yoki nozik hissiyot darajalari) aniqlanadi. Shuningdek, ma’lumotlarning to‘liqligi, dolzarbligi, aniqligi va moslashuvchanligi kabi sifat mezonlari belgilab olinadi. Ushbu talablar asosida, keyingi bosqichlarda ishlatiladigan annotatsiya protokollari, sinov metrikalari va modelning o‘lchov ko‘rsatkichlari ham ishlab chiqiladi, bu esa yaratilayotgan datasetning ilmiy jihatdan mustahkam poydevorini tashkil etadi.

Sentiment tahlil uchun dataset yaratishning dastlabki bosqichi – maqsadga mos funksional va texnik talablarni aniqlashdir. Bu bosqichda quyidagilar ko‘rib chiqiladi: **Maqsadni aniqlash:** Dataset sentiment tahlilning qaysi turi (binary: ijobiy/salbiy, yoki ko‘p toifali: neytral, aralash, emotsional darajalar) uchun mo‘ljallangan? **Lingvistik talablar:** O‘zbek tilining agglutinativligi, dialektlar (Farg‘ona, Toshkent shevalari), kodli almashish (rus/ingliz tillaridan so‘zlar) kabi xususiyatlar hisobga olinadi. **Hajm va diversifikatsiya:** Kamida 100000 ta annotatsiyalangan matn, turli janrlar (ijtimoiy media, rasmiy matnlar, nutq) va

demografik guruhlar (yosh, jins) qamrovini ta’minlash. **Etika va qonuniy talablar:** Shaxsiy ma’lumotlarni olib tashlash (GDPR), mualliflik huquqi (CC BY litsenziya) va anonymizatsiya. **Annotatsiya protokoli:** Inter-annotator kelishuv (Cohen’s kappa > 0.8) va mahalliy ekspertlarni jalb qilish. Datasetni shakllantirishda quyidagi talablarga rioya qilish kerak:

- **Gaplar**: O‘zbek tilidagi real gaplardan iborat bo‘lishi lozim.
- **Sentiment teglari**: Datasetdagi har bir gap "ijobiy", "salbiy", yoki "neytral" kabi teglar vositasida teglanishi lozim.
- **Dataset hajmi**: Katta hajmdagi dataset yaxshi natijalar olishga yordam beradi. "Ijobiy", "salbiy", yoki "neytral" kabi har bir kategoriyadan kamida **1000-5000 ta gap** bo‘lishi maqsadga muvofiq.

2. Dataset manbalari.

Ushbu bosqichda, kerakli ma’lumotlarni yig‘ish uchun foydalaniladigan axborot manbalari aniqlanadi va ularning ilmiy asosda tahlil qilinishi amalga oshiriladi. Tadqiqot doirasida, ijtimoiy tarmoqlar, forumlar, bloglar, onlayn sharhlar va yangilik portallari kabi turli manbalar orqali xom ma’lumotlar to‘planadi. Har bir manbaning ishonchliligi, dolzarbligi va ma’lumotlarning strukturaliligi baholanib, ularning datasetga qo‘shilishi mumkinligi aniqlanadi. Shuningdek, manbalar orasida lingvistik va madaniy farqlarni inobatga olish zarur, chunki bu sentiment tahlilning yakuniy natijalariga sezilarli ta’sir ko‘rsatadi. Manbalarni tanlash jarayonida, mavjud ilmiy adabiyotlar, statistik tahlillar va avvalgi tajribalar asosida optimal manba to‘plami shakllantiriladi. Matnlar **turli manbalardan**, **turli mavzulardan** va **turli kontekstlardan** bo‘lishi muhim. Datasetni quyidagi manbalardan shakllantirish mumkin:

- **Ijtimoiy tarmoqlar**: Telegram, Twitter, Facebook, yoki Instagram ijtimoiy tarmoqlardagi o‘zbek tilidagi postlar yoki izohlar (odamlarning fikrlari va sharhlari).
- **Sharhlar**: Mahsulot sharhlari (masalan, Oson, Uzum, ZoodMall kabi onlayn do‘konlar yoki restoran sharhlari) va kino sharhlarini shakllantirish (ijobiy va salbiy fikrlarni ajratish uchun).
- **Yangilik saytlari**: kun.uz, daryo.uz, gazeta.uz kabi onlayn nashrlar (neytral matnlar uchun ideal manba).
- **Bloglar va maqolalar**: O‘zbek tilidagi bloglardan yoki yangiliklar saytlaridan gaplarni to‘plash zarur (odamlar haqqoniy fikr bildiradigan manba).
- **YouTube izohlari**: Video sharhlari orqali real his-tuyg‘ularni aniqlash.
- **Sun’iy yozilgan gaplar**: Agar haqiqiy ma’lumotlar yetarli bo‘lmasa oddiy gaplar yozib, ularga sentiment teg berish kerak.

3. Datasetni shakllantirish.

Ushbu bosqichda, yig‘ilgan xom ma’lumotlar asosida datasetni strukturaviy, semantik va statistik jihatdan boyitish jarayoni amalga oshiriladi. Avvalo, xom ma’lumotlar tozalash, normalizatsiya, tokenizatsiya va lemmatizatsiya kabi boshlang‘ich ishlov berish jarayonlari orqali mos va yagona formatga keltiriladi. Keyin, annotatsiya bosqichida, har bir matn sentiment bo‘yicha (ijobiy, salbiy, neytral yoki yanada nozik darajalar) qo‘lda yoki yarim avtomatik tarzda belgilab qo‘yiladi. Ushbu jarayon uchun inter-annotator agreement kabi sifat nazorati mexanizmlari qo‘llaniladi. Shuningdek, datasetni balanslash bosqichi orqali, har bir sentiment kategoriyasining teng taqsimlanishini ta’minlash uchun statistik metodlar, masalan, over-sampling yoki under-sampling usullari qo‘llaniladi. Natijada, tayyor dataset trening, test va validatsiya to‘plamlariga ajratilib, keyingi modellashtirish bosqichlariga tayyor holda ilmiy asosda shakllantiriladi. Datasetni shakllantirishda qo‘lda yoki avtomatik vositalardan foydalanish mumkin. Bu bosqichda quyidagi amallar bajariladi:

- **Ma’lumotlarni yig‘ish**: Saytlar va platformalardan ma’lumotlarni yig‘ish uchun **web scraping** usulidan foydalanish mumkin (BeautifulSoup, Selenium, Twitter API, YouTube API).
- **Ma’lumotlarni tozalash**: Avtomatik top`langan matnlarni tozalash lozim. Bu jarayonda matnlardan noto‘g‘ri belgilar, html teglar, va ortiqcha so‘zlar olib tashlanadi.
- **Ma’lumotlarni teglash**: Har bir gap uchun sentimentni aniqlash va uni teglash. Bu jarayon ham qo‘lda yoki avtomatik tarzda amalga oshiriladi. Odatda kichik dataset uchun matnlar qo‘lda teglanadi. Kattaroq hajmdagi dataset matnlari dastlab avtomatik teglanadi va keyin qo‘lda qayta ko‘rib chiqiladi.
- **Matnlarning xilma-xilligi. Mavzu bo‘yicha muvozanatni ta`minlash uchun** matnlar turli sohalarni qamrab olishi kerak. Masalan, mahsulot sharhlari – elektronika, kiyim-kechak, kitoblar; xizmatlar – ta’lim, mehmonxonalar, restoranlar; siyosiy va ijtimoiy sharhlar.
- **Tasnif bo‘yicha ajratish (Sentiment bo‘yicha muvozanat)**: Har bir kategoriya uchun ijobiy, salbiy va neytral gaplarni teng (tekis) taqsimlash lozim. Masalan quyidagicha tasniflash usulidan foydalanish mumkin:
  - *juda ijobiy (20-25%)*
  - *ijobiy (20-25%)*
  - *neytral (20-25%)*
  - *salbiy (20-25%)*
  - *juda salbiy (5-10%)*
- **Matn uzunligi:** Datasetni shakllantirishda matn uzuligi ham muhim ahamiyat kasb etadi. Masalan quyidagicha tasniflash usulidan foydalanish mumkin:

- *qisqa matnlar (1-2 gap) (30-35%)*
- *o‘rta matnlar (3-5 gap) (30-35%)*
- *uzun matnlar (5+ gap) (30-35%)*

▪ **Saqlash**: Datasetni .csv yoki .json formatida yoki ma`lumotlar bazasi jadvallarida saqlash lozim.

Ushbu tizimli yondashuv orqali yaratilgan dataset, sentiment tahlilning yuqori aniqligi va ishonchliligini ta’minlashda muhim omil bo‘lib, kelgusidagi ilmiy tadqiqotlar va amaliy ilovalar uchun mustahkam metodologik poydevor sifatida xizmat qiladi. Datasetni ochiq manbali platformalar (Hugging Face, Zenodo) ga joylashtirish va litsenziya (CC BY-SA 4.0) bilan ta’minlash mumkin.

Dataset namunasi

1-jadval. O‘zbek tili gaplarini teglash formati

№	Gap	Sentiment
.	Bu kitob juda qiziqarli ekan.	ijobiy
.	Xizmat sifati meni qoniqtirmadi.	salbiy
.	Bugungi ob-havo juda yaxshi edi.	neytral
.	Bu film vaqtni behuda ketkazish edi.	salbiy
.	Ovqat juda mazali edi.	ijobiy
.	Pulimni qaytarishmadi, xizmat yomon!	salbiy
.	Yetkazib berish kechikdi, lekin mahsulot yaxshi.	salbiy

Xulosa

O‘zbek tilida sentiment tahlilni rivojlantirishda katta hajmdagi datasetni yaratishning asosiy bosqichlari (talablarni belgilash, manbalarni tanlash, shakllantirish) tilning lingvistik xususiyatlari va amaliy muammolarni hisobga olgan holda ishlab chiqilishi kerak. Tadqiqot shuni ko‘rsatdiki, ma’lumotlarni yig‘ishda ijtimoiy tarmoqlar, mavjud korpuslar va demografik diversifikatsiya muhim rol o‘ynaydi. Datasetni shakllantirish bosqichida morfologik normalizatsiya (lemmatizatsiya, stemming) va kodli almashishni hisobga oladigan tozalash algoritmlari samaradorlikni oshirishga yordam beradi. Annotatsiya jarayonida mahalliy ekspertlar ishtiroki va inter-annotator kelishuv (Cohen’s kappa >0.8) dataset sifatini ta’minlaydi. Yaratilgan datasetning muvaffaqiyati, shuningdek, sinov to‘plamlari orqali tekshirilgan statistik muvozanat va dinamik yangilash mexanizmlariga bog‘liq. Natijada, o‘zbek tiliga mo‘ljallangan keng ko‘lamli dataset nafaqat sentiment tahlil, balki chat-botlar, avtomatik tarjima kabi NLP vazifalari uchun asos bo‘lishi mumkin. Biroq, dialektlar, yangi so‘zlar va semantik noaniqliklar kabi muammolar datasetni doimiy yangilab borishni talab qiladi. Kelajakda ko‘p tilli transformer modellarni (masalan, BERT) o‘zbek tiliga

moslashtirish va ochiq manbali hamjamiyatni rivojlantirish sohaga yangi imkoniyatlar yaratadi. Xulosa qilib, ushbu ish o‘zbek tilining raqamli resurslarini boyitishga qaratilgan muhim qadam hisoblanadi.

Foydalanilgan adabiyotlar:

1. Hovy, E., & Lavid, J. (2010). Towards a ‘science’of corpus annotation: a new methodological challenge for corpus linguistics. International journal of translation, 22(1), 13-36.
2. Cambria, E., Hazarika, D., Poria, S., Hussain, A., & Subramanyam, R. B. V. (2018). Benchmarking multimodal sentiment analysis. In Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II 18 (pp. 166-179). Springer International Publishing.
3. Elov, B., & Xusainova, Z. (2024). Til korpuslarini lingvistik teglash bosqichlari. Computer Linguistics: problems, solutions, prospects, 1(1).
4. Boltayevich, E. B., Yuldashevna, X. Z., Mamurjonovna, U. S., Ermamatovich, N. S., qizi, A. Sh. A., & Shavkatjon, M. (2024, October). Algorithms for Parsing Roots and Stems of Words in Uzbek Language. In 2024 9th International Conference on Computer Science and Engineering (UBMK) (pp. 126-130). IEEE.
5. Kuriyozov, E., Matlatipov, S., Alonso, M. A., & Gómez-Rodríguez, C. (2019, May). Construction and evaluation of sentiment datasets for low-resource languages: The case of Uzbek. In Language and Technology Conference (pp. 232-243). Cham: Springer International Publishing.
6. Mengliev, D. B., Akhmedov, E. Y., Barakhnin, V. B., Hakimov, Z. A., & Alloyorov, O. M. (2023, November). Utilizing lexicographic resources for sentiment classification in Uzbek language. In 2023 IEEE XVI International Scientific and Technical Conference Actual Problems of Electronic Instrument Engineering (APEIE) (pp. 1720-1724). IEEE.
7. Bobicev, V., & Sokolova, M. (2017). Inter-annotator agreement in sentiment analysis: machine learning perspective. In Recent Advances in Natural Language Processing (pp. 97-102).
8. Hamdamov, O‘., Elov, B., & Alayev, R. (2024). Ma’lumotlar to‘plamini o‘qitish, baholash va test to‘plamlariga ajratish usullari. Digital transformation and artificial intelligence, 2(6), 107-116.

LEXICON-BASED VA MACHINE LEARNING-BASED MODELLARI ASOSIDA SENTIMENT TAHLILNI AMALGA OSHIRISH

Elov Botir Boltayevich,
Texnika fanlari bo‘yicha falsafa doktori PhD, dotsent
elov@navoiy-uni.uz
ToshDO‘TAU

Malika Suyunova Odil qizi,
Tayanch doktarant
malikasuyunova0@gamil.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada sentiment tahlili bo‘yicha mavjud yondashuvlar, xususan, leksikonga asoslangan va mashinani o‘rganishga asoslangan metodlar tahlil qilinadi hamda ularning samaradorligi o‘rganiladi. Sentiment tahlili tabiiy tilni qayta ishlashning muhim yo‘nalishlaridan biri bo‘lib, u matnlarning hissiy kayfiyatini (ijobiy, salbiy yoki neytral) aniqlash uchun qo‘llaniladi. Maqolada leksikona asoslangan yondashuv so‘zlarning oldindan belgilangan lug‘atlar va sentiment resurslari (masalan, WordNet) asosida baholanishini tushuntiradi. Ushbu metodda so‘zlarning hissiy qiymati oldindan belgilanib, ulardan foydalanib matnning umumiy sentimenti aniqlanadi. Shuningdek, tadqiqotda mashinani o‘rganishga asoslangan yondashuv keng yoritilib, unda model katta hajmdagi annotatsiyalangan matnlar asosida o‘qitilishi va sentimentni avtomatik ravishda aniqlashi tushuntiriladi. Ushbu usulda NV va SVM kabi algoritmlar qo‘llanilib, sentimentni bashorat qilish uchun matn xususiyatlari o‘rganiladi. Maqolada har ikki yondashuvning natijalari taqqoslanib, ularning kuchli va zaif tomonlari muhokama qilinadi. Tadqiqot natijalariga ko‘ra, agar matnni sentiment tahlil qilish uchun yetarli hajmda ma’lumot mavjud bo‘lsa, mashina o‘rganishga asoslangan yondashuv aniqroq natijalar bera oladi. Biroq, leksikonga asoslangan yondashuv ko‘p resurs talab qilmaydi va turli tillarga oson moslashtirilishi mumkin. Shu sababli, maqolada ushbu ikkala yondashuvni birlashtirgan kombinatsiyalangan yondashuv taklif qilinib, uning sentiment tahlilidagi samaradorligi ko‘rsatilgan.

Annotation. This article analyzes existing approaches to sentiment analysis, particularly lexicon-based and machine learning-based methods, and examines their effectiveness. Sentiment analysis is one of the key areas of natural language processing, used to determine the emotional tone of texts (positive, negative, or neutral). The article explains that the lexicon-based approach evaluates words based on predefined dictionaries and sentiment resources (such as WordNet). In this method, the emotional value of words is predetermined and used to determine the overall sentiment of the text. The study also extensively discusses a machine learning-based approach that involves training a model on large volumes of

annotated text to automatically detect sentiment. In this method, algorithms such as Naive Bayes (NB) and Support Vector Machines (SVM) are used to analyze text features for sentiment prediction. The article compares the results of both approaches and discusses their strengths and weaknesses. According to the research findings, if a sufficient amount of data is available for sentiment analysis, the machine learning-based approach can provide more accurate results. However, the lexicon-based approach requires fewer resources and can be easily adapted to different languages. Therefore, the article proposes a combined approach that integrates both methods and demonstrates its effectiveness in sentiment analysis.

Абстрактный. В статье анализируются существующие подходы к анализу настроений, в частности методы, основанные на лексике и машинном обучении, а также изучается их эффективность. Анализ тональности текста является одной из важных областей обработки естественного языка и используется для определения эмоционального настроя (положительного, отрицательного или нейтрального) текстов. В статье объясняется, что подход на основе лексики оценивает слова на основе предопределенных словарей и ресурсов тональности (например, WordNet). В этом методе эмоциональная ценность слов предопределена и используется для определения общего настроения текста. В исследовании также подробно рассматривается подход к машинному обучению, объясняющий, как модель может обучаться на больших объемах аннотированного текста и автоматически определять тональность. Этот метод использует такие алгоритмы, как NV и SVM, для изучения особенностей текста и прогнозирования настроений. В статье сравниваются результаты обоих подходов и обсуждаются их сильные и слабые стороны. Согласно исследованию, если имеется достаточный объем данных для анализа настроений текста, подход на основе машинного обучения может обеспечить более точные результаты. Однако подход, основанный на лексике, не требует больших ресурсов и может быть легко адаптирован к разным языкам. Поэтому в статье предлагается комбинированный подход, объединяющий эти два подхода и демонстрирующий его эффективность в анализе настроений.

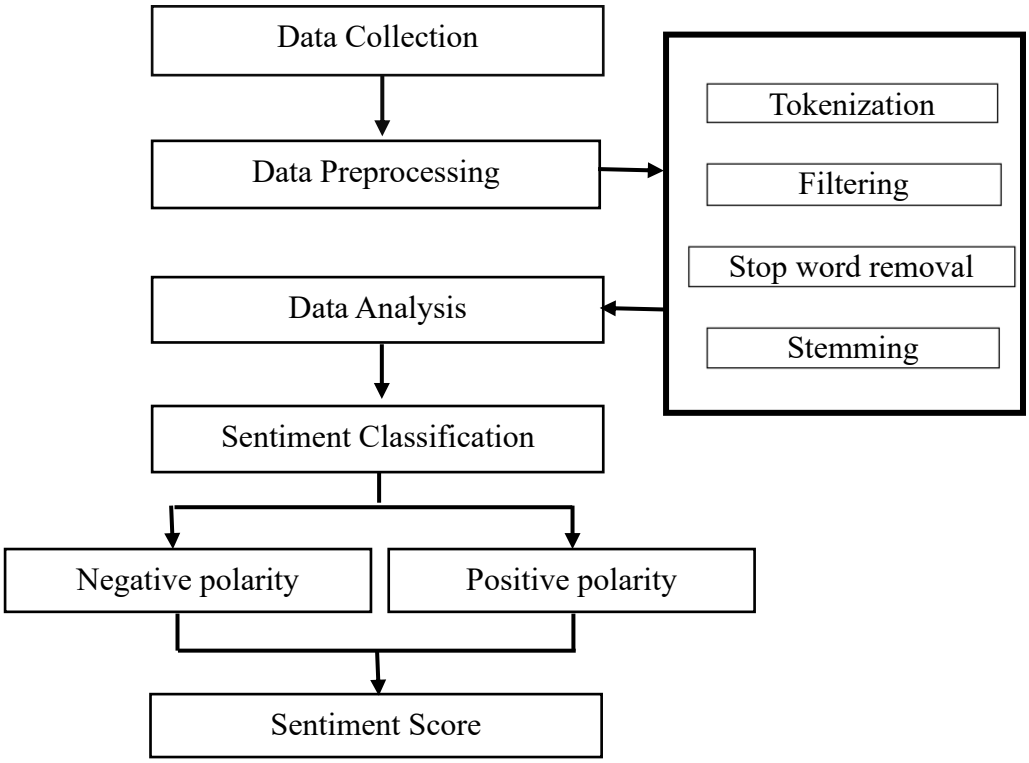
Kalit so‘zlar: *sentiment tahlili, leksikonga asoslangan yondashuv, mashina o‘rganishga asoslangan yondashuv, Naïve Bayes, Support Vector Machine (SVM), sentiment klassifikatsiyasi.*

Kirish

Sentiment tahlili – bu so‘nggi yillarda ko‘plab sohalarda qo‘llanilayotgan muhim texnologiyalardan biri bo‘lib, insonlarning fikrlarini, hissiyotlarini yoki baholarini avtomatik tarzda tahlil qilish imkonini beradi. Ijtimoiy tarmoqlar iste’molchilar va tashkilotlar o‘rtasidagi aloqa vositasi bo‘lib, hozirda keng qo‘llaniluvchi Twitter, Facebook va boshqa tarmoqlar foydalanuvchilarga turli mavzular, shaxslar,

mahsulotlar va xizmatlar haqidagi fikrlarini, xohish-istaklarini ifodalash imkonini bermoqda. Sentiment tahlil esa insonlarning ijtimoiy tarmoqlarda bildirgan ijobiy, salbiy yoki neytral munosabatlarini aniqlash uchun ishlatiladi. Sentiment tahlilni bugungi kunda barcha sohalarda, jumladan, ijtimoiy tarmoqlarda – foydalanuvchilarning fikrlarini tahlil qilishda, marketing sohasida – mahsulot yoki brendga nisbatan mijozlarning munosabatini o‘rganishda, siyosatda – jamoatchilik fikrini o‘rganishda, servis, xizmat, sog‘liqni saqlash sohalarida esa mijozlarning xizmatlar haqidagi fikrlarini tahlil qilishda ishlatamiz. Sentiment tahlil turli ko‘rinishda ifodalanishi mumkin. Masalan, ikkilik shaklida (ijobiy va salbiy), uchlik ko‘rinishida (ijobiy, salbiy, neytral) yoki baho qiymatida (-5 dan 5 gacha). Ushbu ma’lumotlar turli maqsadlar uchun, jumladan, bozor tadqiqotlari, brendni boshqarish va qaror qabul qilish uchun foydali bo‘lishi mumkin.

Sentiment tahlilning jarayon oqimi quyidagicha[1]:



1-rasm. Sentiment tahlil ko‘rinishi

Bu jarayon quyidagicha kechadi:

1. **Ma’lumotlarni yig‘ish** – turli ijtimoiy tarmoqlardan to‘plangan matnli ma’lumotlar to‘planadi;
2. **Ma’lumotlarni oldindan qayta ishlash** – bu bosqichda matnni tayyorlash ishlari amalga oshiriladi:
   - Tokenizatsiya – matn so‘zlarga ajratiladi;

- Filtrlash – keraksiz elementlar chiqarib tashlanadi;
- Stop-so‘zlarni olib tashlash – muhim bo‘lmagan so‘zlar chiqarib tashlanadi;
- Stemming – so‘zlar asosiy asos qismga keltiriladi.

3. **Ma’lumotlarni tahlil qilish** – tozalangan matn ma’lumotlari sentiment tahlili uchun tayyorlanadi;
4. **Sentiment klassifikatsiyasi** – matn ijobiy yoki salbiy polaritetga ega ekanligi aniqlanadi:
   - Salbiy polaritet;
   - Ijobiy polaritet.
5. **Sentiment bahosi** – bu matnning ijobiy yoki salbiy ekanligi haqidagi yakuniy baho.

Ijtimoiy tarmoqlardagi sentiment matnlari 3 ta asosiy elementdan iborat bo‘ladi: opinion holder – fikr bildiruvchi, object – obyekt ya’ni fikr bildirilayotgan narsa yoki shaxs, opinion – matn shaklida ifodalangan munosabat [2]. Ushbu elementlarning har biri matnning aniq tahlil qilinishini ta’minlaydi. Shunday qilib, sentiment tahlili nafaqat individual fikrlarni baholash, balki ijtimoiy jarayonlarning umumiy yo‘nalishini tushunish uchun ham kuchli vositadir.

Sentiment tahlilni 3 usul vositasida amalga oshirish mumkin:

1. **Lexicon-based** (leksikaga asoslangan sentiment tahlili);
2. **Machine learning-based** (mashina o‘rganishga asoslangan sentiment tahlili);
3. **Deep learning** (gibrid yondashuv). Ushbu tadqiqot leksikaga asoslangan va mashina o‘rganishga asoslangan yondashuvlarning kuchli va zaif tomonlarini taqqoslaydi.

Adabiyotlar sharhi

N. Imanina Zabha va boshqalar (2019)[3] Twitter ma’lumotlari asosida leksikonga asoslangan usullar yordamida ko‘p tilli hissiy tahlilni o‘rgangan. Natijada esa klassifikator hissiyotlarni to‘g‘ri aniqlashga muvaffaq bo‘lgan. A. Mitra va boshqalar (2020)[4] esa filmlar bo‘yicha ma’lumotlar to‘plamida leksikonga asoslangan yondashuvdan foydalangan holda qaysi filmlar ommaga ko‘proq yoqishini aniqlashga harakat qilishgan. Bu tadqiqot kino prodyuserlari va teatr egalari uchun foydali hisoblanadi.

M.Sani va boshqalar (2022)[5] Hausa tilidagi tvitlarning hissiy holatini mashinani o‘qitish usullaridan foydalanib tahlil qilishgan. Ushbu tadqiqotda machine learning-based va lexicon-based Hausa matnlarini tasniflashni qanday yaxshilashi mumkinligi o‘rganilgan.

M.S.Hajrahimova, M.I.Ismaylova va boshqalar (2021)[6] COVID-19 pandemiyasi davrida Twitter foydalanuvchilarining munosabatlarini machine learning yordamida tahlil qilishgan. Bu tadqiqotda pandemiya bilan bog‘liq tvitlarning hissiyotlari Supporty Vector Machine, Naive Bayes, Random Forest va Neural Network neyron tarmoqlari yordamida baholangan. Bunda Random Forest klassifikatori 79.37% aniqlik bilan eng yaxshi natijani ko‘rsatgan.

N.M.Sham, A.Mohamed va boshqalar (2022)[7] iqlim o‘zgarishlari bilan bog‘liq tvitlarning hissiy tahlilini o‘rganishgan. Ular yettita leksikonga asoslangan usul, uchta mashina o‘qitish usulini va gibrid algoritmlarni solishtirganlar. Eng samarali usul sifatida 75.3 % natija bilan gibrid yondashuv F1-Score bo‘yicha aniqlangan. Lemmatizatsiya mashina o‘qitish va gibrid usullarni 1.6% ga oshirgan. Shuningdek, T.Islam, Md.Alif Sheakh va boshqalar (2024)[8] turli tadqiqotlarda ishlatilgan yondashuvlarni taqqoslab, ulardan eng maqbuli sifatida LSTM modelining 98% aniqlik darajasiga ega ekanligini isbotladi.

2-jadval. Qisqa matnlar bilan bog‘liq mavjud ishlarga taqqoslama jadvali

Manba	Hissa	Ma’lumotlar to‘plami	Algoritmlar	Eng yaxshi aniqlik
Ushbu ish	Twitter sentiment analizida leksikon va chuqur o‘rganish yondashuvlarini solishtirish	1,6 million tvitdan iborat **Sentiment140** to‘plami	Leksikon asosidagi yondashuv va **Long Short-Term Memory (LSTM)**	**98% (LSTM)**
[11]	Hausa tilidagi Twitter sentiment analizini mashinani o‘rganish va leksikon usullari yordamida o‘rganish	**BBC Hausa Twitter API**	**Logistic Regression**, **Multinomial Naive Bayes (MNB)** + Count Vectorizer va TF-IDF	**86% (LR)**
[12]	Haqiqiy vaqt rejimidagi sentiment analizini SVM, K-SVM va MNB yordamida amalga oshirish	**Twitter ma’lumotlari** va bahsli hududlar yangilik saytlari	**Support Vector Machine (SVM)**, **Kernel SVM (K-SVM)**, **Multinomial Naive Bayes**	**80% (K-SVM)**
[13]	Sentiment analiz uchun leksikon va mashinani o‘rganish usullarini solishtirish	**TF-IDF**	**Logistic Regression**, **Support Vector Machine (SVM)**, **AFINN**, **Vader lexicon**	**96.3% (SVM)**
[15]	Twitter foydalanuvchilarining COVID-19 bo‘yicha munosabatlarini mashinani o‘rganish yordamida tahlil qilish	**Kaggle'dagi COVID-19 tvitlari**	**Naive Bayes**, **Random Forest**, **SVM**, **Neural Network**	**79.37% (Random Forest)**
[17]	Hindistondagi COVID-19 va vaksina bo‘yicha tvitlarni chuqur o‘rganish va leksikon yordamida tahlil qilish	**COVID-19 va vaksina tvitlari (Hindiston)**	**Recurrent Neural Network (RNN)** + **Bi-LSTM va GRU texnikalari**	**93.03% (GRU)**

Chuqur o‘rganish asosidagi usullar, ayniqsa, LSTM-ga asoslangan modellar, matn ketma-ketliklaridagi kontekstual ma’lumotlarni va bog‘liqliklarni olishda sezilarli salohiyatni ko‘rsatdi[9].

Lexicon-based yondashuvi

Leksikon asosidagi sentiment tahlili usuli, matndagi so‘zlarni va ifodalangan his-tuyg‘ularni tahlil qilish uchun oldindan tuzilgan lug‘atlar yoki leksikonlardan foydalanadi. Bu leksikonlarda har bir so‘zga tegishli bo‘lgan ijobiy yoki salbiy his-tuyg‘u (polaritet) ko‘rsatilgan. Bu usulda, matndagi so‘zlar yoki iboralar tekshiriladi va ular leksikonda mavjud bo‘lgan his-tuyg‘u polariteti bilan solishtiriladi[10]. Leksikonga asoslangan yondashuv hissiy tahlil uchun oddiy va tushunarli mexanizmni taqdim etadi. Avvaldan belgilangan leksikondagi so‘zlarni hissiy polaritet bilan bog‘lash orqali, u matnning hissiyotini ijobiy yoki salbiy iboralar mavjudligiga asoslanib tezda tasniflaydi [11]. Biroq, leksikonlar kontekstga bog‘liq kayfiyat, ironiyani yoki jargonni ushlay olmaydi, bu esa xato natijalarga olib kelishi mumkin.

Leksikonga asoslangan yondashuvda hissiyotni bildiruvchi so‘z yoki iboralar oldindan lug‘at holida shakllantirilgan bo‘ladi. Ushbu lug‘at orqali sentiment tahlilning ijobiy yoki salbiy ekanligi ishonchli tahlil qilinadi. Matnning hissiy yo‘nalishini aniqlash uchun leksikon hujjat bilan taqqoslanadi va undagi ijobiy va salbiy so‘zlarning soni aniqlanadi. Keyin esa quyidagi formula yordamida matnning hissiy yo‘nalishi belgilanadi [12]:

$$x1\ (s) = 1 \text{ if sum } (1,s) > 0 \qquad (1)$$

$$x1\ (s) = 0 \text{ if sum } (1,s) = 0 \qquad (2)$$

$$x1\ (s) = -1 \text{ if sum } (1,s) < 0 \qquad (3)$$

Bu yerda sum (1,s) – hujjatda topilgan ijobiy va salbiy hissiyotga ega so‘zlarning umumiy farqini ifodalaydi. Agar ijobiy so‘zlar soni salbiy so‘zlardan ko‘p bo‘lsa natija x1 (s) = 1 (matn ijobiy), agar ijobiy va salbiy so‘zlar teng bo‘lsa x1 (s) = 0 (matn neytral), agar salbiy so‘zlar soni ijobiylardan ko‘p bo‘lsa x1 (s) = -1(matn salbiy) deb qaraladi. Leksikonlarni kengaytirish uchun sinonimlar, WordNet, SenticNet, emojilardan foydalanish mumkin. Sentiment tahlilni amalga oshiruvchi leksikon lug‘atlar 3 usul orqali yaratiladi[13,14]:

a) *qo‘lda yondashuv:* bunda hissiyotni ifodalovchi so‘zlar shaxsiy bilim va va til tushunchasi orqali qo‘lda tuziladi. Bu juda ko‘p vaqt talab qiladi va insonning til bo‘yicha bilimiga bog‘liq bo‘ladi. Odatda, ushbu yondashuv

avtomatlashtirilgan usullar bilan birlashtiriladi, chunki bular ba’zi xatolarga yo‘l qo‘yishi mumkin.

b) *lug‘at asosidagi yondashuv:* hissiyotni bildiruvchi so‘zlar turli lug‘aviy manbalardan yig‘iladi, onlayn lug‘at va leksikografik resurslardan foydalanadi, sinonim, antonimlar orqali yangi hissiy so‘zlar aniqlanadi[15].

c) *Korpus asosidagi yondashuv:* bunda so‘zlarning gap ichidagi o‘rni so‘zlar o‘rtasidagi munosabati va birga ishlatilish chastotasini tahlil qilish orqali hissiy so‘zlarni aniqlaydi[16].

Lexicon-based usulning afzalliklari

- Oddiy: leksikonga asoslangan algoritmlar murakkab machine learning-based usuliga qaraganda kam resurs talab qiladi.
- Tezkor: hissiy yo‘nalishni ifodalovchi so‘z yoki iboralar oldindan mavjudligi sababli natijalar juda tez chiqadi.
- Yaxshi natija beradi: agar lug‘at yetarlicha boy bo‘lsa, sentimentni aniqlash natijalari ancha aniq bo‘ladi.
- Turli usullar bilan kengaytirish mumkin: lug‘at WordNet, SenticNet, emoji va sinonimlar bilan kengaytirilishi mumkin.

Lexicon-based usulining kamchiliklari

- Ironiya va sarkazmni tushunmaydi: masalan, *“Xizmat ko‘rsatishga gap bo‘lishi mumkinmas, 4 soat kutdim!”* kabi iboralarni noto‘g‘ri tushunishi mumkin.
- Til bilan bog‘liq muammo: har bir til uchun alohida sentiment lug‘ati bo‘lishi shart.
- Kontekstni tushunmaydi: masalan, *“Havo juda sovuq, bu yomon”,* degan gapning hissiyotini aniq tushunmaydi.

Leksikon asosidagi yondashuv juda oddiy, samarali va tushunarli hisoblanadi. Bunda mashina o‘rganish talab qilinmaydi. Lekin bu usulda so‘zning kontekstda anglatayotgan ma’nosi turlicha bo‘lishi hisobga olinmaydi. Masalan, *yaxshi* va *yaxshi emas* so‘zini ijobiy deb tahlil qilishi mumkin. Shuningdek, biror mahsulotga baho berishda qo‘shtirnoq ichida teskari ma’no anglatuvchi birikma yoki gaplar ham noto‘g‘ri tahlil qilinishi mumkin: *“Gap bo‘lishi mumkin emas”* – bu, umuman, yaxshi yoki juda yomon ma’nolarida ham ishlatiladi.

Machine learning-based yondashuvi

Bu yondashuv mashina o‘rganishga asoslangan usul bo‘lib, sentiment tahlilda katta hajmdagi ijtimoiy tarmoq postlari, mahsulot sharhlari va xabarlar

ustida ishlashda samarali hisoblanadi. Bu usul kompyuter tizimlariga aniq dasturlashtirishsiz o‘z tajribasi aosida o‘rganish va qaror qabul qilish imkonini beradi. Ushbu yondashuv ikki asosiy turga bo‘linadi[17]:

1. **Nazoratli o‘rganish** – oldindan belgilangan toifalarga ega bo‘lgan matn ma’lumotlari asosida modelni o‘rgatish usulidir. Uning afzalligi yuqori aniqlik bilan yaxshi ishlash imkoniyatiga ega. Taqdim etilgan sohaga mos ravishda optimallashtirish mumkin. Kamchiligi esa ko‘p miqdordagi belgilangan o‘quv ma’lumotlari talab qiladi, yangi yoki kontekstga bog‘liq so‘zlarni tushunmasligi mumkin.
2. **Nazoratsiz o‘rganish** – avvaldan belgilangan ma’lumotlarga tayanmagan holda sentimentni aniqlaydi. Afzalligi annotatsiyalangan ma’lumotlarni talab qilmaydi. Yangi yoki noma’lum mustaqil so‘zlarning hissiy belgisini mustaqil o‘rganishi mumkin. Kamchiligi esa natijalarni oldindan bashorat qilganligi sababli aniqlik kamroq bo‘lishi mumkin.

Shuningdek, ba’zi adabiyotlarda *mustahkamlovchi o‘rganish* turi haqida ham so‘z boradi. Mustahkamlovchi o‘rganishda mukofot va jazolar asosida aniq bir qaror qabul qilishni o‘rganadi. U tahlil natijasida mukofot yoki jazo olib, o‘z strategiyasini yaxshilaydi.

ML yondashuvining afzalliklari:

- Yuqori aniqlikka ega (katta hajmdagi ma’lumotlarga moslashib mustaqil qaror qabul qiladi);
- So‘zlarning kontekstda anglatayotgan ma’nosini tushuna olish qobiliyatiga ega;
- Mashina o‘rgangan sari chiqargan natijalarini o‘zgartirib, sifatini oshirib boradi.

ML yondashuvining kamchiliklari:

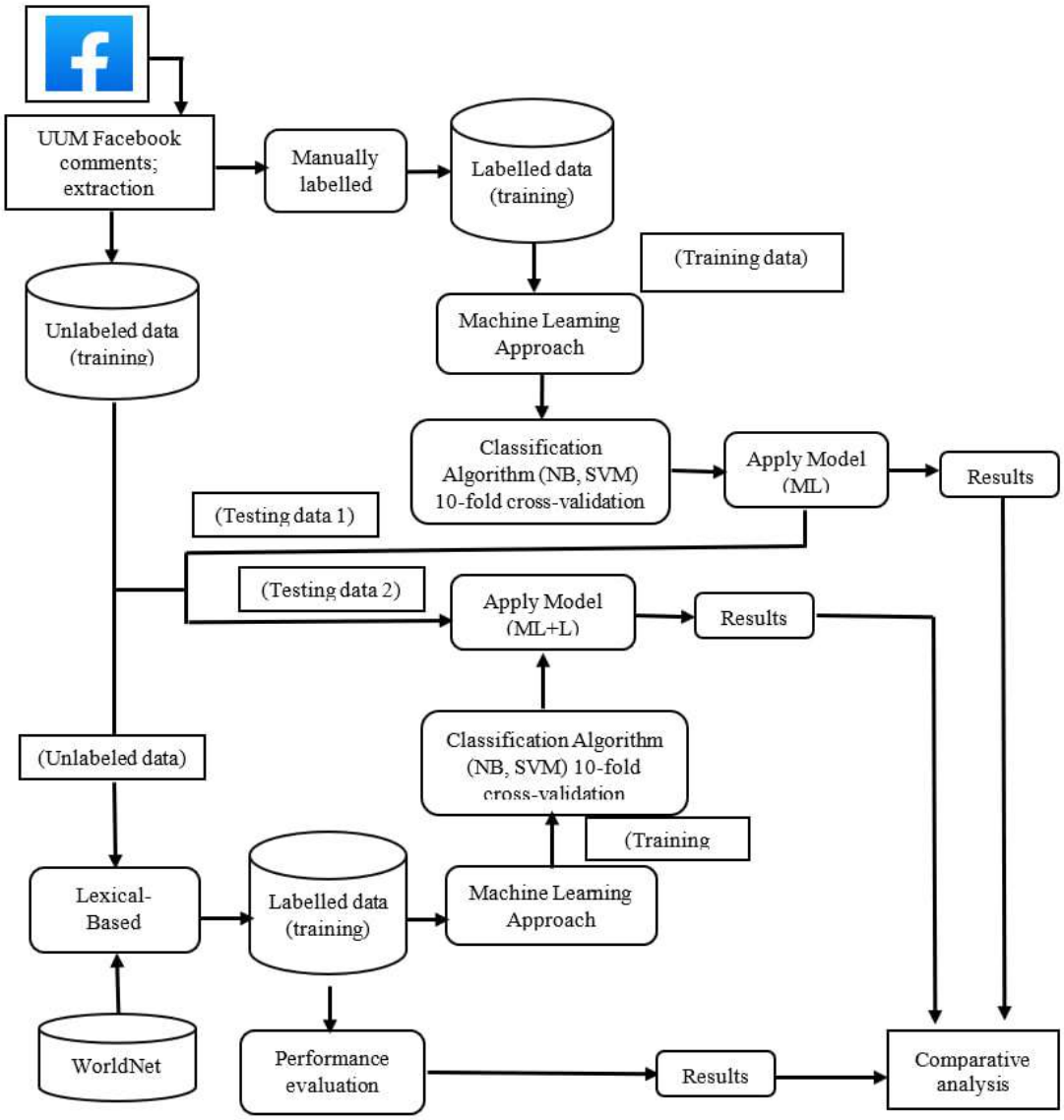
- Katta hajmdagi yorliqlangan ma’lumotlar talab qiladi;
- Tahlil qilish jarayoni ko‘p vaqt talab qilishi mumkin;
- Ma’lumotlar tili yoki hissiy bahosi o‘zgarganda model moslashishi qiyin bo‘lishi mumkin.

Mashina o‘rganish usuli bugungi kunda sun’iy intellekt, tibbiyot, moliya, transport, marketing kabi ko‘plab sohalarda qo‘llanilmoqda.

Chuqur o‘rganishga asoslangan yondashuv (gibrid)

Islam T., Sheakh Md.A., Sadik Md.R., Tahosin Mst.S., Foysal Md.M.R., Ferdush J. va M. Begumlar leksikonga asoslangan va chuqur o‘rganishga asoslangan

yondashuvlarni taqqoslashgan (2024)[8]. Ushbu tadqiqotda ikki mashhur yondashuv – leksikon asosidagi va chuqur o‘rganish asosidagi yondashuvlarni 1,5 milliondan ortiq ma’lumotga ega **sentiment140** deb nomlangan Twitter datasetidan foydalanib taqqoslashgan. Asosiy e’tiborni chuqur o‘rganishning ketma-ketlik modeli – *Long Short-Term Memory (LSTM)* ga qaratadilar. Ushbu tadqiqot natijasida leksikon asosidagi yondashuv 95% aniqlikka, chuqur o‘rganish algoritmlari model 98% aniqlikka ega ekanligini aniqlashgan. Shuningdek, Alaa Thamer Mahmood, Siti Sakira Kamaruddin va boshqalar (2020)[18] leksikonga asoslangan yondashuv va mashina o‘rganishga asoslangan yondashuvni birlashtirib Facebook izohlarida sentiment tahlilni amalga oshirdilar.



3-rasm. Gibrid yondashuv asosidagi sentiment tahlil

Bunda UUM Facebook sahifasidan izohlar olinadi.

1. Mashina o‘rganish yondashuvi:

- izohlarga ijobiy yoki salbiy yorliqlari qo‘shiladi;
- yorliqlangan ma’lumotlar mashina o‘rganish modeli uchun o‘quv ma’lumoti sifatida ishlatiladi;
- NB va SVM algoritmlari yordamida 10 marta o‘qitiladi;
- 1-natija tayyor bo‘ladi.

2. Leksikonga asoslangan yondashuv:

- Facebook izohlari WordNet lug‘atidan foydalanib baholanadi;
- lug‘at asosida har bir so‘z uchun sentiment aniqlanadi;
- bu ma’lumotlar yorliqlangan shaklda saqlanadi va mashina o‘rganish model uchun ma’lumot sifatida ishlatiladi;
- mashina o‘rganish yondashuvi qayta ishlanib 10 marta qayta tekshirilib o‘qitiladi;
- ushbu yondashuv bo‘yicha 2-natija hosil bo‘ladi.

3. Gibrid yondashuv:

- leksikonga asoslangan yondashuvdan olingan natijalar mashina o‘rganish modeli bilan birlashtiriladi;
- birlashtirilgach 3-natija hosil bo‘ladi.

4. Natijalarni taqqoslash:

- 1-natija, 2-natija va 3-natija o‘zaro taqqoslanib, qaysi yondashuv eng aniq sentiment tahlilini amalga oshirgani baholanadi [19].

Xulosa

Bugungi kunda ijtimoiy tarmoqlarda insonlar tomonidan bildirilayotgan fikrlar va sharhlar katta hajmli bo‘lib, bu ma’lumotlarni inson omili vositasida hissiy tahlil qilish juda ko‘p vaqt va resurs talab qiladi. Ushbu ma’lumotlarning sentimentini aniqlash uchun sun’iy intellekt va avtomatlashtirilgan tahlil usullari tobora keng qo‘llanilmoqda. Ayniqsa, o‘zbek tilida ham sentiment tahlili bo‘yicha katta ishlar amalga oshirilmoqda. S.Allanazarova, D.Elova (2024)[20] sentiment tahlil uchun matndagi valentlik, hissiyot va boshqa ta’sirchan holatlarni aniqlaganlar. Ilyos Rabbimov (2021)[21] boshchiligidagi bir guruh izlanuvchilar o‘zbek filmlariga qoldirilgan sharhlarni emojilar asosida o‘rganishgan. Ushbu tadqiqotda esa sentiment tahlilni amalga oshirish uchun eng mashhur ikki yondashuv tahlil qilindi. Har ikkala usul afzallik va kamchiliklarga ega. Faqat bitta usulga tayanish sentiment tahlil uchun yetarli emas, chunki har bir yondashuvning cheklovlari mavjud. Leksikonga asoslangan yondashuvni mashina o‘rganish modeli bilan birlashtirish sentimentlarni yanada aniqroq tahlil qilish imkonini beradi.

Foydalanilgan adabiyotlar:

1. Verma, B., & Thakur, R. S. (2018). Sentiment analysis using lexicon and machine learning-based approaches: A survey. In *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017* (pp. 441-447). Springer Singapore.
2. Verma, B., & Thakur, R. S. (2018). Sentiment analysis using lexicon and machine learning-based approaches: A survey. In *Proceedings of International Conference on Recent Advancement on Computer and Communication: ICRAC 2017* (pp. 441-447). Springer Singapore.
3. Zabha, N. I., Ayop, Z., Anawar, S., Hamid, E., & Abidin, Z. Z. (2019). Developing cross-lingual sentiment analysis of Malay Twitter data using lexicon-based approach. *International Journal of Advanced Computer Science and Applications*, *10*(1).
4. Mitra, A., & Mohanty, S. (2020). Sentiment analysis using machine learning approaches. *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS*, *2*, 63-68.
5. Sani, M., Ahmad, A., & Abdulazeez, H. S. (2022). Sentiment analysis of Hausa language tweet using machine learning approach. *Journal of Research in Applied Mathematics*, *8*(9), 07-16.
6. Hajrahimova, M. S., & Ismaylova, M. I. (2021). Machine learning-based sentiment analysis of Twitter data.
7. Sham, N. M., & Mohamed, A. (2022). Climate change sentiment analysis using lexicon, machine learning and hybrid approaches. *Sustainability*, *14*(8), 4723.
8. Islam, T., Sheakh, M. A., Sadik, M. R., Tahosin, M. S., Foysal, M. M. R., Ferdush, J., & Begum, M. (2024). Lexicon and deep learning-based approaches in sentiment analysis on short texts. *Journal of Computer and Communications*, *12*(1), 11-34.
9. Bhowmik, N. R., Arifuzzaman, M., & Mondal, M. R. H. (2022). Sentiment analysis on Bangla text using extended lexicon dictionary and deep learning algorithms. *Array*, *13*, 100123.
10. Catelli, R., Pelosi, S., & Esposito, M. (2022). Lexicon-based vs. Bert-based sentiment analysis: A comparative study in Italian. *Electronics*, *11*(3), 374.
11. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.
12. Hota, H. S., Dinesh K. Sharma, and Nilesh Verma. "Lexicon-based sentiment analysis using Twitter data: a case of COVID-19 outbreak in India and abroad." *Data science for COVID-19*. Academic Press, 2021. 275-295.

13. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams engineering journal, *5*(4), 1093-1113.

14. Khan, A., Baharudin, B., Lee, L. H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. Journal of advances in information technology, *1*(1), 4-20.

15. Hardeniya, T., & Borikar, D. A. (2016). Dictionary based approach to sentiment analysis-a review. International Journal of Advanced Engineering, Management and Science, *2*(5), 239438.

16. Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert systems with applications*, *36*(3), 6527-6535.

17. Zhang, H., Gan, W., & Jiang, B. (2014, September). Machine learning and lexicon based methods for sentiment classification: A survey. In *2014 11th web information system and application conference* (pp. 262-265). IEEE.

18. Mahmood, A. T., Kamaruddin, S. S., Naser, R. K., & Nadzir, M. M. (2020). A combination of lexicon and machine learning approaches for sentiment analysis on Facebook. *Journal of System and Management Sciences*.

19. Surroop, K., Canoo, K., & Pudaruth, S. (2016). A Novel Position-based Sentiment Classification Algorithm for Facebook Comments. *International Journal of Advanced Computer Science and Applications*, *7*(10).

20. Allanazarova, S. (2024). SENTIMENT TAHLIL UCHUN MATNDAN VALENTLIK, HISSIYOT VA BOSHQA TA’SIRCHAN HOLATLARNI ANIQLASH. *Uzbekistan language and culture*, *2*(02).

21. Allanazarova, S. (2021). Jamoatchilik fikrini organish yoki sentiment tahlil. *Computer linguistics: problems, solutions, prospects*, *1*(1).

22. B.Elov, Sh.Hamroyeva, O.Abdullayeva, M.Uzoqova. O‘zbek tilida pos tegging masalasi: muammo va takliflar. O`zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 45-63 b.

SENTIMENT TAHLILI UCHUN O‘RGATILGAN RESURLARI KAM TILLARDA MA’LUMOTLAR OMBORINI TUZISH

Allanazarova Saboxat Yusupboyevna,
Tayanch doktorant
allanazarova.sabosh@gmail.com
ToshDO‘TAU

Annotatsiya. Kam resursli tillar uchun sentiment tahlili tizimlarini rivojlantirish tabiiy tilni qayta ishlash (NLP) sohasidagi dolzarb muammolardan biridir. Ushbu maqolada kam resursli tillarda, xususan, o‘zbek tilida sentiment tahlili uchun lingvistik ta’minot yaratish jarayoni tahlil qilinadi. Leksik-semantik lug‘atlar, belgilangan korpuslar va annotatsiya usullarini ishlab chiqish, so‘z va iboralarning polaritesini aniqlash hamda ularni avtomatlashtirilgan tahlil tizimlariga integratsiya qilish usullari ko‘rib chiqiladi. Tadqiqotda o‘zbek tilining morfologik va sintaktik o‘ziga xosliklari hisobga olinib, mavjud xalqaro sentiment resurslari bilan taqqoslash amalga oshiriladi. Shuningdek, kam resursli tillar uchun lingvistik ta’minotni yaratishning asosiy muammolari va ularni bartaraf etish strategiyalari muhokama qilinadi. Ushbu ish natijalari o‘zbek tili uchun sifatli sentiment tahlili tizimlarini yaratishga hissa qo‘shishi bilan birga, kam resursli tillar uchun NLP infratuzilmasini rivojlantirishga ham xizmat qiladi.

Annotation. The development of sentiment analysis systems for low-resource languages is one of the critical challenges in the field of Natural Language Processing (NLP). This paper explores the process of creating linguistic resources for sentiment analysis in low-resource languages, specifically in Uzbek. It examines the development of lexicosemantic dictionaries, annotated corpora, and annotation methods, as well as automated techniques for determining the polarity of words and phrases, followed by their integration into sentiment analysis systems. The research takes into account the morphological and syntactic characteristics of the Uzbek language and compares it with existing international sentiment resources. Furthermore, the main challenges in creating linguistic resources for low-resource languages are discussed, along with strategies to overcome them. The findings of this study contribute to the development of high-quality sentiment analysis systems for the Uzbek language and support the advancement of NLP infrastructure for low-resource languages.

Аннотация. Развитие систем анализа сентимента для языков с ограниченными ресурсами является одной из актуальных проблем в области обработки естественного языка (NLP). В данной статье рассматривается процесс создания лингвистического обеспечения для анализа сентимента в языках с ограниченными ресурсами, в частности в узбекском языке. Исследуются методы разработки лексико-семантических словарей,

размеченных корпусов и аннотационных подходов, а также автоматизированные методы определения полярности слов и фраз с их последующей интеграцией в системы анализа сентимента. В работе учитываются морфологические и синтаксические особенности узбекского языка, проводится сравнительный анализ с международными сентиментными ресурсами. Кроме того, обсуждаются основные проблемы, возникающие при создании лингвистического обеспечения для языков с ограниченными ресурсами, а также предлагаются стратегии их преодоления. Полученные результаты способствуют созданию качественных систем анализа сентимента для узбекского языка, а также развитию NLP-инфраструктуры для языков с ограниченными ресурсами.

Kalit so‘zlar: *sentiment tahlili, kam resursli tillar, tabiiy tilni qayta ishlash, lingvistik resurslar, belgilangan korpus, leksik-semantik lug‘at, hissiy toifa, annotatsiya, NLP infratuzilmasi, avtomatlashtirilgan tahlil.*

1. KIRISH

So‘nggi yillarda tabiiy tilni qayta ishlash (NLP) va sun’iy intellekt texnologiyalarining rivojlanishi matnlarning avtomatlashtirilgan tahlilini keng qo‘llash imkonini bermoqda. Ushbu yo‘nalishlardan biri bo‘lgan sentiment tahlili matnlardagi ijobiy, salbiy yoki neytral kayfiyatni aniqlashga qaratilgan bo‘lib, ijtimoiy tarmoqlar, yangiliklar, sharhlar va boshqa matn manbalarini tahlil qilishda muhim ahamiyat kasb etadi. Biroq, sentiment tahlili tizimlarini yaratishda lingvistik resurslarning mavjudligi hal qiluvchi omillardan biri hisoblanadi. Ingliz, xitoy va boshqa ko‘p resursli tillar uchun keng ko‘lamli leksik-semantik lug‘atlar, belgilangan korpuslar va annotatsiya qilingan ma’lumotlar bazasi mavjud bo‘lsa-da, kam resursli tillarda, jumladan, o‘zbek tilida bunday infratuzilma yetarlicha rivojlanmagan.

Kam resursli tillar – tabiiy tilni qayta ishlash (NLP) va sun’iy intellekt tizimlari uchun yetarli darajada lingvistik resurslarga ega bo‘lmagan tillardir [1]. Ushbu tillarda belgilanadigan korpuslar, leksik-semantik lug‘atlar, annotatsiyalangan ma’lumotlar bazasi va yetarli miqdordagi o‘quv materiallari mavjud emasligi sababli, mashinani o‘rganish modellari samarali ishlashi qiyinlashadi. Kam resursli tillarga odatda rasmiy tillar sifatida kamroq qo‘llaniladigan yoki iqtisodiy va ilmiy sohalarda yetarli darajada raqamlashtirilmagan tillar kiradi.

Kam resursli tillarning xususiyatlari quyidagilarga qarab aniqlanadi:

Matn resurslarining cheklanganligi – onlayn va bosma manbalarda bunday tillarda ma’lumotlarning yetishmovchiligi.

Belgilanmagan korpuslar muammosi – annotatsiyalangan va belgilangan NLP ma’lumotlar bazasining yo‘qligi yoki juda kichik hajmda bo‘lishi.

Leksik va morfologik murakkablik – ayrim kam resursli tillar murakkab morfologiyaga ega bo‘lib, NLP modellarining samarali ishlashiga ta’sir ko‘rsatadi.

Ko‘p tilli NLP modellariga integratsiya qiyinligi – ko‘p hollarda bu tillar uchun maxsus NLP modellarining yo‘qligi sababli boshqa tillardan transfer o‘rganish usullari talab etiladi.

Masalan, o‘zbek tili NLP jamoatchiligi tomonidan kam resursli til sifatida baholanadi, chunki sentiment tahlili, mashinada tarjima va nutqni avtomatik tanib olish uchun katta hajmli va sifatli korpuslar yetarli emas. Ingliz, xitoy yoki nemis tillari bilan solishtirganda, o‘zbek tilida NLP modellarini o‘rgatish va qo‘llashda sezilarli qiyinchiliklar kuzatiladi. Shu sababli, kam resursli tillarda sentiment tahlili uchun lingvistik ta’minot yaratish dolzarb ilmiy muammolardan biri hisoblanadi.

Kam resursli tillarda sentiment tahlilining samaradorligini oshirish uchun maxsus lingvistik resurslarni ishlab chiqish zarur. Bu jarayon o‘z ichiga leksik-semantik lug‘atlar yaratish, belgilangan korpuslar tayyorlash, polarite va hissiyot darajalarini aniqlash kabi muhim bosqichlarni oladi. O‘zbek tilining morfologik va sintaktik jihatdan boy va murakkab tuzilishi sentiment tahlilini yanada qiyinlashtiradi. Shunday ekan, ushbu tadqiqotda o‘zbek tili misolida kam resursli tillar uchun sentiment tahlili tizimlarini rivojlantirishga qaratilgan lingvistik ta’minotni yaratish masalasi yoritiladi.

2. Adabiyotlar tahlili

So‘nggi yigirma yilda sentiment tahlili tabiiy tilni qayta ishlash (NLP) sohasining dolzarb yo‘nalishlaridan biriga aylandi va ushbu sohada turli nazariy hamda amaliy tadqiqotlar olib borildi. Xususan, sentiment tahlili uchun maxsus leksik resurslar va tasniflovchi modellar ko‘plab yirik tillar, jumladan, ingliz, rus va turk tillari uchun ishlab chiqilgan. Biroq, kam resursli tillar, ayniqsa, agglutinativ tuzilishga ega bo‘lgan o‘zbek tili uchun bunday imkoniyatlar cheklangan bo‘lib, mavjud metod va resurslarning bevosita qo‘llanishi samarasiz bo‘lishi mumkin. Ayrim yondashuvlar sentiment lug‘atlarini tarjima qilish orqali foydalanishni taklif qiladi, ammo tillar orasidagi semantik, sintaktik va pragmatik tafovutlar natijasida bunday yondashuv yetarlicha aniqlikni ta’minlay olmaydi. Shu bois, har bir tilning morfologik va sintaktik o‘ziga xosliklarini inobatga olgan holda, maxsus sentiment lug‘atlari hamda tasniflash modellarini ishlab chiqish zarurati ortib bormoqda.

Rus tilida sentiment tahliliga oid tadqiqotlar lingvistik yondashuvlardan tortib, mashinani o‘rganish va chuqur neyron tarmoqlar asosidagi modellargacha bo‘lgan keng doirani qamrab oladi. Ayniqsa, oxirgi yillarda transformer modellar va kontekstga bog‘liq tahlil usullarining rivojlanishi sentimentni yanada aniqroq

aniqlashga imkoniyat yaratmoqda. O‘zbek tili uchun sentiment tahlili borasida esa hali ham yetarli tadqiqotlar mavjud emas, ammo so‘nggi yillarda ayrim ilmiy izlanishlar boshlangan.

D.B. Mengliyev boshchiligidagi tadqiqotchilar guruhi o‘zbek tilidagi sentiment tahlilini amalga oshirish uchun maxsus algoritm ishlab chiqib, har bir so‘z, ibora va frazeologizmlarga -3 dan +3 gacha bo‘lgan sentiment ballarini belgilash mexanizmini taklif etishadi[2]. Sinov natijalari ushbu algoritmning aniqligi 85–95% oralig‘ida ekanligini ko‘rsatgan, biroq tadqiqot davomida so‘z shakllarini yanada mukammal qayta ishlash va lug‘at bazasini kengaytirish zarurati aniqlangan.

E.Y. Akhmedov o‘z tadqiqotida sentiment tahlili jarayonida o‘zbek tilining leksik, grammatik va kontekstual xususiyatlarini hisobga olish lozimligini ta’kidlaydi. Xususan, chuqur o‘rganish modellaridan—LSTM, CNN va mBERT kabi ilg‘or yondashuvlardan foydalanish o‘zbek tilidagi sentiment tahlilining samaradorligini oshirish imkonini berishi ko‘rsatilgan[3].

K. Niyozmetova tomonidan olib borilgan tadqiqotda Toshkent shahridagi restoranlarga oid Google Maps’dagi foydalanuvchilar fikrlari tahlil qilingan. Tadqiqotchilar sentiment tasnifi uchun logistic regression modelidan foydalangan bo‘lib, oldindan qayta ishlash bosqichida stemming kabi agglutinativ tillar uchun muhim bo‘lgan usullarning samaradorligi kuzatilgan. Tadqiqot natijalari sentiment tahlili xizmat sifatini baholashda va iste’molchi sadoqatini oshirishda samarali vosita ekanligini tasdiqlaydi[4].

S. Matlatipov tadqiqotida o‘zbek tilidagi restoran sharhlaridan iborat sentiment datasetini yaratish va tahlil qilish jarayoni yoritilgan. O‘zbek tilining kam resursga ega tillardan biri ekani inobatga olinib, logistic regression, support vector machines (SVM) hamda chuqur o‘rganish modellari (RNN, CNN) sinovdan o‘tkazilgan. Ma’lumotlarni yig‘ish, oldindan qayta ishlash va baholash bosqichlari batafsil tahlil qilingan bo‘lib, stemming kabi metodlarning qo‘llanilishi natijalarning yaxshilanishiga xizmat qilgan. Eng samarali model 91% aniqlikka erishgan[5].

E. Kuriyozov tadqiqotida o‘zbek tilida ko‘p-toifali yangiliklar tasnifi uchun dataset yaratish va modellarni baholash jarayoni o‘rganilgan. Shu maqsadda 10 ta yangilik saytidan ma’lumotlar yig‘ilib, 15 toifani qamrab oluvchi dataset shakllantirilgan. Sinov natijalariga ko‘ra, RNN va CNN qoidaga asoslangan usullarga nisbatan samaraliroq ekanligi kuzatilgan, eng yuqori natijalar esa BERTbek modelida qayd etilgan. Ushbu natijalar o‘zbek tilidagi matn tasniflash bo‘yicha tadqiqotlar uchun muhim asos bo‘lib xizmat qilishi mumkin[6].

I. Rabbimovning tadqiqotida o‘zbek tilida ko‘p-toifali tasniflash amalga oshirilib, “Daryo” yangiliklar saytining 10 toifasiga tegishli dataset yaratilgan. Tadqiqotchilar SVM, DTC, RF, LR va MNB kabi mashinani o‘rganish modellaridan foydalanib, sentiment tasnifini amalga oshirishgan. Belgilar ajratishda TF-IDF va n-

gram metodlari qo‘llanilgan. 5-fold cross-validation orqali hiperparametrlar optimallashtirilgan va natijada 86.88% aniqlikka erishilgan. Ushbu tadqiqot natijalari o‘zbek tilidagi matnlarni avtomatik tasniflash uchun samarali usullarni ko‘rsatadi[7].

A. Yusufu va hamkor tadqiqotchilar guruhi kam resursli til sifatida o‘zbek tilida sentiment tahlilini o‘rganish maqsadida Google Play Store’dan 27,985 ta o‘zbekcha sharh yig‘ib, User Experience Honeycomb modelining olti asosiy komponenti bo‘yicha 43,712 jumlani annotatsiya qilishgan. Sentiment qutblanishini baholash uchun oldindan o‘qitilgan modellar hamda GCN asosida integratsiyalangan yondashuv ishlab chiqilgan. Multi-klassifikatsiya uchun F1 ko‘rsatkichida 0.30, binar tasniflash uchun esa 0.43 yaxshilanish kuzatilgan. Ushbu tadqiqot kam resursli tillarda foydalanuvchi tajribasini chuqur tahlil qilish uchun istiqbolli yondashuvlarni taqdim etadi[8].

Mazkur tadqiqotlar shuni ko‘rsatadiki, o‘zbek tilida sentiment tahlilini rivojlantirish uchun lingvistik resurslarni kengaytirish, maxsus sentiment lug‘atlarini ishlab chiqish hamda chuqur o‘rganish modellarini tatbiq etish muhim ahamiyatga ega. Ushbu yo‘nalishda tadqiqotlarni davom ettirish va NLP texnologiyalarini takomillashtirish kelajakda samarali natijalarga olib kelishi mumkin.

3. Metodologiya

Kam resursli tillar uchun sentiment tahlili tizimini yaratishda lingvistik ta’minotning yetishmovchiligi asosiy muammolardan biri hisoblanadi. Ushbu tadqiqotda o‘zbek tilida sentiment tahlili uchun zarur bo‘lgan lingvistik resurslarni yaratish metodologiyasi ishlab chiqildi. Metodologiya uch asosiy bosqichni o‘z ichiga oladi: (1) belgilangan korpus yaratish va annotatsiyalash, (2) leksik-semantik lug‘at ishlab chiqish, (3) sentiment tahlili modellarini sinovdan o‘tkazish va baholash. Baholash yarim avtomatik tarzda amalga oshirildi.

Belgilangan korpus yaratish va annotatsiyalash

Belgilangan korpus sentiment tahlili modelini o‘rgatish va baholash uchun asosiy manba hisoblanadi. O‘zbek tilida bunday korpuslarning mavjud emasligi sababli, qo‘lda annotatsiyalash va yarim-avtomatlashtirilgan yondashuvlardan foydalanish zarur. Annotatsiya jarayonida quyidagi xalqaro amaliy tajribalar tahlil qilindi:

- **SentiWordNet** (Esuli & Sebastiani, 2006)
- **SenticNet** (Cambria va boshqalar, 2012)

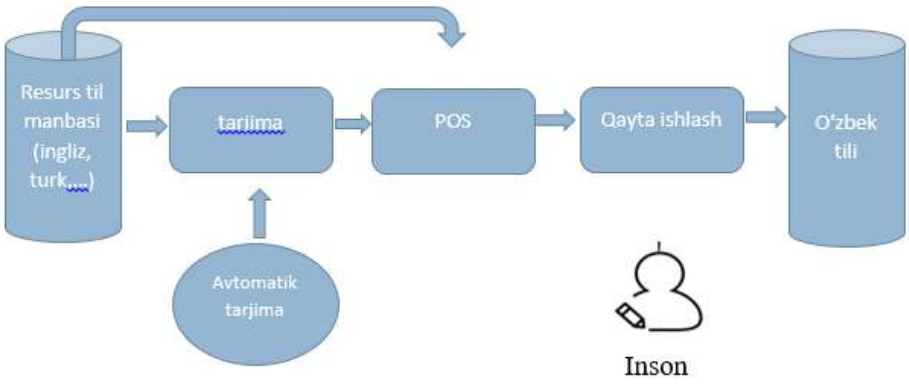
SentiWordNet, SenticNet score lari kiritildi. Bularni hammasidan foydalanib SentiUzNet lingvistik ta’minoti uchun pos, neg, obj score lari mashinaviy o‘qitish algoritmlari yordamida qo‘yildi.

O‘zbek tili uchun belgilangan korpus yaratishda ushbu ishlardan ilhomlanib, **ijtimoiy tarmoqlar (Twitter, Facebook), yangiliklar portallari (Kun.uz, Daryo.uz) va blog postlari** kabi turli manbalardan matnlar yig‘ildi. Tanlangan matnlar qo‘lda annotatsiyalanib, ijobiy, salbiy va betaraf toifalarga ajratildi. 5 nafar annotatorlar o‘rtasidagi kelishuv darajasini baholash uchun Cohen’s Kappa koeffitsienti (Cohen, 1960) ishlatildi. Belgilangan Cohen's kappa toifalar mosligi 0.79 ni ko‘rsatdi.

Leksik-semantik lug‘at ishlab chiqish

Leksik-semantik lug‘atlar sentiment tahlilining muhim komponenti hisoblanadi. Ushbu bosqichda o‘zbek tilida foydalanish mumkin bo‘lgan lug‘at yaratish uchun quyidagi usullardan foydalanildi:

Tarjima asosida lug‘at yaratish: SentiTurkNet lug‘atidagi fizika, kimyo, tibbiyot sohalariga oid 500 dan ortiq terminlar o‘zbek tiliga tarjima qilindi va ularning polaritesi lingvist mutaxassislar tomonidan tekshirildi.

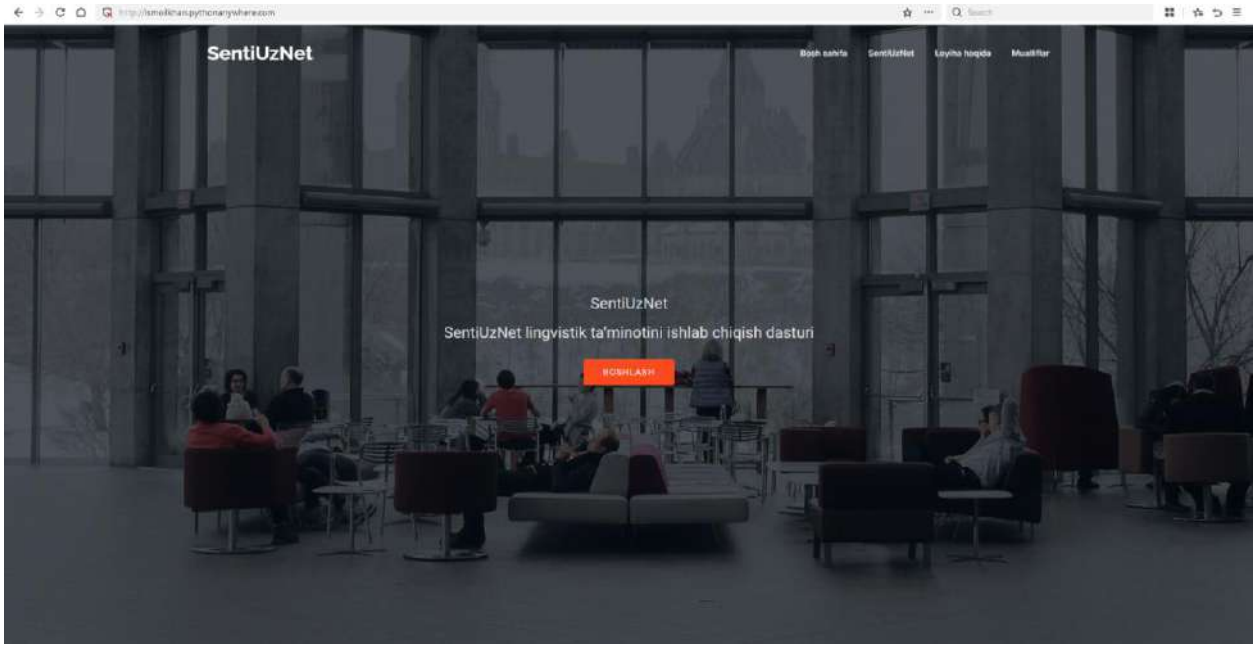


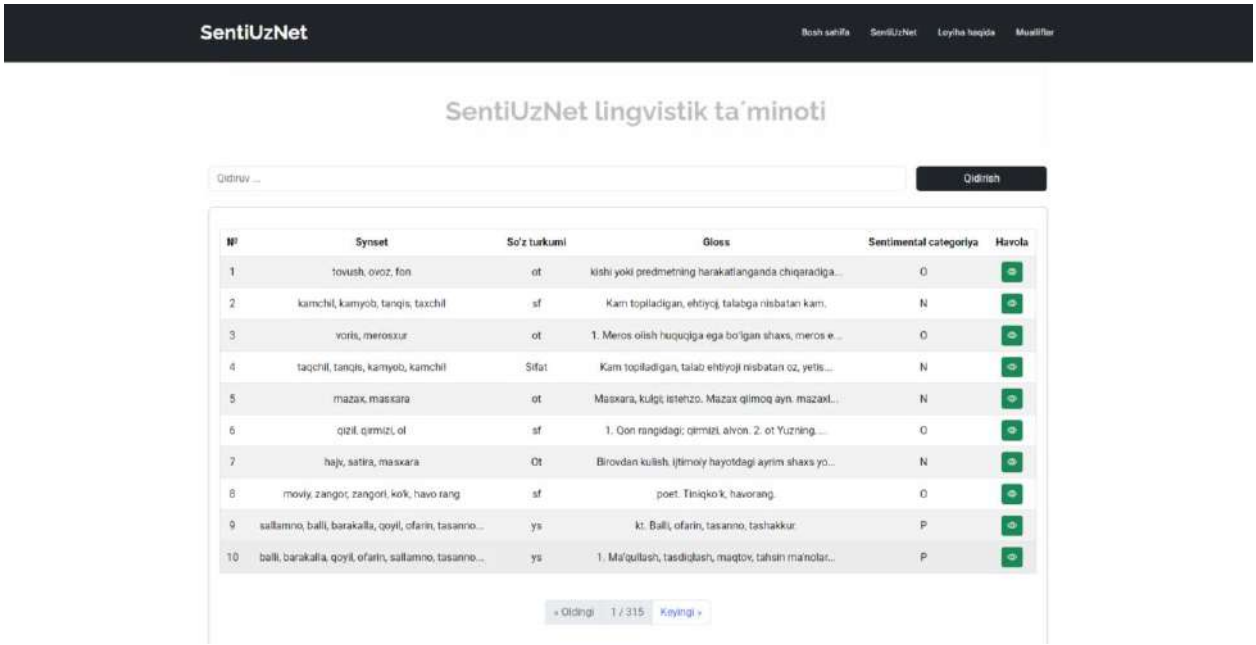
1- rasm. Tarjimaga asoslangan usulning bosqichlari.

O‘zbek tili uchun ishlab chiqilgan sentiment lug‘at ingliz (SentiWordNet), (SenticNet) har bir tillar orasidagi hissiy tasniflash farqlari statistik tahlil qilindi.

NATIJALAR

Har bir so‘z, sinset, gloss uchun PosPMI, NegPMI hisoblab chiqildi. Bunda o‘zbek tilidagi barcha sentimental tahlil datasetlaridan foydalanildi, jami 384 mingta sharh (review) yordamida hisoblandi.





O‘zbek tilida sentiment tahlilini baholash natijalari boshqa tillar uchun erishilgan natijalar bilan taqqoslandi. Masalan, O‘zbek tili uchun ishlab chiqilgan sentiment lug‘at ingliz tilidagi (SentiWordNet), (SenticNet) lug‘atlari orasidagi hissiy tasniflash farqlari statistik tahlil qilindi. Jami 11 ta ML algoritmi ishlatilgan bo’lib, 71-91% aniqlik ko‘rsatgan.

Model	Accuracy	Precision	Recall	F-Measure	Training Time (s)	Test Time (s)
Random Forest	0,9141	0,9137	0,9141	0,9139	1,5148	0,0090
XGBoost	0,9141	0,9137	0,9141	0,9138	0,9549	0,0060
LightGBM	0,9094	0,9089	0,9094	0,9091	0,8835	0,0040
Stacking Classifier	0,9094	0,9088	0,9094	0,9090	12,8707	0,1130
Logistic Regression	0,8951	0,8950	0,8951	0,8949	0,2822	0,0000
Linear Discriminant Analysis	0,8871	0,8864	0,8871	0,8865	0,0269	0,0010
Naive Bayes	0,8808	0,8853	0,8808	0,8811	0,0170	0,0010
Neural Network	0,8776	0,8787	0,8776	0,8779	3,2221	0,0010
SVM (SMO)	0,8760	0,8798	0,8760	0,8730	2,8455	0,0660
Decision Tree	0,8665	0,8659	0,8665	0,8650	0,0840	0,0010
K-Nearest Neighbors	0,7059	0,7042	0,7059	0,7022	0,1912	0,0191

Xulosa

Ushbu tadqiqot kam resursli tillar, xususan, o‘zbek tili uchun sentiment tahlili tizimlarini rivojlantirishning asosiy jihatlarini yoritdi. Tadqiqot natijalari shuni ko‘rsatadiki, sentiment tahlili uchun maxsus leksik-semantik lug‘atlar, belgilangan korpuslar va avtomatlashtirilgan annotatsiya tizimlari zarur. O‘zbek tilining agglutinativ tuzilishi va murakkab morfologiyasi mavjud modellarni to‘g‘ridan-to‘g‘ri qo‘llashda qiyinchiliklar tug‘dirishi aniqlangan. Ushbu tadqiqot natijalari nafaqat o‘zbek tilida, balki boshqa kam resursli tillar uchun ham NLP infratuzilmasini rivojlantirishga hissa qo‘shadi.

Foydalanilgan adabiyotlar:

1. Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. *Proceedings of ACL*, 93-103. https://doi.org/10.18653/v1/2020.acl-main.560
2. Mengliev, D.B., Abdurakhmonova, N.Z., Barakhnin, V.B., Vasliddinova, K., Rahimov, H., & Djalolova, K. (2024). Enhancing Sentiment Analysis in Uzbek Language Texts through Weighted Lexical Features. 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), 2450-2453.
3. Akhmedov, E.Y., Palchunov, D.E., Khaitboeva, D.Z., Ibragimov, M.F., Sultanov, O.R., & Rakhimova, L.S. (2024). Sentiment Analysis in Uzbek Language Texts: a Study Using Neural Networks and Algorithms. 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), 2460-2464.
4. Kumushoy, N., Farogat, Y., & Ogabek, S. (2024). Sentiment Analysis In Uzbek Texts In The Restaurant Field. *2024 IEEE 3rd International Conference on Problems of Informatics, Electronics and Radio Engineering (PIERE)*, 1470-1473.
5. Matlatipov, S., Rahimboeva, H., Rajabov, J., & Kuriyozov, E. (2022). Uzbek Sentiment Analysis Based on Local Restaurant Reviews. *ALTNLP*.
6. Kuriyozov, E., Salaev, U., Matlatipov, S., & Matlatipov, G. (2023). Text classification dataset and analysis for Uzbek language. *ArXiv, abs/2302.14494*.

7. Rabbimov, I., Kobilov, S.S., & Mporas, I. (2021). Opinion Classification via Word and Emoji Embedding Models with LSTM. *International Conference on Speech and Computer*.

8. **Yusufu, A., Ainiwaer, A., Li, B., Li, F., Yusufu, A., & Ji, H.** (2025). *Analysis of user experience in low-resource languages: A case study of the Uzbek language Google Play reviews*. **Information Processing and Management, 62**, 104015. https://doi.org/10.1016/j.ipm.2024.104015

9. **Cohen, J.** (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

10. Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings of LREC*, 417–422.

11. Cambria, E., Havasi, C., & Hussain, A. (2012). SenticNet 2. In *Proc. AAAI IFAI RSC-12*.

SENTIMENT TAHLILI VA UNING JAHON TILSHUNOSLIGIDA O‘RGANILISHI

Malika Suyunova Odil qizi,
Tayanch doktarant
malikasuyunova0@gamil.com
ToshDO‘TAU

Annotatsiya. Ushbu maqola kompyuter lingvistikasi sohasining bir yo‘nalishi bo‘lgan sentiment tahlili va uning jahon tilshunosligida o‘rganilishi masalalarini o‘rganadi. Kompyuter lingvistikasi inson tilini kompyuterlar yordamida tushunish va ishlov berishga qaratilgan ilmiy soha bo‘lib, sentiment tahlilining ahamiyati kundan-kunga ortib bormoqda. Matndagi hissiy holatlarni aniqlash va ularni tasniflash orqali ko‘plab amaliy sohalarda, jumladan, ijtimoiy tarmoqlar, mijozlar sharhlari va boshqa matnlar ustida tahlil qilish mumkin. Maqolada sentiment tahlilining asosiy usullari va metodlari, shuningdek, bu tahlilning tilshunoslikda o‘rganilishiga oid yondashuvlar haqida batafsil ma’lumot berilgan. Bunda, tilning semantik xususiyatlari, kontekstual tahlil, va kross-til tadqiqotlarining muhimligi ta’kidlanadi.

Annotation. This article examines the issues of sentiment analysis, which is a branch of computational linguistics, and its study in the context of world linguistics. Computational linguistics is a scientific field focused on understanding and processing human language with the help of computers, and the importance of sentiment analysis is growing day by day. By detecting and classifying emotional states in text, it is possible to analyze various practical areas, including social networks, customer reviews, and other texts. The article provides detailed information on the main methods and techniques of sentiment analysis, as well as approaches to its study in linguistics. It emphasizes the importance of the semantic characteristics of language, contextual analysis, and cross-lingual research.

Аннотация. Эта статья изучает вопросы компьютерной лингвистики, связанные с анализом сентимента и его исследованием в контексте мировой лингвистики. Компьютерная лингвистика — это научная область, направленная на понимание и обработку человеческого языка с помощью компьютеров, и значение анализа сентимента с каждым днем растет. Определение эмоционального состояния в тексте и его классификация позволяют проводить анализ в различных практических областях, включая социальные сети, отзывы клиентов и другие тексты. В статье представлены основные методы и подходы анализа сентимента, а также подходы к его исследованию в лингвистике. В частности, подчеркивается важность

семантических характеристик языка, контекстуального анализа и кросс-языковых исследований.

Kalit so‘zlar: *sentiment tahlil, tilga ishlov berish, emotsional tahlil, natural language processing (NLP), kross-til tadqiqotlari, semantika, tilshunoslik, ijtimoiy tarmoqlar, sun’iy intellekt, grammatik tahlil.*

Kirish.

Bugungi kunda axborot texnologiyalariga bo‘lgan talab kundan-kunga oshib bormoqda. Dunyodagi barcha sohalarni, korxona va firmalarni kompyuter texnologiyalari qamrab oldi. Texnologiyalar rivojlangani sari ishlab chiqarish mahsulotlarni, xizmatlarni ijtimoiy tarmoqlarda targ‘ib etish orqali raqobat ham kuchayib bormoqda. Mahsulotga bo‘lgan talabning ko‘payishi yoki pasayishi ijtimoiy tarmoqlarda unga berilgan fikrlarning ijobiy yoki salbiy ekanligiga bog‘liq bo‘lmoqda. Bugungi kunda ijtimoiy tarmoqlardagi mahsulotlar va xizmatlarga berilgan fikr-mulohazalarning oqimi nihoyatda kuchayib bormoqda. Bu esa har bir bildirilgan fikrni ijobiy, salbiy yoki neytral ekanligini aniqlashda ko‘p ishchi kuchini talab qiladi. Inson omili bu darajadagi katta ma’lumotni bir kunda yoki bir necha kunda tahlil qila olmaydi. Vaholanki, har soatda ma’lumotlar, sharhlar soni ko‘payib boradi. Shu va shunga o‘xshash til bilan bog‘liq masalalarni kompyuter modellari yordamida hal qilish tabiiy tilni qayta ishlash (NLP) sohasida dolzarb masalardan biri bo‘lib turibdi. Ma’lum bir mahsulot, xizmat yoki brendga inson tomonidan bildirilgan his-tuyg‘ularini tahlil qilish sentiment tahlil deyiladi. Bu tabiiy tilni qayta ishlashning eng faol tadqiqot yo‘nalishlaridan biridir. Sentiment tahlili matndagi hissiy holatlarni aniqlash va kompyuter lingvistikasi sohasida muhim yo‘nalishlardan biridir. Bu metod yordamida matnlarda ifodalangan musbat, salbiy yoki neytral hissiyotlar aniqlanadi va shu asosda turli sohalarda, jumladan, mijozlar sharhlari, ijtimoiy tarmoqlar, reklama materiallari va boshqa matnlar bo‘yicha tahlillar o‘tkaziladi. Sentiment tahlilining o‘rganilishi jahon tilshunosligida ham keng tarqalgan va bu sohada olib borilgan izlanishlar nafaqat tahlil metodlarini, balki tilning semantik, sintaktik va kontekstual xususiyatlarini chuqur tushunishga yordam beradi [1].

Sentiment tahlilining asosiy maqsadi – matnning yoki so‘zning emotsional holatini aniqlashdir. Masalan, biror mijozning mahsulotga yozgan sharhi, ijobiy yoki salbiy munosabatni bildiradi. Sentiment tahlilida kompyuter lingvistikasi texnologiyalaridan foydalanib, matndagi so‘zlarning o‘rni, grammatik tuzilishi, va ularning ma’nosi tahlil qilinadi [2].

Sentiment tahlili asosan uchta asosiy kategoriyaga ajratiladi:

1. Ijobiy: Matn yoki so‘zning mohiyati yaxshi yoki ijobiy hissiyotini bildiradi.
2. Salbiy: Matn yoki so‘z salbiy yoki norozilikni ifodalaydi.

3. Neytrallik: Matn yoki so‘zning hissiy mazmuni aniq ko‘rsatilmagan yoki neytral holatda qoladi.

Bu tahlilni amalga oshirish uchun turli metodlar qo‘llaniladi:

- So‘zga asoslangan tahlil: Bu usulda har bir so‘zning hissiy qiyofasi (ijobiy yoki salbiy) aniqlanadi.
- Fraza va kontekstga asoslangan tahlil: Matnni to‘liq tahlil qilishda so‘zlar o‘rtasidagi kontekst va grammatik tuzilish hisobga olinadi.
- Mashina o‘qish texnologiyalari: Sun'iy intellekt va chuqur o‘qish (deep learning) metodlari yordamida sentiment tahlilining aniqligi va samaradorligi oshiriladi.

Tilshunoslar turli tillarda so‘zlarning semantik (ma’no) jihatlarini tahlil qilishadi. Sentiment tahlili uchun bu tilning o‘ziga xos xususiyatlarini tushunish muhimdir, chunki har bir til o‘ziga xos grammatik va leksik xususiyatlarga ega. Har bir tilning semantik va leksik tizimi turlicha bo‘lgani uchun, sentiment tahlilining translingvistik (tillararo) tadqiqotlari jahon tilshunosligida katta ahamiyatga ega. Tilshunoslar bir nechta tillarda ijobiy va salbiy so‘zlarni qanday aniqlashni va bu so‘zlarning madaniy kontekstda qanday o‘zgarishini o‘rganadilar. Sentiment tahlili, ayniqsa, tillar o‘rtasida farqlar mavjud bo‘lsa, faqat so‘zlarga qarab bajarilsa, xato natijalar berishi mumkin. Shuning uchun, jahon tilshunosligida matnni butunlay tahlil qilishga e’tibor qaratiladi, chunki so‘zlarning ma’nosi ularning konteksti, tuzilishi va madaniy xususiyatlariga bog‘liqdir. Sentiment tahlili nafaqat til strukturasini, balki muallifning niyati, uning madaniy va sotsiologik kontekstini ham o‘rganadi [3]. Tilshunoslar, shu bilan birga, ijtimoiy tarmoqlarda yoki onlayn platformalarda yuzaga kelgan hissiy tahlillarni qanchalik aniq va ishonchli qilish mumkinligini o‘rganadilar. Kompyuter lingvistikasi sohasida sentiment tahlilining o‘rganilishi tilshunoslik va kompyuter fanlari o‘rtasidagi muhim ko‘prikni tashkil qiladi. Bu soha tilning tabiiy tuzilishini kompyuter yordamida tahlil qilish, matnlar orasidagi hissiy holatlarni aniqlash va xalqaro o‘zaro aloqalarni yanada rivojlantirish imkoniyatini beradi. Jahon tilshunosligi uchun sentiment tahlili o‘zaro tilshunoslik tadqiqotlari, madaniy va ijtimoiy nuqtai nazardan tahlil qilish imkoniyatlarini yaratib, tilning emotsional va kontekstual omillarini tushunishga yordam beradi.

Tadqiqotning ahamiyati. Ushbu tadqiqot kompyuter lingvistikasi va sentiment tahlili sohalaridagi ilmiy izlanishlar va amaliy qo‘llanilishini chuqur tahlil qilishga qaratilgan. Bugungi kunda texnologiyalar va sun’iy intellektning rivojlanishi bilan birga, sentiment tahlilining imkoniyatlari kengayib, uning tahliliy va ilmiy jihatlari ham yuksalmoqda. Kompyuter lingvistikasi yordamida matnlarni tahlil qilish, hissiy holatlarni aniqlash va ularni tasniflash jarayonlari turli sohalarda, jumladan, marketing, ijtimoiy tarmoqlar, mijozlar fikrini o‘rganish va mahsulotlar haqida mulohazalar olishda keng qo‘llanilmoqda. Tadqiqotning ahamiyati shundaki, u sentiment tahlilining jahon tilshunosligida qanday o‘rganilayotganini va tilning

semantik va kontekstual xususiyatlarini aniqlashda bu tahlilning rolini ko‘rsatadi. Tilshunoslik va kompyuter fanlari o‘rtasidagi o‘zaro aloqalarni chuqurlashtirish va kross-til tadqiqotlarini rivojlantirish, zamonaviy tilshunoslik va tilni ishlov berish texnologiyalarining yuksalishiga xizmat qiladi. Shu bilan birga, sentiment tahlilining turli tillarda va madaniy kontekstlarda qanday farq qilishi va uning ilmiy va amaliy sohalarga ta’siri tadqiqotning dolzarbligini oshiradi.

Bugungi kunda o‘zbek tilidagi matnlarni sentiment tahlil qilish dolzarb masalardan biridir. Bu esa, o‘z navbatida, ushbu ma’lumotlarni turli maqsadlarda, jumladan, marketing, siyosat, ijtimoiy tadqiqotlar va boshqa sohalarda qo‘llash imkonini beradi. Sentiment tahlil nafaqat matndagi umumiy hissiy yo‘nalishni, balki muallifning muayyan obyektga yoki mavzuga nisbatan munosabatini ham aniqlashga imkon beradi.

Adabiyotlar tahlili. Kompyuter lingvistikasi va sentiment tahlili sohalarida so‘nggi yillarda bir qator muhim tadqiqotlar amalga oshirilgan. Ushbu tadqiqotlarda tilni avtomatik tahlil qilish va hissiy holatlarni aniqlashning turli metodlari va amaliy qo‘llanilishlari o‘rganilgan. Adabiyotlar tahlilida, kompyuter lingvistikasi va sentiment tahlilining asosiy nazariy va amaliy jihatlari ko‘rib chiqilgan bo‘lib, bu sohalarning rivojlanishi va qo‘llanilish imkoniyatlari haqida bir qancha muhim ma’lumotlar taqdim etilgan. Kompyuter lingvistikasi asosan tilni avtomatik tarzda tahlil qilishga qaratilgan. Bu soha bo‘yicha ilk ishlar 1950-yillarda boshlanib, oxirgi yillarda chuqur o‘qish (deep learning) va sun'iy intellekt texnologiyalarining jadal rivojlanishi bilan yanada kuchaygan. Allen (1995) va Jurafsky va Martin (2020) tomonidan olib borilgan tadqiqotlarda, kompyuter lingvistikasi metodlari, jumladan, tilni ishlov berish (NLP) va uning amaliy qo‘llanilishi haqida ko‘plab muhim ma’lumotlar keltirilgan. Bu metodlar yordamida matnlarni tahlil qilish, grammatik tuzilmalarni aniqlash, va semantik ma’nolarni chiqarish mumkin [4, 5].

Jahonda tabiiy tillarni qayta ishlash jarayonida tilning modelini qurish va sun’iy intellekt algoritmlari asosida sentiment tahlil qilish hamda dasturiy modul va vositalari ishlab chiqish bo‘yicha bir qancha tadqiqotlar olib borilgan. Sentiment tahlilning nazariy asoslari S.Yu.Toldova, AA.Bonch-Osmolovskaya, T.Sadikov ishlarida tavsiflangan. Shu jumladan, Stanford universiteri professorlari Christopher Manning, Dan Jurafsky va Percy Liang (Natural Language Processing Group, AQSH), Janubiy Kaliforniya universiteti professorlari Robin Jia va Jesse Thomason(NLP guruhi, AQSH), Oregon universiteti professori Liang Huang (NLP guruhi AQSHda tadqiqotlar olib borilmoqda. Undan tashqari, Heidelberg universiteti (Germaniya) va boshqalarning bu boradagi ishlari qiyosiy o‘rganib chiqildi. Standford universiteti professori Cristopher Manningning tabiiy tilni jarayoni sohasiga qo‘shgan hissasi ilg‘or lingvistik modellar va mashinani o‘rganish usullarining samaradorligini ko‘rsatib, SVMga asoslangan sentiment tahlil qilishda yuqori natijalarni berdi. Uning tadqiqotlari yozib qoldirilgan fikrlarni to‘g‘ri tahlil

qilishda murakkab algoritmlar va keng qamrovli lingvistik tushunchalarning muhimligini ta’kidlab, sentiment tahlil qilishni tushunish va qo‘llashni sezilarli darajada rivojlantirdi[6]. Sentiment tahlilining rivojlanishi 2000-yillarda boshlangan va tezda o‘zining amaliy qo‘llanilishiga ega bo‘lgan. Pang va Lee (2008) tomonidan taqdim etilgan sentiment tahlilining asosiy metodlari, shu jumladan, so‘zga asoslangan tahlil va fraza tahlilining samaradorligi haqida batafsil ma’lumot berilgan. Ian Maas, Ray Yeung va Richard P.L. (2011) larning IMDB dataset bilan amalga oshirilgan ishlari sentiment tahlili sohasidagi asosiy ishlardan biri hisoblanadi. Ular RNN (rekurent neyron tarmoqlari) va SNN (konvolyutsion neyron tarmoqlari) kabi chuqur o‘qitish metodlarini qo‘llaganlar[7]. Socher va boshqalar (2013) Stanford Sentiment Treebank yaratib, bu datasetni sentiment tahlili uchun ishlatganlar. Ular RNN va uning variatsiyalaridan (masalan, LSTM) foydalanib, ma’lum bir tuzilishga ega bo‘lgan matnlarni tahlil qilishni yanada rivojlantirdilar[8]. Zadeh va boshqalar (2016) SMU-MOSI datasetini yaratgan va multimodal sentiment tahlilini o‘rganganlar. Ular matn, rasm va ovoz ma’lumotlarini birlashtirib, ko‘p omilli tahlilni amalga oshirdilar[9].

Sentiment tahlilining dastlabki yondashuvlari, so‘zlarning hissiy qiymatini tahlil qilishga qaratilgan bo‘lsa, keyinchalik chuqur o‘qish va mashina o‘qish texnologiyalari yordamida matnning to‘liq kontekstini tushunishga harakat qilindi (Devlin va boshqalar, 2018). Sentiment tahlili, tilshunoslikdagi eng murakkab masalalardan biri bo‘lib, har bir tilning o‘ziga xos semantik va leksik tuzilmalari bor [10]. Shuning uchun, kross-til sentiment tahlili tadqiqotlari juda muhim ahamiyatga ega. Shen va boshqalar (2018) tomonidan olib borilgan tadqiqotlar kross-til sentiment tahlilining samaradorligini va turli tillar o‘rtasidagi farqlarni aniqlashga qaratilgan. Bu tadqiqotlarda, bir tilning semantik xususiyatlari boshqa tilga qanday ta’sir qilishi, va tilshunoslikning translingvistik tahlili qanday amalga oshirilishi masalalari o‘rganilgan [11].

Sentiment tahlilining tilshunoslikda o‘rganilishi, nafaqat semantik va leksik tahlilni, balki tilning ijtimoiy, madaniy va pragmatik jihatlarini ham o‘z ichiga oladi. Hoveyda va Falk (2017) tomonidan taqdim etilgan tadqiqotlar, sentiment tahlilining madaniy va ijtimoiy kontekstdagi o‘zgarishlarini o‘rganishga qaratilgan [12]. Ular sentiment tahlilining har bir madaniyatdagi hissiy holatlarni qanday aniqlashini va turli til va madaniyatlarning sentiment tahliliga ta’sirini ko‘rsatgan. Sentiment tahlilining kross-til (tillararo) tadqiqotlari amalga oshirildi. Bu metodda, bir nechta tillarda sentimentni aniqlash jarayoni ko‘rib chiqildi. Har bir tilning leksik va semantik xususiyatlari, shuningdek, tilning madaniy konteksti hisobga olindi. Kross-til sentiment tahlilida turli tillarda mavjud bo‘lgan ijobiy va salbiy so‘zlarning farqlari va o‘xshashliklari tahlil qilindi. Sentiment tahlilining samaradorligini o‘lchash uchun bir qator eksperimentlar o‘tkazildi. Ushbu eksperimentlarda, matnlarning turli turdagi sentimentlarini (ijobiy, salbiy, neytral) aniqlash uchun

yuqoridagi metodlar asosida modellar qurildi. Natijalar sinovdan o‘tkazildi va ularning aniqligi (accuracy) va samaradorligi baholandi [13,14]. Eksperimentlar davomida matnlarning turli uzunliklari, mavzulari va tillari hisobga olinib, modellarni sinovdan o‘tkazish jarayoni amalga oshirildi.

O‘zbek tilida ham sentiment tahlil bo‘yicha bir qator ilmiy izlanishlar olib borilmoqda. Masalan, S. Matlatipov “O‘zbek tilining sentiment tahlili” mavzusida ish olib borgan. U o‘zbek tilida raqamli matn ko‘rinishida yozilgan taklif hamda fikrlarni sentiment tahlil qilish masalasini yechish uchun berilganlarning intellektual tahlili usullarini ishlab chiqqan[15]. Samarqand davlat universitetida Ilyos Rabbimov boshchiligidagi bir guruh izlanuvchilar o‘zbek filmlariga qoldirilgan sharhlarni emojilar asosida o‘rganishgan. Sentiment tahlil bo‘yicha, ayniqsa, Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti tadqiqotchilari ilmiy izlanishlar olib bormoqdalar. D. Elova, S. Allanazarova o‘zbek tili matnlarni hissiy jihatdan o‘rganish masalasi bo‘yicha tadqiqotlar olib bormoqda [16].

Munozara. Ushbu tadqiqotda kompyuter lingvistikasi va sentiment tahlili metodlari o‘rganildi. Tadqiqotning asosiy maqsadi, sentiment tahlilining samaradorligini oshirish va turli metodlarni taqqoslash orqali eng maqbul yondashuvni aniqlash edi. Tadqiqot jarayonida so‘zga asoslangan tahlil, mashina o‘qish, chuqur o‘qish va kross-til tahlili kabi turli metodlar sinovdan o‘tkazildi. Har bir metodning afzalliklari va kamchiliklari tahlil qilindi va shu orqali sentiment tahlilining samaradorligini oshirish yo‘llari aniqlanmoqda. Bu metodning oddiyligi va tezligi uning eng katta afzalliklaridan biridir. Lekin uning kamchiligi shundaki, u matndagi kontekstual ma’lumotlarni hisobga olmaydi. Bu esa ba’zan xatolarga olib kelishi mumkin, chunki bir so‘z turli kontekstlarda turlicha ma’no va hissiyotni anglatishi mumkin. Shu sababli, sentiment tahlilining aniqligi va samaradorligini oshirish uchun faqat so‘zga asoslangan metodlar yetarli bo‘lmaydi [17].

Chuqur o‘qish metodlari, masalan, RNN, LSTM yoki Transformer kabi ilg‘or tarmoqlar, sentiment tahlilining aniq va samarali bajarilishini ta’minlaydi. Ular matnning kontekstual ma’nolarini va semantik tuzilishini to‘liq tushunishga imkon beradi. Biroq, bu metodlar katta hisoblash resurslarini talab qiladi va ba’zan yuqori murakkablikka ega bo‘lishi mumkin. Shuningdek, bu metodlarni o‘qitish vaqt va kuch talab qiladi. Kross-til sentiment tahlili, turli tillarda ishlashda juda foydali va global miqyosda keng qo‘llanilishi mumkin. Bu metodlar yordamida bir nechta tillarda matnlarni tahlil qilish va turli madaniyatlardagi hissiy holatlarni aniqlash mumkin. Biroq, tilning o‘ziga xos leksik va semantik xususiyatlarini hisobga olishda qiyinchiliklar mavjud. Har bir tilning o‘ziga xos tuzilishi, jumladan, so‘zlar va frazalar o‘rtasidagi farqlarni hisobga olish, sentiment tahlilini murakkablashtiradi.

Har bir metodning o‘z afzalliklari va kamchiliklari mavjud va ularni bir-biri bilan taqqoslaganda, har biri o‘zining ma’lum holatlarida samarali ishlaydi. So‘zga asoslangan tahlil oddiy vazifalar uchun yaxshi, lekin chuqur o‘qish va mashina o‘qish metodlari murakkab va uzun matnlarni tahlil qilishda yanada yuqori aniqlikni ta’minlaydi. Kross-til tahlili esa turli tillarda va madaniy kontekstlarda ishlashda muhim ahamiyatga ega. Kelajakdagi tadqiqotlar uchun sentiment tahlilining aniqligini oshirish va turli metodlarning integratsiyasi bo‘yicha yangi yondashuvlarni rivojlantirish zarur. Masalan, so‘zga asoslangan tahlil va mashina o‘qish metodlarini birlashtirish, matnning kontekstual va leksik xususiyatlarini yaxshiroq aniqlash imkonini berishi mumkin [18]. Shuningdek, kross-til tadqiqotlari orqali turli tillarda va madaniyatlarda universal sentiment tahlili tizimlarini ishlab chiqish mumkin. Umuman olganda, sentiment tahlili bugungi kunda kompyuter lingvistikasi va tilshunoslik sohalarida keng qo‘llanilayotgan muhim metod bo‘lib, uning ilmiy va amaliy ahamiyati kundan-kunga ortib bormoqda. Shuning uchun, bu sohada yanada chuqurroq izlanishlar olib borish va ilg‘or texnologiyalarni joriy etish zarur bo‘lib, bu nafaqat tilni tushunish jarayonini yaxshilaydi, balki ko‘plab ijtimoiy va iqtisodiy sohalarda ham samarali foydalanishga imkon yaratadi.

Xulosa. Ushbu tadqiqot kompyuter lingvistikasi va sentiment tahlili sohasidagi turli metodlarni o‘rganish va ularning samaradorligini taqqoslashga qaratilgan. Tadqiqotda so‘zga asoslangan tahlil, mashina o‘qish, chuqur o‘qish va kross-til tahlili kabi metodlar keng ko‘lamda tahlil qilindi. Har bir metodning afzalliklari va kamchiliklari aniqlanib, ularning turli vaziyatlardagi samaradorligi ko‘rsatilgan. So‘zga asoslangan tahlil metodlari tez va oson ishlashiga qaramay, matnning kontekstini hisobga olmaydi, shuning uchun ba’zan xatolarga olib kelishi mumkin. Mashina o‘qish metodlari ko‘proq ma’lumotlarga asoslangan holda aniq natijalar beradi, ammo ularning samarali ishlashi uchun katta hajmdagi o‘rgatish ma’lumotlari talab etiladi. Chuqur o‘qish metodlari, xususan Transformer tarmoqlari, matnlarning aniq va to‘liq kontekstini tahlil qilishda juda samarali bo‘lib, yuqori aniqlikni ta’minlaydi, ammo bu metodlar hisoblash resurslarini ko‘p talab qiladi. Kross-til sentiment tahlili esa bir nechta tillarda ishlashda samarali bo‘lib, turli madaniyatlarni hisobga olish imkonini beradi, ammo tillar o‘rtasidagi farqlarni hisobga olishda ba’zi qiyinchiliklarga duch kelishi mumkin. Tadqiqot natijalari shuni ko‘rsatadiki, sentiment tahlilining samaradorligi tanlangan metodga va tahlil qilinayotgan matnning xususiyatlariga bog‘liq. Shu sababli, turli metodlarning kombinatsiyasini qo‘llash, sentiment tahlilining aniqligini oshirishga yordam beradi. Kelajakdagi tadqiqotlarda, metodlarning integratsiyasi va yangi yondashuvlarni ishlab chiqish, xususan, kross-til tahlili va chuqur o‘qish texnologiyalarining birgalikda qo‘llanilishi orqali sentiment tahlilining yanada samarali tizimlarini yaratish mumkin.

Foydalanilgan adabiyotlar:

1. Ismoilov, R. (2020). Kompyuter lingvistikasi: Nazariy asoslar va amaliy qo‘llanilishini o‘rganish. O‘zbekiston Respublikasi Fanlar Akademiyasi.
2. Sattorov, F. (2018). Matnni qayta ishlash va sentiment tahlilining metodologiyasi. Tilshunoslik va kompyuter lingvistikasi, 12(3), 45-59.
3. Azimov, M., & Akhmedov, S. (2021). Tilshunoslik va sentiment tahlili: Yangi yondashuvlar va izlanishlar. O‘zbek tilshunosligi jurnalida 15(2), 118-134.
4. Allen, J. (1995). Natural Language Understanding. Addison-Wesley.
5. Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson.
6. G‘.S.Matlatipov. O‘ZBEK TILINING SENTIMENT TAHLILI. Avtoreferat. 6 b.
7. Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. *The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).b*
8. Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts (2013).*Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank.*
9. Zadeh, A., Chen, M., Liang, J., Poria, S., & Morency, L. P. (2016). CMU-MOSI: A Multimodal Sentiment Analysis Dataset. *Proceedings of the 2016 IEEE International Conference on Data Mining (ICDM)*, 2016, 1-10. https://doi.org/10.1109/ICDM.2016.0042.
10. Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.
11. Shen, D., Xu, Y., & Sun, Y. (2018). Cross-lingual Sentiment Analysis: A Survey. IEEE Access, 6, 10506-10516.
12. Hovy, D., & Falk, C. (2017). Sentiment Analysis and Its Application in Language and Culture. Journal of Linguistic Research.
13. Pang, B., & Lee, L. (2008). Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, 2(1-2), 1-135.
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. NIPS.
15. G‘.S.Matlatipov. O‘ZBEK TILINING SENTIMENT TAHLILI. Avtoreferat.
16. D.Elova, S.Allanazarova (2023). O‘ZBEK TILI MATNLARIDA SENTIMENT TAHLIL USULLARI. O‘zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 1(6). 65-76 b.

17. Davronov, A. (2020). Kompyuter lingvistikasi va sentiment tahlilining o‘zbek tilida qo‘llanilishi. O‘zbekiston fan va texnologiyalar instituti.

18. Karimov, M., & Tashpulatov, D. (2021). Kompyuter lingvistikasi va uning turli tillarda ishlatilishi: Sentiment tahlilining amaliy qo‘llanilishi. Xalqaro ilmiy ishlanmalar 3(5), 44-53.

19. B.Elov, Sh.Hamroyeva, O.Abdullayeva, M.Uzoqova. O‘zbek tilida pos tegging masalasi: muammo va takliflar. O`zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4). 45-63 b.

O‘ZBEK TILIDAGI FRAZEOLOGIK BIRLIKLARNING SINONIMLIK HODISASI

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Omarova Durdona Tajatdin qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
omarovadurdona@gmail.com
ToshDO‘TAU

Annotatsiya Ushbu maqolada o‘zbek tilidagi frazeologik birliklarning sinonimlik hodisasi tahlil qilinadi. Frazeologik sinonimlarning turlari, ularning nutqdagi o‘rni va stilistik imkoniyatlari ko‘rib chiqiladi. Shuningdek, frazeologik birliklarning mazmuniy, uslubiy va differensial sinonimlari tahlil qilinib, ularning til boyligini oshirishdagi ahamiyati yoritiladi.

Annotation. This article analyzes the phenomenon of synonymy in phraseological units of the Uzbek language. It examines the types of phraseological synonyms, their role in speech, and their stylistic potential. Additionally, semantic, stylistic, and differential synonyms of phraseological units are analyzed, highlighting their importance in enriching the language.

Аннотация. В данной статье анализируется явление синонимии фразеологических единиц в узбекском языке. Рассматриваются типы фразеологических синонимов, их место в речи и стилистические возможности. Также анализируются смысловые, стилистические и дифференциальные синонимы фразеологических единиц, освещается их значение в обогащении языка.

Kalit so‘zlar: *frazeologik birliklar, sinonimiya, uslubiy sinonimlar*.

KIRISH

Til qurilishining lug‘at sathida leksik va frazeologik qatlamlar mavjud. Lug‘at sathidagi frazeologik qatlam (uni o‘rganuvchi soha ham) frazeologiya deb yuritiladi. Ilmiy tadqiqotlarda frazeologiyaning hajmini keng va tor tushunish kuzatiladi. Keng tushunishda frazeologiya doirasiga so‘zlarning turg‘un birikmalaridan tashqari maqol, matal, aforizm ham mansub deb qaraladi[1]. Frazeologiya sifatida so‘zlarning semantik bog‘lanishlarini tushunish ham mavjud. Bu hodisalardan faqat bir turini – so‘zlarning turg‘un birikmalarini – til birligi deb baholash mumkin. Olimlar aniq ilmiy kriteriyalarga asoslanib, so‘zlarning barcha turdagi turg‘un birikmalarini emas, balki shulardan bir qismini frazeologik birlik sifatida tan oladilar [2]. Frazeologik birlik bo‘lish uchun so‘zlarning turg‘un

birikmasi obrazli ma’no, ko‘chma ma’no anglatishi shart. Frazeologik birlikdan (iboradan) yaxlitligicha anglashiladigan ma’no uning tarkibidagi so‘z-komponentlarga xos leksik ma’nolarning oddiy yig‘indisiga teng bo‘lmaydi, iboraning ma’nosi umumlashma ma’no, maxraj ma’no sifatida namoyon bo‘ladi. Ma’no yaxlitligiga ega bo‘lgan va nutqiy jarayonga qadar ikki va undan ortiq so‘zlarning barqaror munosabatidan tashkil topgan, nutqqa tayyor holda olib kiriluvchi ko‘chma ma’nodagi barqaror birikmalar *frazeologizm* (*frazeologik birikma*)lardir.

Barqaror birikmalarning eng xarakterli belgilari quyidagilar:

1. Nutq jarayoniga qadar tilda mavjudlik: nutqqa tayyor holda olib kiriladi.

2. Ma’no yaxlitligi.

3. Tuzilishi va tarkibining barqarorligi [3].

Tilning boyligi va ifoda imkoniyatlari uning frazeologik birliklarida o‘z aksini topadi. O‘zbek tilida frazeologizmlar nutqni ta’sirchan, obrazli va ifodali qilishda muhim rol o‘ynaydi. Ushbu tezisda frazeologik birliklarning sinonimlik hodisasi tahlil qilinadi. Frazeologik sinonimlar - bu ma’nosi bir xil yoki juda yaqin bo‘lgan, bir xil so‘z turkumiga mansub, odatda o‘xshash yoki bir xil tuzilishga va birikish xususiyatiga ega bo‘lgan, biroq bir-biridan obrazli asosi, shuningdek, ma’nosi yoki stilistik bo‘yoqlari (yoki ikkalasi birgalikda) bilan farqlanadigan birikmalardir.

MUHOKAMA VA NATIJALAR. Frazeologik birliklar ibora bilan va so‘z bilan ham sinonim bo‘lishi mumkin.

- Shaklan har xil iboralarning bir-biriga yaqin ma’nolarni anglatib kelishi *sinonim iboralar*ni hosil qiladi: *Kayfi buzuq – ta’bi tirriq; toqati toq bo‘lmoq – sabr kosasi to‘lmoq. Yaxshi ko‘rmoq – ko‘ngil bermoq.*

- Lug‘aviy birlik sifatida iboralar so‘zlar bilan ham sinonimik munosabatda bo‘ladi: *tomdan tarasha tushganday – kutilmaganda; og‘zi qulog‘ida – xursand, joni chiqmoq – g‘azablanmoq; ta’bi xira – xafa; quling o‘rgilsin – ajoyib*

Ikki iborani o‘zaro sinonim deyish uchun ular ayni bir ma’noni anglatishi shart. Sinonimlar, odatda, bir yoki bir necha jihatdan farqlanadi, shulardan birima’no qirrasidagi farq hisoblanadi. Masalan, *yer bilan yakson bo‘lmoq – yer bilan yakson qilmoq* iborasi *kulini ko‘kka sovurmoq – kuli ko‘kka sovurildi* iborasiga sinonim: ayni bir ma’noni anglatadi («butunlay yemirmoq, yo‘q qilmoq»). Bu sinonimlar, boshqa belgi-xususiyatlarida umumiylik hosil qilishidan qat’i nazar, ma’no qirrasida farq qiladi: ikkinchisida ma’no bir qadar kuchli ifodalanadi. Shuningdek, *ko‘ngl(i) bo‘sh* va *ko‘ngl(i) yumshoq – yumshoq ko‘ngil* iboralari o‘zaro sinonim bo‘lib, «rahmdil, ko‘ngilchan» ma’nosini anglatadi; birinchi iborada «ko‘ngilchan»lik

ma’no qirrasi, ikkinchi iborada esa «rahmdil»lik ma’no qirrasi ustun. Umuman, frazeologik sinonimlarda farqli ma’no qirralari rang-barang bo‘lib, aniq yondashib izohlashni talab qiladi [2]. Sinonimiya muammosi o‘zining ko‘p qirrali va kognitiv tabiati tufayli kuzatuvchilarning doimiy e'tiborini qozonib bormoqda. Tilshunoslikda sinonimiya bo‘yicha ko‘plab ishlar mavjud bo‘lib, ulardan ko‘pchiligi leksik sinonimiyaga bag‘ishlangan, lekin lingvistlar frazeologik birliklarning sinonimiyasini ham faol o‘rganmoqdalar. “Frazeologizmlarning sinonimligi muammosi tadqiqotchilar tomonidan turlicha hal qilingan. Mashhur rus olimi V.V. Vinogradov shunday deb yozgan edi: "Umummilliy, milliy til stilistikasi tegishli til hodisalarini funktsional farqlanish, o‘zaro bog‘liqlik va ta’sir doirasida ko‘rib chiqadi. U o‘xshash, o‘zaro bog‘liq, parallel yoki sinonimik ifoda vositalarining bir-biriga yaqin ma’nolarni ifodalashdagi o‘rnini hamda turli nutq hodisalarining ekspressiv bo‘yoqlari va ohanglari mos kelishini tahlil qiladi" [4]. Leksika va frazeologik birliklar tilimizning so‘z boyligini tashkil qiladi. Agar lug‘atda til birligi so‘z deb qaralsa, frazeologik tizimda frazeologik birlik (FB) yoki ibora til birligi sifatida qaraladi. So‘z ham, frazeologizm ham leksik birlik sanaladi. Har qanday tilning boyishi faqat yangi so‘zlar hisobiga emas, balki barqaror birikmalar, ya’ni frazeologizmlar hisobiga ham sodir bo‘ladi. Nutq so‘zlovchi yoki yozuvchi tomonidan to‘g‘ri til birliklarini emas, balki muloqot vaziyatiga mos keladigan, fikrlarning aniq va to‘g‘ri tushunilishini ta’minlovchi birliklarni tanlash orqali shakllanadi. Bunday tanlov natijasida bir xil yoki yaqin ma’noni ifodalovchi til birliklari bir-biri bilan bog‘lanib, sinonimik qatorni hosil qiladi. Ma’lumki, ikki yoki undan ortiq so‘zlarning sinonim ekanligini aniqlash uchun ularning ma’nolarini asos qilib olish to‘g‘ri yo‘ldir [5]. Shu sababli, o‘zaro sinonim bo‘lgan so‘zlarning ma’no jihatidan qanday munosabatga ega ekanligini aniqlash asosiy masaladir. Agar masalaga shu nuqtai nazardan yondashilsa, sinonimik qatorlar, sifatlarning semantik xususiyatlari, ma’no nozikliklari, ekspressivlik – baholovchanlik va frazeologik sinonimlar (FS) muayyan uslubiy xususiyatlarga qarab farqlanadi.

Ma’noli so‘zlar va frazeologizmlar badiiy tasvir vositalarining eng muhim usullaridan biri hisoblanadi. Frazeologik sinonimiya badiiy uslubda keng rivojlangan bo‘lib, mualliflarning badiiy til vositalaridan foydalanishdagi ijodiy yondashuvlari, yangi iboralar yaratishi frazeologik sinonimlarning ko‘payishiga, til lug‘at tarkibining boyishiga olib keladi. Shuning uchun har bir frazeologik sinonim turli semantik va uslubiy ma’nolarni ifodalashi bilan ajralib turadi hamda ularning tilning frazeologik tizimida yashashiga imkon yaratadi. Frazeologizmlarning semantik tuzilishi frazeologik ma’no hamda ma’no nozikliklaridan iborat. Harakat yoki hodisa haqida ma’lumot beruvchi frazeologik ifodalar frazeologik ma’no deb ataladi. Frazeologizmlar ham so‘zlar kabi butun bir (yagona) ma’noni ifodalasa ham, ularning frazeologik ma’nosi ko‘p jihatdan leksik ma’nodan farqlanadi. Shu sababli, frazeologizmlar ba’zan so‘zlarga sinonim sifatida ishlatilgan taqdirda ham, ularning

leksik ma’nosi bilan frazeologik ma’nosi teng kelmaydi. Shu o‘rinda Hojiyev fikrini keltirish kerak: “Agar frazeologizmlar o‘z sinonimlari bo‘lgan so‘zlarga nisbatan kuchliroq ma’no darajasini ifodalasa, ular majoziy ma’noga ega bo‘ladi” [6]. Ammo inson faoliyati va ruhiyatida shunday holatlar borki, ularni faqat so‘z orqali tasvirlash juda qiyin. Til bu kabi nozik holatlarni ifodalashda frazeologizmlarga muhim o‘rin beradi. Bunday nozik ma’nolarni aks ettiruvchi frazeologik birliklar odatda bitta so‘z bilan ifodalanmaydi. Frazeologik sinonimlar kuchli va obrazli ifodadir. Ular so‘zlarga nisbatan obrazli tasvir bilan uyg‘unlashib, his-tuyg‘ularni ifodalash uchun xizmat qiladi.

XULOSA

O‘zbek tilida frazeologik birliklarning sinonimlik hodisasi til boyligini ifodalovchi muhim hodisalardan biridir. Frazeologik sinonimlar o‘zaro yaqin yoki bir-birini to‘ldiruvchi ma’noga ega bo‘lib, badiiy nutqning ekspressivligini saqlashga xizmat qiladi. Frazeologik sinonimiya leksik, semantik va stilistik o‘ziga xos xususiyat ega. Ba’zi frazeologizmlar to‘liq sinonimik bog‘lanishda bo‘lsa, ayrimlari emotsional yoki uslubiy rang-barangligi bilan farqlanadi. Bu esa hujjat nutqda qo‘llanish holatini aniqlash va ma’no ifodalashning aniqroq bo‘lishini ta’minlaydi. Frazeologik sinonimlar tilni obrazli oshirib, xalq tafakkuri va madaniyatini aks ettiradi. Ularning tizimli o‘rganilishi tabiiy lingvistik natijalar uchun, balki tarjima, tilni o‘qitish va badiiy asarlarni tahlil qilish uchun ham muhim ahamiyat kasb etadi. Shu bois, o‘zbek tilida frazeologik birliklarning sinonimlik hodisalarini chuqur tadqiq etish ilmiy va amaliy muammolardan biri sanaladi.

Foydalanilgan adabiyotlar:

1. Bashieva S. K. phraseological Stylistic component values: author. dis. d-RA filol.sciences'. – Krasnodar, 1995, – p. 42.
2. Рахматуллаев Ш. Некоторые вопросы узбекской фразеологии (фразеологическая синонимия): Тез.докл. / Ташк. ун-т. Гум.науки. Ташкент, 1968, с.267-269.
3. Sh.Rahmatullayev, N.Mahmudov Z.Xolmanova, I.O‘razova, K.Rixsiyeva. O‘zbek tili frazeologik lug‘ati: /– T.: G‘afur G‘ulom nomidagi nashriyot matbaa ijodiy uyi, 2022. – 636 bet.
4. В. В. Виноградов Об основных типах фразеологических единиц в русском языке (Виноградов В. В. Избранные труды. Лексикология и лексикография. - М., 1977- С. 140-161)
5. Mukhayyo Vafoyeva. Semantic and Stylistic Features of Phraseological Synonyms International Journal for Social Studies Available at https://journals.pen2print.org/index.php/ijss
6. Hajiev A. Explanatory Dictionary of linguistic terms. - Tashkent: National Encyclopedia of Uzbekistan, 2002, - 168 b.

UDK: 811.512.133`42:004.89

GAPNING SEMANTIK STRUKTURASINING ONTOLOGIK MODELINI YARATISH

Xudoykulova Nargiza Ravshanovna,
Magistrant
n.khudaykulova@newuu.uz
TATU

Eshmurodova Madina Dilshod qizi,
Magistrant
madinadilshodovna313@gmail.com
TATU

Annotatsiya. Ushbu tezisda gapning semantik strukturasi tushunchasi va uning ontologik modelini yaratish zarurati tahlil qilinadi. Semantik komponentlarni aniqlash, ularning o‘zaro bog‘liqligi, rollar va munosabatlar tizimini shakllantirish orqali til birliklarining kompyuterda qayta ishlanishiga imkon beruvchi konseptual model ishlab chiqiladi. Tezisda ontologik yondashuv asosida gap tuzilmasining formal tavsifi hamda uni tabiiy tilni qayta ishlash (NLP) tizimlariga integratsiya qilish imkoniyatlari ko‘rib chiqiladi.

Abstract. This thesis examines the concept of the semantic structure of a sentence and the necessity of developing its ontological model. Through identifying semantic components, their interrelations, and forming a system of roles and relations, a conceptual model that enables the computer processing of linguistic units is developed. The formal description of sentence structure based on ontological approaches, and the possibilities of integrating it into Natural Language Processing (NLP) systems, are explored. The thesis also discusses how such a model can be applied in various NLP tasks, such as automatic translation, information extraction, and text understanding.

Аннотация: В данной работе рассматривается концепция семантической структуры предложения и необходимость разработки её онтологической модели. Путём определения семантических компонентов, их взаимосвязей и формирования системы ролей и отношений разрабатывается концептуальная модель, которая позволяет обрабатывать языковые единицы с помощью компьютера. Также анализируется формальное описание структуры предложения на основе онтологического подхода и возможности интеграции модели в системы обработки естественного языка (NLP). В работе рассматриваются способы применения данной модели в таких задачах NLP, как автоматический перевод, извлечение информации и понимание текста.

Kalit so‘zlar: *semantik struktura, ontologik model, gap tahlili, tabiiy tilni qayta ishlash, semantik rollar, konseptual model*

Zamonaviy tilshunoslik va kompyuter lingvistikasi sohasidagi taraqqiyot tabiiy tilning chuqur semantik tahliliga bo‘lgan ehtiyojni sezilarli darajada oshirgan. Ayniqsa, sun’iy intellekt (AI), mashinaviy tarjima, dialog tizimlari, va matnlarni avtomatik tushunish kabi yo‘nalishlar kontekstida gapning semantik strukturasining tahlili markaziy o‘rin tutadi. Gapdagi semantik komponentlar, ularning o‘zaro bog‘liqligi, rollari va munosabatlar tizimi til birliklarining mazmun jihatdan to‘g‘ri qayta ishlanishini ta’minlashga yordam beradi.

Shu sababli, gapning semantik strukturasi va uning ontologik modelini yaratish masalasi, ayniqsa, tabiiy tilni qayta ishlash (NLP) tizimlarida katta ahamiyat kasb etadi. Ushbu modelni yaratish orqali kompyuterlar uchun tilni "tushunish" imkoniyati yaratiladi, bu esa o‘z navbatida o‘zbek tilining raqamli vositalarda qo‘llanilishini kengaytiradi va uning sun’iy intellekt texnologiyalaridagi ishtirokini kuchaytiradi.

Gapning semantik strukturasi va uning ontologik modelini yaratish zarurati, ayniqsa, quyidagi sohalarda dolzarbdir:

1. **Tabiiy tilni qayta ishlash (NLP)**: NLP sohasida, tilning semantik tuzilmasini to‘g‘ri qayta ishlash kompyuterlar uchun tabiiy tilni tushunish va ishlov berish imkonini yaratadi. Masalan, matnni avtomatik tarjima qilish, matnni tahlil qilish va ma’lumotni ajratib olish tizimlarida gapning semantik tuzilmasi to‘g‘ri aniqlanishi zarur.

2. **Sun’iy intellekt (AI) va mashinaviy o‘rganish**: AI tizimlarida tilni tushunish va unga asoslangan qarorlar qabul qilish jarayonida gaplarning semantik tuzilmasi muhim rol o‘ynaydi. Ontologik yondashuv bu tizimlarga yanada yuqori darajada aniq va ishonchli natijalar berishda yordam beradi.

3. **Mashina tarjimasi**: Xususan, avtomatik mashina tarjimasi tizimlarida biror tilda olingan gapning to‘g‘ri semantik tarjimasi, mantiqiy tushunchalar va ularning o‘zaro bog‘liqligini to‘g‘ri aniqlashni talab qiladi. Bu esa gapning semantik strukturasini chuqur tahlil qilishni zarur qiladi.

4. **Korpus lingvistikasi**: Korpus tahlili va matnlarni indeksatsiya qilishda, tilning semantik strukturasini to‘g‘ri aniqlash va tuzish orqali, til birikmalarining mazmunini yaxshiroq tahlil qilish mumkin bo‘ladi. Bu esa tilshunoslik sohasidagi yangi ilmiy yondashuvlarni ishlab chiqish uchun asos bo‘ladi.

5. **Dialog tizimlari**: Sun’iy intellektga asoslangan chat-botlar va boshqa dialog tizimlarida, gaplarning semantik strukturasini to‘g‘ri aniqlash, foydalanuvchining so‘rovlariga aniq va tegishli javoblarni bera olish imkonini beradi.

6. **Bilim bazalari va ontologiyalar**: Ontologik modellardan foydalangan holda bilim bazalarini yaratish, turli sohalarda maxsus bilimlarni tizimli ravishda saqlash va ulardan samarali foydalanish imkonini beradi.

Gapning semantik strukturasini aniqlash va uning ontologik modelini yaratish, tabiiy tilni qayta ishlash (NLP) tizimlari uchun muhim vazifa hisoblanadi. Bugungi kunda tabiiy tilni tushunish va ishlov berish jarayonlarida til birikmalarining semantik tarkibini aniq ifodalash zarurati mavjud. Bu esa, tizimlarga yanada mukammal ishlov berish imkonini beradi. Tabiiy tilda mavjud bo‘lgan shakldoshlarni (ko‘p ma’nolilik) va sintaktik murakkabliklarni aniqlashda, gapning semantik strukturasining to‘g‘ri tashkil etilishi kerak.

Ushbu masala, asosan, quyidagi muammolarni hal qilishni talab etadi:

1. **Shakldoshlarni aniqlash**: Gaplarning bir nechta ma'nolarga ega bo‘lishi, NLP tizimlari uchun muammo yaratadi. Gapning semantik strukturasini to‘g‘ri modellashtirish orqali bunday ambiguitetlarni bartaraf etish mumkin.

2. **Mantiqiy tuzilmalarni aniqlash**: Gapdagi ob’yektlar, ularning rollari, o‘zaro munosabatlarini to‘g‘ri belgilash, NLP tizimlariga yuqori darajadagi semantik tahlil imkoniyatini yaratadi.

3. **Syntagmatik va paradigmatik munosabatlar**: Gapdagi so‘zlarning o‘zaro bog‘lanishlari va ular o‘rtasidagi munosabatlarni aniq aniqlash, tahlil qilish va semantik ma’nolarni tuzishda zarurdir.

Semantik struktura va ontologik modellarning yaratilishida quyidagi texnologiyalar va yondashuvlar qo‘llaniladi:

1. **Natural Language Processing (NLP) texnologiyalari**: NLP sohasidagi texnologiyalar gapni semantik va sintaktik tahlil qilishda, so‘zlar orasidagi munosabatlarni aniqlashda va ularni aniq semantik struktura sifatida ifodalashda ishlatiladi. NLP tizimlari ko‘plab algoritmlar, masalan, ma'lumotlarni ajratib olish, avtomatik tarjima, sentiment tahlili va boshqalar yordamida semantik ma'lumotlarni ishlov berish imkoniyatini beradi.

2. **Ontologiyalar yaratish**: Ontologiyalar yaratish uchun maxsus metodlar va asboblar (masalan, Protégé, OWL) mavjud. Bu asboblar semantik struktura va uning ichidagi tushunchalar o‘rtasidagi munosabatlarni aniq tasvirlashga yordam beradi. Ontologiya yaratish jarayonida semantik bog‘lanishlarni tahlil qilish va modellashtirishga imkon beruvchi texnologiyalar qo‘llaniladi.

3. **Semantik tahlil algoritmlari**: Semantik tahlil uchun maxsus algoritmlar ishlab chiqilgan, ular gapdagi so‘zlar o‘rtasidagi semantik munosabatlarni aniqlaydi. Misol uchun, “*dependency parsing* va *semantic role labeling”* kabi algoritmlar yordamida gapdagi har bir so‘zning ma'nosi va uning boshqa so‘zlar bilan bo‘lgan o‘zaro bog‘lanishlari aniqlanadi.

4. **Mashina o‘rganish (Machine Learning) metodlari**: Mashina o‘rganish algoritmlari gaplarning semantik strukturasini to‘g‘ri modellashtirishda va o‘rganish jarayonida ishlatiladi. Model, oldindan o‘rganilgan ma'lumotlarga asoslanib, yangi gaplarning semantik strukturasini aniqlashga yordam beradi.

5. **Bilim bazalari**: Bilim bazalari va ontologik modellar, tizimlarning semantik ma'lumotlarini saqlash va ishlov berishda qo‘llaniladi. Ular til birikmalarining to‘g‘ri tuzilishini va ularning mantiqiy bog‘lanishlarini tashkil etadi.

Kompyuter lingvistikasi va tabiiy tilni qayta ishlash (NLP) sohasidagi rivojlanishlar, asosan, tilning semantik tuzilmalari va ularning kompyuter tizimlari

tomonidan tushunilishi bilan bog‘liqdir. Bu jarayon, ayniqsa, gapning semantik strukturasining formal tavsiflanishi va uni ontologik model asosida ifodalashni talab qiladi.

Semantik tahlil — bu tabiiy tilni kompyuterlar tomonidan tushunish va qayta ishlashga yo‘naltirilgan ilmiy soha bo‘lib, unda gapdagi ma'no va uning tarkibiy qismlari (predikatlar, sub’ektlar, ob’yektlar va ularning o‘rtasidagi munosabatlar) aniqlanadi. Ontologiya esa bu ma'no va munosabatlarni rasmiylashtirish va tizimlashtirish uchun ishlatiladigan modeldir. Ontologiyalar kompyuter tizimlariga ma'lumotlarni tahlil qilish, qayta ishlash va saqlashda aniq ma'no va munosabatlarni yaratishga yordam beradi.

Gapning semantik tuzilmasi, tilshunoslikda, asosan, gapdagi elementlarning bir-biriga qanday bog‘lanishini va ular orasidagi munosabatlarni tushunishni o‘z ichiga oladi. Masalan, sub’ekt va predikat o‘rtasidagi bog‘liqlik, ob’yekt va predikat o‘rtasidagi munosabatlar va shu kabilar. Tabiiy tilni qayta ishlashda bu tuzilmalarni aniqlash va to‘g‘ri talqin qilish juda muhim.

Ontologik modellashtirish tabiiy tilni qayta ishlashda muhim rol o‘ynaydi. Bu jarayon, tabiiy tilni kompyuterlar tomonidan tushuniladigan va semantik jihatdan to‘g‘ri qayta ishlanishi uchun talab qilinadi. Ontologiyalar tilning semantik strukturasini tizimli va rasmiy tarzda tasvirlashga imkon beradi, shuningdek, semantik tahlil va natural language processing (NLP) tizimlarining samaradorligini oshiradi.

Tilshunoslikda gap tuzilmasi va semantik tahlil masalalari bo‘yicha ko‘plab tadqiqotlar amalga oshirilgan. Xususan:

- **Chomsky N.** (Transformatsion-generativ grammatika asoschisi) – Chomsky gap tuzilmasini formal model asosida tahlil qilishga harakat qilgan. Biroq, uning modeli ko‘proq sintaktik jihatga e’tibor beradi, semantik struktura va ularning ontologik tasnifi yetarli darajada yoritilmagan. Bu, tabiiy tilni qayta ishlashda sintaksis va semantikaning to‘liq integratsiyasining zarurligini ko‘rsatadi.

- **Fillmore Ch.** (Case Grammar muallifi) – Fillmore gapdagi rollarni (agent, object, location va b.) tahlil qilishga harakat qilgan. Biroq, bu model hozirgi NLP va ontologik tizimlarga to‘liq integratsiyalash uchun yetarlicha mos emas, chunki u struktura va semantikaning formal integratsiyasini bermaydi. Bu esa, tilni tushunish va ishlov berish tizimlarida tizimli va chuqur tahlil zarurligini ta’kidlaydi.

- **Paducheva E.V., Apresyan Yu.D.** – Paducheva va Apresyan rus tilshunosligida semantik rollar va tavsiflar masalasini chuqur o‘rganishgan. Ularning ishlari ko‘plab ontologik lug‘atlar va avtomatik tahlil tizimlariga asos bo‘lgan. Biroq, ularning ishlari umumiy modeldan ko‘ra ko‘proq konkret lug‘aviy birliklar

doirasida cheklangan, bu esa semantik tahlilni kengroq doiraga tatbiq etish imkoniyatlarini cheklaydi.

- **Jurayev O., Erkinov T.** kabi o‘zbek tilshunoslari – O‘zbek tilidagi gap tuzilmasi, semantik rollar va morfologik jihatlar bo‘yicha tadqiqotlar olib borgan. Lekin bu ishlarning aksariyatida ontologik modellashtirish usuli chuqur qo‘llanilmagan, ya’ni tilni kompyuter tomonidan tushunishga imkon beradigan formal tizimlar yaratishda yetarli rivojlanish kuzatilmagan.

Yuqoridagi ishlarda semantik struktura masalasi turli jihatlarda yoritilgan bo‘lsa-da, ontologik model yaratish, ya’ni formal, kompyuterga mos tushuncha va munosabatlar tizimini ishlab chiqish ko‘p hollarda yetarli darajada bajarilmagan. Shuningdek, hozirgi zamon NLP tizimlari bilan uzviy bog‘lanish, ma’lumotlar ontologiyasiga asoslangan avtomatik semantik tahlil imkoniyatlari hali to‘liq o‘z yechimini topmagan.

Bunday vaziyatda, ontologik modelni yaratish va uning NLP tizimlari bilan integratsiyasini rivojlantirish, bugungi kunda ilmiy tadqiqotlar va amaliyotda katta ahamiyat kasb etadi. Garchi ba'zi dasturiy yechimlar mavjud bo'lsa-da, ular ko‘pincha aniq, keng qamrovli va kompleks semantik tahlilni taqdim eta olmaydi. Bu esa, shuningdek, mashina o‘rganish va tabiiy tilni qayta ishlash tizimlarining samaradorligini pasaytirishi mumkin.

Semantik tahlilni takomillashtirish uchun, ontologik modellar nafaqat grammatika va sintaksisga asoslangan, balki tilning ma'naviy va mantiqiy aspektlarini o'z ichiga olishi kerak. Yangi avlod ontologik modellarni yaratish, shu bilan birga, NLP texnologiyalarining imkoniyatlarini kengaytirishga xizmat qilishi mumkin. Aynan shu nuqtada, bugungi kunda semantik strukturaning chuqur tahlili va murakkab tushunchalar o'rtasidagi aloqalarni o'rganish, kelajakdagi ilmiy va amaliy tadqiqotlarning asosiy yo'nalishlaridan biriga aylanishi kutilmoqda.

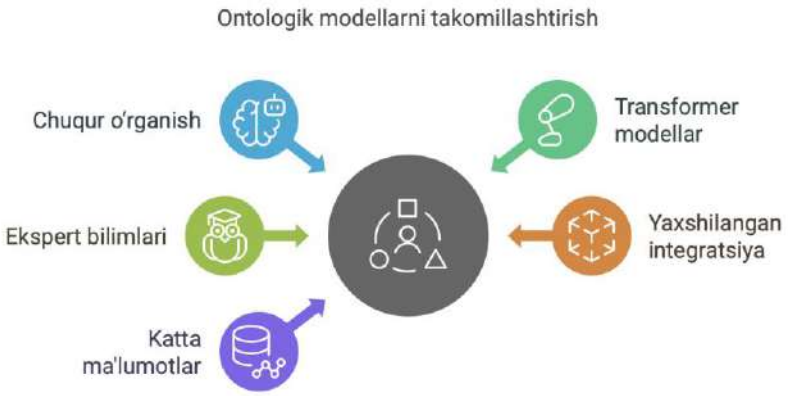
Kelajakda semantik tahlil va ontologik modellarni takomillashtirish uchun bir nechta yo‘nalishlar mavjud.

Birinchidan, tabiiy tilni qayta ishlash sohasidagi yangi yondoshuvlar, masalan, chuqur o‘rganish (deep learning) va transformer modellari, ontologik tahlilni avtomatlashtirishda katta imkoniyatlar yaratmoqda. Ushbu texnologiyalar yordamida, tilning yanada murakkab va nozik jihatlarini tushunish va avtomatik tarzda ma’lumotlarni tahlil qilish mumkin bo‘ladi.

Ikkinchidan, ontologik modellarni joriy qilishda kollaborativ va interaktiv yondoshuvlar muvaffaqiyatli bo‘lishi mumkin. Bunda, ekspert tizimlari yoki foydalanuvchilarning bilimi bilan to‘ldirilgan dinamik ontologiyalar yaratish mumkin. Bu orqali modelning aniqligi va moslashuvchanligi oshadi, chunki

foydalanuvchilar yoki ekspertlar tizimga yangi tushunchalar va munosabatlarni qo‘shish imkoniyatiga ega bo‘ladi.

Uchinchidan, global va dinamik kontekstda semantik tahlilni rivojlantirish uchun, ontologiyalar va ma’lumotlar bazalari o‘rtasida yaxshilangan integratsiya zarur. Ma’lumotlarning doimiy o‘zgarishi va yangi bilimlarning paydo bo‘lishi sharoitida, ontologik modellarni yangilash va moslashtirish tizimi samarali ishlashi kerak.



Shu bilan birga, tilni qayta ishlash sohasida yirik ko‘lamli ma’lumotlarni tahlil qilishda, ma’lumotlar bazalaridan va real vaqt tizimlaridan foydalanish muhim ahamiyatga ega. Bunday yondoshuvlar NLP va ontologik modellarning faoliyatini kengaytirishga yordam beradi, shu bilan birga, ular aniq va tezkor qarorlar qabul qilishda qo‘llaniladi.

Bundan keyin, olingan nazariy bilimlarni amaliy jihatdan qo‘llash masalasi turli sohalarda katta ahamiyatga ega. Ontologik modellarning samarali ishlashi uchun ularning turli ilovalarda, masalan, tibbiyot, moliya, ta’lim va boshqa sohalarda qanday ishlashini ko‘rib chiqish muhimdir. Shuningdek, bu modellarning avtomatik semantik tahlilga asoslangan tizimlar bilan birgalikda ishlash imkoniyatlari hamda ularni real vaqt tizimlarida qo‘llash masalalari o‘rganilishi zarur.

Ontologik modellarni qo‘llashda, ayniqsa, tillarni qayta ishlashda va ma’lumotlarni tahlil qilishda, bunday yondoshuvlar yordamida yangi bilimlar va tushunchalarni kiritish uchun zarur bo‘lgan ilmiy-uslubiy bazani yaratish lozim. Shuningdek, ilg‘or algoritmlar, masalan, mashinani o‘rganish va sun’iy intellekt metodlari, ontologik modellarni tahlil qilish va ishlash jarayonini yanada samarali va aniq qiladi.

Amaliyotda, ontologik modelni yaratishda quyidagi bosqichlar muhim ahamiyatga ega:

1. **Tushunchalar va munosabatlarni aniqlash**: Tilshunoslik va semantik tahlil bo‘yicha mavjud bilimlar asosida, aniq tushunchalar va munosabatlarni shakllantirish zarur. Bu bosqichda tilning o‘ziga xos xususiyatlarini inobatga olish muhimdir.

2. **Modelni yaratish va integratsiya qilish**: Yaratilgan ontologik modelni real tizimga integratsiya qilish, ya'ni ma’lumotlarni avtomatik tarzda tahlil qilish imkoniyatlarini yaratish. Bu bosqichda ma’lumotlar bazalari va NLP tizimlari o‘rtasida samarali integratsiya muhimdir.

3. **Test va optimallashtirish**: Modelni sinovdan o‘tkazish va optimallashtirish. Bu bosqichda ma’lumotlar tahlili natijalari asosida, modelning aniqligini oshirish va uning ishlashini takomillashtirish zarur.

Ontologik modellarni yaratish va ularni semantik tahlil tizimlariga integratsiya qilish jarayonida bir qator muammolar va cheklovlar mavjud. Birinchidan, til va tushunchalarning murakkabligi, inson tilining noaniqliklari va ko‘plab ma’lumotlar turlicha talqin qilinishi ontologiyaning aniqligini kamaytiradi. Ayniqsa, natural tilni qayta ishlash (NLP) sohasidagi tizimlar, o‘z vaqtida yangi bilimlarni integratsiya qilish va yangilash masalalarida qiyinchiliklarga duch keladi.

Ikkinchidan, ontologiyaning kengaytirilishi va uning dinamikasi bilan bog‘liq muammolar ham mavjud. Ma’lumotlarning doimiy o‘zgarib turishi va yangilanishi modelni muntazam ravishda qayta tuzishni talab qiladi, bu esa ishlab chiqish va saqlash xarajatlarini oshiradi. Bundan tashqari, turli sohalarda ishlatiladigan ontologiyalarni moslashtirish va umumiy tizimga integratsiya qilish murakkab bo‘lishi mumkin.

Bundan tashqari, ontologik modellarning samarali ishlashi uchun kuchli va samarali texnologik infratuzilma talab etiladi. Bu tizimlar ko‘pincha katta hajmdagi ma’lumotlar bilan ishlashni talab qiladi, shuning uchun yuqori sifatli serverlar va hisoblash resurslari kerak bo‘ladi.

Biroq, bu muammolar va cheklovlar yangi texnologiyalar, masalan, mashinani o‘rganish va sun’iy intellektning rivojlanishi bilan bartaraf etilishi mumkin. Kelajakda ontologik modellarning ishlashi, ayniqsa, avtomatik semantik tahlil tizimlari bilan birgalikda, yanada aniqroq va samarali bo‘ladi. Shuningdek, yangi algoritmlar va texnologiyalar yordamida ontologiyaning kengaytirilishi va integratsiyasi jarayonlari osonlashadi.

Yuqoridagi ishlarda, semantik struktura va ontologik modellarni yaratishning ilmiy va amaliy ahamiyati ko‘rib chiqildi. Tizimlar va ma’lumotlarni aniq va samarali tashkil qilish uchun ontologiyalar yirik rol o‘ynaydi, chunki ular

bilimlarni tizimli va tushunarli shaklda ifodalash imkonini beradi. Bu, o‘z navbatida, avtomatik semantik tahlilni amalga oshirishda va tabiat tilini qayta ishlashda yanada samarali natijalarni ta’minlaydi.

Biroq, ontologik modellarni ishlab chiqishda va ularni mavjud tizimlar bilan integratsiya qilishda hali ko‘plab muammolar va qiyinchiliklar mavjud. Xususan, tilning noaniqligi, ma’lumotlarning ko‘pligi va tez o‘zgarishi, shuningdek, ilg‘or texnologiyalarning ishlab chiqilishi uchun zarur bo‘lgan infratuzilmaning mavjud emasligi cheklovlar yaratadi. Bu masalalar hali to‘liq hal qilinmagan.

Shunga qaramay, kelajakda yangi texnologiyalar va metodologiyalarning rivojlanishi bilan ontologik modellar va semantik tahlil tizimlarining samaradorligi yanada oshadi. Sun’iy intellekt va mashina o‘rganish sohasidagi yuksak yutuqlar ontologiyalarni tezda yangilash va kengaytirish imkonini beradi, shuningdek, ularni yanada kuchliroq va moslashuvchan qiladi.

Kelajakdagi izlanishlar ontologiyalarni turli sohalarga moslashtirish, ular orasidagi aloqalarni yanada aniqroq va samaraliroq tashkil etishga qaratilgan bo‘lishi kerak. Shu bilan birga, ontologik modellarni amaliyotga keng joriy etish va mavjud tizimlarga integratsiyalash jarayonlari o‘rganilishi va takomillashtirilishi zarur.

Foydalanilgan adabiyotlar:

1. Jurafsky, D., & Martin, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2021.
2. Goldberg, Y. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers, 2017.
3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL-HLT, 2019.
4. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv, 2019.
5. Ляпин, А., Стернин, И., & Михайлова, Л. *Обработка текстов на естественном языке: Современные подходы и методы*. Наука, 2016.
6. Садовничий, В. А., & Ли, В. В. *Семантика и синтаксис: Теории и модели*. ИТМ и ВШЭ, 2020.
7. Гречушкин, С. В., & Казаков, О. И. *Методы синтаксического анализа текста для вычислительных лингвистов*. Физматлит, 2019.

8. Казаков, И. Л., & Попов, И. А. *Нейросетевые методы в обработке текста*. БХВ-Петербург, 2018.
9. Shukurov, B. A. *Tabiiy tilni qayta ishlash: Muammolar va yechimlar*. Fan, 2017.
10. Rakhmatullayev, M., & Tashpulatov, B. *Kompyuter lingvistikasi va tabiiy tilni qayta ishlash*. Universitet, 2019.
11. Xamrayev, B., & Baqoyev, R. *Algoritmlar va tabiiy tilni qayta ishlash texnologiyalari*. Innovatsiya, 2018.
12. Tojiboev, N. *Uzbek tilida semantik tahlil va uning amaliyotda qo'llanilishi*. Qomus, 2020.

SINONIM IBORALARNI ANIQLASH MASALASI TADQIQI

Omarova Durdona Tajatdin qizi
Kompyuter lingvistikasi mutaxassisligi magistranti
omarovadurdona@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada sinonim iboralarni aniqlash masalasi nazariy va amaliy jihatdan o‘rganiladi. Tilshunoslik va sun’iy intellekt nuqtayi nazaridan perefraz tushunchasi, uning usullari va qo‘llanilish sohalari tahlil qilinadi. Shuningdek, tabiiy tilga ishlov berish (NLP) sohasida perefrazlashning roli va uning avtomatlashtirilgan tizimlar uchun ahamiyati muhokama qilinadi. Tadqiqot davomida turli uslubiy yondashuvlar solishtiriladi hamda zamonaviy modellar va algoritmlar yordamida sinonim iboralarni aniqlashning samarali usullari taklif etiladi. Olingan natijalar lingvistik tadqiqotlar, tarjima jarayonlari va axborot qidiruv tizimlarida qo‘llanilishi mumkin.

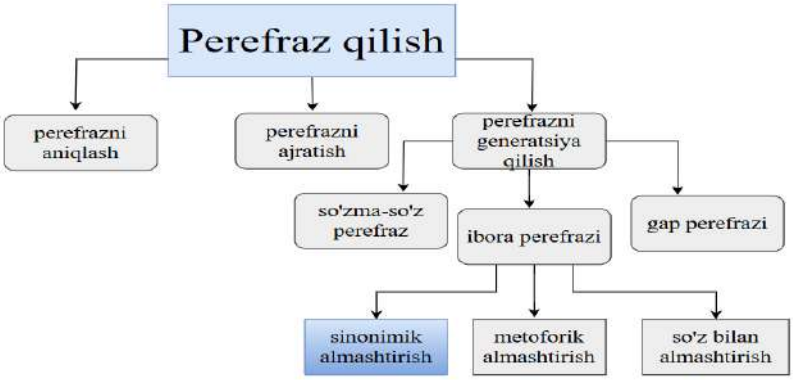
Annotation. This article examines the problem of identifying synonymous phrases from both theoretical and practical perspectives. The concept of paraphrasing is analyzed in terms of linguistics and artificial intelligence, along with its methods and application areas. The role of paraphrasing in the field of natural language processing (NLP) and its importance for automated systems is also discussed. Various methodological approaches are compared during the study, and effective methods for identifying synonymous phrases using modern models and algorithms are proposed.

Аннотация. В данной статье теоретически и практически рассматривается задача определения синонимичных фраз. Анализируется понятие перефразирования с точки зрения лингвистики и искусственного интеллекта, его методы и области применения. Также обсуждается роль перефразирования в области обработки естественного языка (NLP) и его значение для автоматизированных систем. В ходе исследования сравниваются различные методологические подходы, а также предлагаются эффективные методы определения синонимичных фраз с помощью современных моделей и алгоритмов. Полученные результаты могут быть применены в лингвистических исследованиях, процессах перевода и системах информационного поиска.

Kalit so‘zlar: *iboralarnining sinonimlarini aniqlash, tabiiy tilga ishlov berish, NLP, semantik tahlil, tarjima tizimlari, matnni qayta ishlash.*

KIRISH. Barcha tillarning boy lingvistik imkoniyatlarini ko‘rsatib beruvchi vositalardan biri frazeologik birliklardir. Frazeologiya har bir tilning muhim qatlami hisoblanib, uning o‘ziga xos xususiyatlarini yaqqol ko‘zga tashlanishida muhim rol

o‘ynaydi. Frazeologizmlarning tilda qanchalik ko‘p bo‘lishi til imkoniyatlarining shunchalik kengligini ko‘rsatadi. Biroq, turli nutqiy vaziyatlarda iboralarni o‘zgartirish, ularni boshqacha yo‘sinda ifodalash – ya’ni iboralarni perefraz qilish zarurati tug‘iladi. Perefraz qilish – bu matn yoki iboraning ma’nosini saqlab, uni boshqa shaklda qayta ifodalash jarayonidir. Bu jarayon tilshunoslikda keng o‘rganilib, tarjima nazariyasi, pragmatika, stilistika va sun’iy intellekt sohalarida qo‘llaniladi. Ayniqsa, avtomatik tarjima tizimlari va matnlarni qayta ishlash texnologiyalari rivojlanishi bilan iboralarni sinonim iboralarni aniqlash muammosi yanada dolzarb ahamiyat kasb etmoqda. Ushbu maqolada sinonim iboralarni aniqlashning nazariy va amaliy jihatlari tahlil qilinadi, shuningdek, turli tillardagi iboralarni o‘zaro taqqoslash va ularning ekvivalentlarini aniqlash masalalariga e’tibor qaratiladi. Shuningdek, sinonim iboralarni aniqlash jarayonida yuzaga keladigan lingvistik va semantik qiyinchiliklar ham muhokama qilinadi. Tadqiqot natijalari tilshunoslik, tarjima amaliyoti va kompyuter lingvistikasi sohalarida yangi yondashuvlarni ishlab chiqishga xizmat qilishi mumkin. Biz tadqiqotimizda sinonim iboralarni aniqlash va almashtirish usullarini ko‘rib chiqamiz. Bu perefraz jarayoning bir qismi hisoblanadi. Quyidagi rasmda perefraz qilish turlari berilgan biz mana shu turlardan sinonimik almashtirish bosqichini o‘rganamiz. Bu bosqichni o‘rganish uchun umumiy perefraz jarayonini o‘rganish, jahon tajribasida qilingan ishlar bilan tanishish, ular foydalangan usullarni qo‘llab ko‘rish va ulardan samaralilarini belgilab olish kerak bo‘ladi.



1-rasm. Perefraz qilish turlari

Perefraz tushunchasining mohiyatini aniq tushunishga bag‘ishlangan juda ko‘p ishlar mavjud. Bir tomondan, an’anaviy til nazariyalari perefraz tushunchasini qadimgi ritorika an’analarigacha kuzatib borish imkonini beradi (yunoncha "paraphrasis" – "qayta hikoya qilish" ma’nosini anglatadi) va uni turli xil nasr, she’r, musiqa asarlari kabi turli janrlarda keng qo‘llaydi. Bu jarayon matnni boyitish va xilma-xillik kiritish — bir xil so‘z va iboralarning takrorlanishini oldini olib, matnni yanada jozibador va o‘qishga yengil qiladi. Fikrni aniqroq yetkazish — ba’zan

sinonimlar yoki boshqa shaklda ifodalangan iboralar orqali murakkab fikrni oddiyroq, tushunarliroq qilish mumkin. Plagiatni oldini olish — matnni o‘zgartirish orqali boshqa manbadan olgan fikrni o‘z uslubingizda qayta ifodalashga yordam beradi.

MUHOKAMA VA NATIJALAR. Ko‘pgina lingvistik manbalarda "perefraz" tushunchasi quyidagicha talqin qilinadi: "so‘z yoki so‘z birikmasining takrorlanishining oldini olish, hikoyaga katta ifodalilik berish uchun ishlatiladigan ta’rifiy ibora" [1]. Zamonaviy lingvistikada ushbu tushunchaning tarkibida bir necha asosiy komponentlar mavjud: ta’rifiy nomlash usuli, istioraviy (kinoya orqali) ma’no ifodalash va tropik ifoda vositasi sifatida qaralishi.

Rossiyalik olim o‘z tadqiqotlarida "perefraz" (yoki "parafraz") tushunchasiga ikkita ta’rif beradi:

1. O‘z o‘rnida boshqa so‘z yoki iborani almashtiruvchi ta’rifiy ifoda.
2. Oddiy so‘zni ta’rifiy ibora bilan almashtirishdan iborat bo‘lgan trop[2].

Iboralarni perefraz qilish jarayoni tilshunoslikda muhim o‘rin tutadi. Ushbu jarayonning nazariy asoslari semantika, pragmatika va stilistikaga tayanadi.

Perefrazning frazeologiyaga aloqadorligi lingvistikada to‘liq aniqlanmagan masalalardan biri hisoblanadi. L.N. Sinelnikova perefrazga xos bo‘lgan quyidagi xususiyatlarni ajratib ko‘rsatadi:

- mustaqil ifoda shakliga ega bo‘lishi,
- semantik jihatdan bo‘linmasligi,
- so‘z o‘rnini bosuvchi birlik bo‘lishi,
- o‘ziga xos kontekst talab qilishi,
- gapning mustaqil a’zosi sifatida ishlatilishi [3].

Semantik jihatdan, iboralarni perefraz qilish ularning asosiy ma’nosini saqlab qolgan holda yangi shaklda ifodalashni talab qiladi. Pragmatik nuqtayi nazardan, iboralarning kontekstga mos kelishi va aniq tushunarli bo‘lishi zarur. Stilistik jihatdan esa, turli uslublarga mos ravishda iboralarning o‘zgartirilishi mumkin[4].

Iboralarni perefraz qilishning bir necha turlari bor. Eng ko‘p uchraydiganlari quyidagilardan iborat:

Sinonimik almashtirish – iboradagi so‘zlarni ularning sinonimlari bilan almashtirish orqali yangi ifoda yaratish.

Misol: Asl gap: *lekin ularning oldilarida otasining boshi bukildi.*

Perefraz: *lekin ularning oldilarida otasining yuzi yerga qaradi.*

Strukturaviy o‘zgarish – iboraning grammatik tuzilishini o‘zgartirib, uning ma’nosini saqlab qolish.

Misol: Asl gap: *shu arava ustida hali hushi o'ziga kelmagan Nazirqul yotibdi.* Perefraz: *Nazirqul shu arava ustida hali hushi o'ziga kelmagan holda yotibdi.*

Qisqartirish va kengaytirish – iborani yanada qisqa yoki kengroq shaklda ifodalash.

Misol: Asl gap: *Olti oy yoz, olti oy qish, kechani kecha, kunduzni kunduz demay mehnat qildi.*

Perefraz:(qisqartilgan) *Olti oy yoz, olti oy qish, kecha- kunduz demay mehnat qildi.*

Perefraz:(kenggaytirilgan) *Olti oy yoz, olti oy qish, kechani kecha, kunduzni kunduz demay, tinim bilmay, vaqt tanlamay mehnat qildi.*

Metaforik almashtirish – iboralarning badiiy ifodalash shaklini o‘zgartirish orqali yangi uslub yaratish.

Asl gap: *U juda g‘azablandi.*

Perefraz: *Uning ichida olov alanga oldi*.

Sinonim iboralarni aniqlash va ularni almashtirish jarayoni avvalo matnlarni turli uslubiy shakllarda ifodalash va tarjima jarayonlarida muqobil variantlarini tanlashga yordam beradi.

Iboralarni perefraz qilish jarayonida turli lingvistik va semantik qiyinchiliklar yuzaga kelishi mumkin. Eng ko‘p uchraydigan qiyinchiliklar quyidagilardir:

- Ko‘p ma’nolilik (polisemantik) muammosi – ayrim iboralar bir nechta ma’noga ega bo‘lgani uchun ularni sinonim iboralarni aniqlashda noaniqlik yuzaga kelishi mumkin.
- Madaniy tafovutlar – turli tillarda yoki madaniyatlarda iboralarning ifoda tarzi va konnotatsiyasi farqlanadi.
- Grammatik moslik – iboralarning perefraz qilingan shakllari grammatik jihatdan to‘g‘ri bo‘lishi kerak.

Iboralarni perefraz qilishda ko'p ma'nolilik, madaniy tafovutlar va grammatik moslik muammolari bo‘lishi tabiiy. Madaniy va pragmatik moslikni ta'minlash, mos sinonimlarni tanlash va grammatik me'yorlarga rioya qilish perefraz jarayonining sifatini oshirishrishga yordam beradi. Shunday ekan iboralarni sifatli

va to'g'ri qayta ifodalash, sinonimlarini aniqlash va ularni almashtirishda ilmiy izlanishlar, usullar va lingvomadaniy jihatdan tahlil qilib, o‘rganib keyin amalga oshirish lozim.

XULOSA

Ushbu tadqiqotda iboralarni perefraz qilish, sinonim iboralarni aniqlash masalasi atroflicha tahlil qilindi Perefrazlash tilshunoslik va sun’iy intellekt sohalarida muhim ahamiyat kasb etib, tabiiy tilni qayta ishlash, tarjima qilish, matn qisqartirish hamda ma’lumotlarni yanada tushunarli shaklda taqdim etishda keng qo‘llaniladi.Tadqiqot davomida iboralarni sinonim iboralarni aniqlash ning **leksik, sintaktik va semantik** darajalari tahlil qilindi. Shuningdek**, sinonimik almashtirish, strukturaviy o‘zgarish, metaforik qayta ifodalash** kabi yondashuvlarning samaradorligi muhokama qilindi. Natijalar shuni ko‘rsatadiki, iboralarni samarali perefraz qilish, sinonim iboralarni aniqlash uchun nafaqat so‘zlarni almashtirish, balki gap qurilishini o‘zgartirish va kontekstga mos muqobil variantlarni tanlash muhim hisoblanadi. Kelajakda ushbu tadqiqotni yanada rivojlantirish maqsadida avtomatlashtirilgan perefrazlash usullari, jumladan, **neyron tarmoqlar va transformer modellar** asosida ishlab chiqilgan yondashuvlarni o‘rganish va ularning tabiiy til bilan ishlash jarayonidagi samaradorligini tahlil qilish lozim. Shu bilan birga, turli tillarda iboralarni sinonim iboralarni aniqlash ning o‘ziga xos xususiyatlarini aniqlash va qiyosiy tahlil qilish istiqbolli yo‘nalishlardan biri sifatida qaralishi mumkin.

Foydalanilgan adabiyotlar:

1. Головенченко, А.Ф. Перифраз // Краткий словарь литературоведческих терминов / ред.- сост. Л.И. Тимофеев, С.В. Тураев. М.: Просвещение, 1985. 5. Грехнева Л.В. С
2. Ахманова О.С. Словарь лингвистических терминов. М.: Сов.энциклопедия, 1969. 608 с
3. Синельникова Л.Н. К вопросу о сущности перифразы как функционально-семантической единицы речи // Вопросы грамматического строя современного русского языка. М., 1972. С. 255–261
4. Гукова Л.Н., Фомина Л.Ф. Топонимическая перифраза: введение в проблему. Ономастические науки. Киев: Ин-т украинского языка НАН Украины. 2006. № 1. С. 72–81
5. Automatically Ranked Russian Paraphrase Corpus for Text Generation Vadim Gudkov1, Olga Mitrofanova2, Elizaveta Filippskikh3 Saint Petersburg State University1;2 st071220@student.spbu.ru1, o.mitrofanova@spbu.ru2 CraftTalk3 efilippskikh@ crafttalk.ru3

KNN-ALGORITMI YORDAMIDA SO‘Z MA’NOSINI ANIQLASH
(sifat so‘z turkumi misolida)

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Xushmuratova Madina Baxtiyor qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
xushmurodovamadina209@gmail.com
ToshDO‘TAU

Annotatsiya: Kompyuter lingvistikasi sohasining muhim vazifalaridan biri so‘z ma’nosini aniqlashdir. So‘z ma’nosini aniqlash atamasi negizida semantik tahlil elementlari aniqlash masalasi yetakchilik qiladi. Bu elementlar omonim, polisemantik, polifunksional so‘zlar, nomlangan obyektlarni tanib olish (NER) meronim, holonim so‘zlardir. Avtomatik semantik tahlilni amalga oshirish turlicha yondashuvlarni talab qiladi. Zamonaviy yondashuvlardan foydalanish yuqori aniqlikka erishishni ta’minlaydi. Ana shunday yondashuvlardan biri Mashinali o‘qitish yondashuvi bo‘lib, bu yondashuv algoritmlari uchta guruhga bo‘linadi: Nazorat ostida o‘qitish, nazoratsiz o‘qitish, yarim nazorat ostida o‘qitish algoritmlari. Ushbu maqolada nazorat ostida o‘qitish algoritmlaridan bir KNN algoritmi yordamida omonim so‘zlarni semantik farqlash ketma-ketligi keltirilgan.

Annotation: One of the key tasks in the field of computational linguistics is word sense disambiguation. The term “word sense disambiguation” is primarily associated with identifying elements of semantic analysis. These elements include homonyms, polysemantic and polyfunctional words, named entity recognition (NER), meronyms, and homonyms. Performing automatic semantic analysis requires various approaches. Using modern methods ensures high accuracy. One such method is the Machine Learning approach, which is divided into three groups: supervised learning, unsupervised learning, and semi-supervised learning algorithms. This article presents a sequence of semantic disambiguation of homonymous words using one of the supervised learning algorithms-the KNN algorithm.

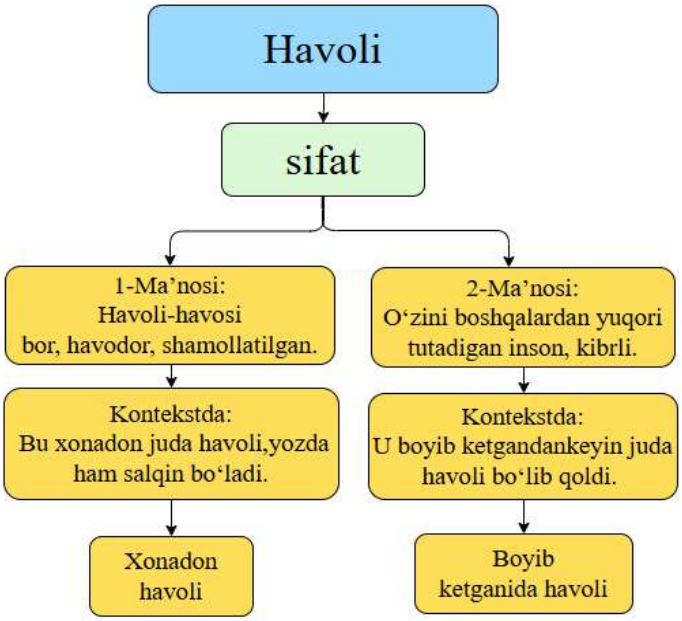
Аннотация: Одной из важных задач в области компьютерной лингвистики является определение значения слова. Термин «определение значения слова» подразумевает выявление элементов семантического анализа. К таким элементам относятся омонимы, полисемантические и полифункциональные слова, распознавание именованных сущностей (NER), меронимы и холонимы. Автоматический семантический анализ требует различных подходов. Применение современных методов обеспечивает

высокую точность. Одним из таких подходов является машинное обучение, которое подразделяется на три группы: обучение с учителем, обучение без учителя и частично контролируемое обучение. В данной статье представлена последовательность семантического различения омонимов с использованием одного из алгоритмов обучения с учителем — алгоритма KNN.

Kalit so‘zlar: *So‘z ma’nosini aniqlash, tabiiy tilga ishlov berish, omonim, mashinali o‘qitish, KNN*

Kirish

So‘z ma’nolarini aniqlash (WSD) tabiiy tilga ishlov berish (Natural Language Processing, NLP) sohasidagi muhim muammolardan biri bo‘lib, axborot izlash, mashina tarjimasi, nutqni tanib olish kabi ko‘plab ilovalar uchun zarurdir. WSD muammosi shundan iboratki, ko‘p ma’noli so‘z (polisemiya), vazifadoshlik (polifunksional), omonim so‘zlar turli kontekstlarda ma’nolarni ifodalaydi va bu so‘zning joriy kontekstdagi ma’nosini aniqlash muhim vazifa hisoblanadi [1]. Mashina odam singari so‘zning ma’nosini to‘g‘ri tushunishi uchun katta hajmdagi ma’lumotlar talab etiladi. Shu sababli, ko‘p hollarda statistik va ehtimollik modellari asosida yondashuvlar ishlab chiqilgan. Buni quyidagi “havoli” so‘zi misolida ko‘rib chiqildi. Bu so‘z bir so‘z turkumida ya’ni sifat so‘z turkumida omonimlikni tashkil qiluvchi so‘z hisoblanadi. Bu omonim so‘z konteksda kelganda qaysi ma’noda kelganini aniqlash uchun avval uning yonidagi qo‘shni so‘zlarini topib, eng ko‘p qaysi so‘z bilan kelishi aniqlanib olib, shu orqali ma’nosi aniqlanishi mumkin:



1-rasm. Havoli so‘zining ma’nolari

WSD muammosini hal qilish uchun turli yondashuvlar mavjud bo‘lib, shulardan eng ko‘p qo‘llaniladigan yondashuv mashinali o‘qitish yondashuvidir.

Mashinali o‘qitish yondashuvi o‘qitiladigan ma’lumotlarning turiga qarab uchta guruhga bo‘linadi:

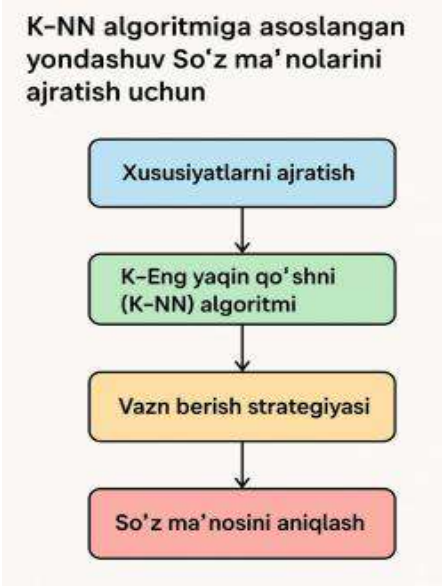
Nazorat ostidagi algoritmlar – annotatsiyalangan (sense-tagged) korpuslardan foydalanib mashinani o‘rgatish [2].

Nazoratsiz algortimlar – belgilanmagan yoki teglanmagan (unlabeled) korpuslar asosida klasterlash usullari orqali ma’nolarni ajratish [2].

Yarim nazorat ostidagi algoritmlar – yuqoridagi ikkala yondashuvni birlashtirgan usullar [2].

Ushbu maqolada nazorat ostidagi algoritmlar guruhiga mansub bo‘lgan KNN algoritmi yordamida so‘z ma’nosini aniqlashni ko‘rib chiqiladi.

K-NN algoritmi. K-NN – nazorat ostidagi mashinali o‘qitish algoritmi bo‘lib, semantik o‘xshashlikka asoslangan klassifikatsiya usulini qo‘llaydi [3]. Noaniq so‘z uchun eng yaqin k ta o‘xshash namunalar topilib, ular asosida ushbu so‘zning ma’nosi aniqlanadi. Bu jarayon ketma-ketligini 2-rasmda ko‘rish mumkin:



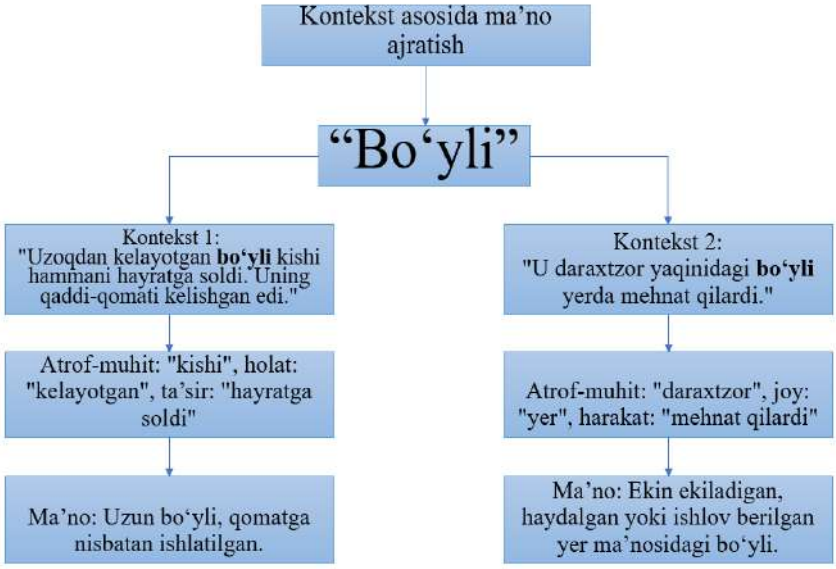
2-rasm. K-NN algoritmi ketma-ketligi

2-rasmdagi jarayonlarni ta’riflab o‘tsak.

Xususiyatlarni ajratish. Bu atama ingliz tilida Feature Extraction ya’ni xususiyatlarni ajratish deb nomlanadi [3]. WSD tizimini ishlab chiqishda xususiyatlarni to‘g‘ri ajratish juda muhim. Xususiyatlar to‘plami quyidagi turlardan iborat:

Ko‘p uchraydigan so‘zlar to‘plami (TF - Term Frequency). Bu usulda, matndagi eng ko‘p uchraydigan so‘zlar tanlanadi. TF asosida noaniq so‘zning biror ma’nosini ifodalaydigan kontekstda eng ko‘p uchraydigan so‘zlar bo‘lib, ular noaniq

so‘zning ma’nosini aniqlashda yordam beradi [4]. Bu jarayonni quyidagi 3-rasmdagi 1 va 2-kontekstlarda uchragan *bo‘yli* omonim so‘zi misolida ko‘rib chiqiladi, bu so‘z sifat so‘z turkumi doirasida omonimlikni hosil qilib, turli kontekstlarda turli ma’nolarni ifodalashi mumkin:



3-rasm. Konteks asosida ma’no ajratish (bo‘yli so‘zi misolida)

“Bo‘yli” so‘zi 1- kontekstda “kishi” so‘zi bilan, 2-kontekstda esa “daraxtzor” so‘zlari atrofida paydo bo‘lgan. Omonim so‘zning eng ko‘p uchraydigan birikuvchilarini hisobga olish orqali kontekst ma’nosi aniqlanadi.

Atrofdagi so‘zlar to‘plami. Noaniq so‘zning atrofidagi so‘zlar hisobga olinadi, masalan, Word2Vec yoki boshqa embedding usullari yordamida [5]. 3-rasmdagi 2 ta kontekstda ham mavjud bo‘lgan “bo‘yli” so‘zining atrofidagi bu so‘z bilan kelgan so‘zlari turlicha. Birinchi jumlada “kishi” va “kelayotgan” so‘zlari bilan, ikkinchi jumlada esa “daraxtzor” va “yer” so‘zlari bilan birikib kelgan. Omonim so‘zning har bir ma’nosi bo‘yicha jamlangan gaplarning soni ortishi bilan ularning birikuvchilari soni ham oshadi, bu esa so‘zning konteks ma’nosi aniqlanishida yordam beradi. Omonim so‘zning kontekst ma’nosini uning birikuvchilari yordamida aniqlashda turli algoritmlardan foydalanish mumkin. Shunday algoritmlardan biri KNN algoritmi.

*K-eng yaqin qo‘shni (K-NN) algoritmi*ning ishlash prinsipi:

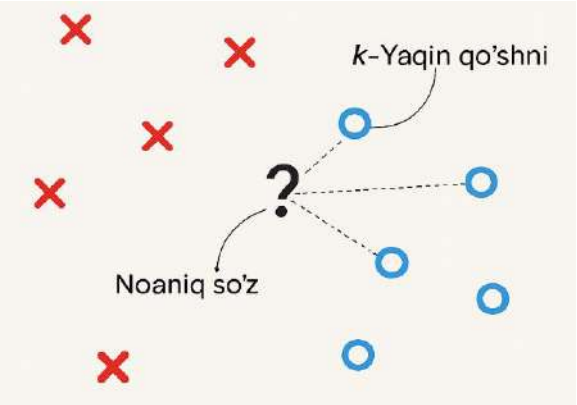
Ma’lumotlarni tayyorlash: Boshqa so‘zlar bilan ajratilgan ma’lumotlar (masalan, matndagi so‘zlar) mavjud bo‘ladi.

Masofa o‘lchovi: So‘zlar orasidagi o‘xshashlikni masofa yordamida o‘lchash mumkin. K-NN algoritmida masofa o‘lchovlari sifatida Evklid masofasi ko‘pincha ishlatiladi.

Eng yaqin k ta qo‘shnini topish: Yangi so‘z uchun eng yaqin k ta qo‘shni so‘zlar tanlanadi.

O‘rganish va klassifikatsiya: Bu k ta qo‘shnilar asosida yangi so‘zning ma’nosi aniqlanadi. Ya’ni, yangi kiritilgan gapdagi noaniq so‘zning ma’nosi, uning atrofidagi k ta qo‘shni so‘zlarning umumiy ma’nosi bilan bog‘lanadi [6].

Sifat so‘z turkumi doirasidagi omonim so‘zlarni knn algoritmi yordamida semantik farqlash jarayonini ko‘rib chiqamiz. O‘zbek tilida faqat sifat so‘z turkumi doirasida omonimlik hosil qiluvchi omonim so‘zlarni “O‘zbek tili izohli lug‘ati” dan yig‘ilganida 82 tani tashkil etib, ular jami 175 ma’noni qamrab oladi. Ushbu omonim so‘zlarning kontekst ma’nosini KNN algoritmi yordamida aniqlashdagi dastlabki jarayon ma’lumotlarni jamlash bo‘lib, omonim so‘zlarning har bir ma’nosini anglatuvchi 100 tacha gaplar jamlandi. Jamlangan gaplar asosida sifat omonimlarni aniqlash uchun KNN algoritmidan foydalanildi. K-NN algoritmi uchun k koeffitsentini tanlash muhim sanaladi. Jahon tajribasiga ko‘ra k=3, 5 kabi belgilash mumkin ekan, asosa k=3 kabi qo‘llanilgan.



4-rasm. Qo‘shni so‘zlarni aniqlash

Vazn berish strategiyasi (Weighting Strategy) so‘zlarning ma’nolarini aniqlashda, ularning matndagi joylashuvi va chastotasi muhim rol o‘ynaydi. Har bir so‘zning vaznini quyidagi formulaga asoslanib hisoblash mumkin:

$$w(k, f_i) = \frac{C(k, f_i)}{C(k)}$$

Bu yerda:

N (k, f-i) – k-ma’nodagi so‘z bilan birga uchraydigan f-i xususiyatli so‘zlar soni,

N (k) – k-ma’nodagi umumiy namunalar soni.

1-jadval.Omonin so‘zlar statistikasi

So‘z	Atrofdagi so‘zlar bilan (%)	Tez-tez uchraydigan so‘zlar bilan (%)	Ikkisini birga ishlatganda (%)
bo‘ydor	90.7	89.7	90.7
bo‘sh	76.8	75.8	76.8
bo‘yli	75.6	72.6	75.6
bog‘li	78.6	82.1	78.6
sirli	61.2	56.9	63.8
belli	70.1	65.7	71.1
O‘rtacha:	75.5	73.8	76.1

O‘rtacha aniqlik: 76.1%

Taklif etilgan metod ilgari ishlab chiqilgan yondashuvlardan yaxshiroq natija ko‘rsatildi.

Xulosa

Mazkur maqolada tabiiy tilni qayta ishlashda muhim hisoblangan so‘z ma’nolarini aniqlash muammosi amaliy jihatdan yoritildi. K-NN algoritmiga asoslangan nazorat qilinuvchi yondashuv yordamida sifat so‘z turkumidagi ko‘p ma’noli so‘zlar kontekst asosida tahlil qilindi. Eksperimentlar natijalari shuni ko‘rsatdiki, kontekst oynasi (Context Window) va so‘z chastotasi asosida xususiyatlar ajratish, shuningdek, vazn berish strategiyalarining qo‘llanishi model aniqligini sezilarli darajada oshiradi. To‘plangan gaplar test qilish natijasida o‘rtacha 76.1% aniqlik kuzatildi, bu esa ilgari mavjud yondashuvlarga nisbatan yuqori ko‘rsatkich hisoblanadi.

Kelajakdagi tadqiqotlarda kontekstga asoslangan chuqur o‘rganish (deep learning) modellarini, jumladan BERT kabi neyron tarmoqlarni qo‘llash, boy va ko‘p tilli korpuslar asosida ma’no ajratish tizimlarini ishlab chiqish, hamda cross-lingual yondashuvlar orqali turli tillarda WSD tizimlarini taqqoslab o‘rganish rejalashtirilgan. Bu esa tilshunoslik va sun’iy intellekt kesishgan sohaga katta ilmiy va amaliy hissa bo‘ladi.

Foydalanilgan adabiyotlar:

1. Boulder, Colorado. Statistical Post-Editing of a Rule-Based Machine Translation System, Proceedings of NAACL HLT 2009: Association for Computational Linguistics, pp 217-224

2. Arabic/English Word Translation Disambiguation using Parallel Corpora and Matching Schemes, 8th EAMT conference, 22-23 September 2008, Hamburg, Germany, pp. 6-11.

3. Abdul-Rauf and Holger Schwenk. On the use of Comparable Corpora to improve SMT performance Sadaf. 12th Conference of the European Chapter of the ACL, pages 16–23, Athens, Greece, 30 March – 3 April 2009.pp 16-23

4. Yarowsky.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 88-95 (1994).

5. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space (2013).

6. Gale, K. Church, and D. Yarowsky.: A Method for Disambiguating Word Senses in a Large Corpus. Computers and Humanities, vol. 26, pp. 415-439 (1992).

7. Baker, P., McEnery, T., Hardie, A. . A glossary of corpus linguistics. Edinburgh: Edinburgh University Press. (2006).

WORDNET ASOSIDA SO‘Z MA’NOSINI ANIQLASH

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Muxammadyakubova Mo‘mina Esojon qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
saidakadirova002@gmail.com
ToshDO‘TAU

Annotatsiya. Tabiiy tilni qayta ishlash (Natural Language Processing, NLP) sohasida so‘zning aniq ma’nosini aniqlash – ya’ni, Word Sense Disambiguation (WSD) – dolzarb va murakkab vazifalardan biridir. Matn kontekstida bitta so‘z turli ma’nolarga ega bo‘lishi mumkinligi sababli, qaysi ma’no aynan shu kontekstga mos kelishini aniqlash talab etiladi. *WSD* nafaqat tilni avtomatik tushunishda, balki tarjima, axborot qidiruvi va boshqa ko‘plab amaliy sohalarda ham muhim rol o‘ynaydi. Hozirgi kunda usullar asosan statistik yondashuvlar va mashinani o‘rganish algoritmlariga tayansa-da, ular katta hajmdagi belgilanganda ma’lumotlarga ehtiyoj sezishi va tilning murakkab semantik munosabatlarini to‘liq aks ettira olmasligi sababli cheklovlarga ega. Shu sababli, so‘nggi yillarda WordNet kabi boy leksik resurslardan, xususan uning bilimga asoslangan imkoniyatlaridan samarali foydalanish g‘oyasi ilgari surildi. Ushbu maqola WordNetning leksik semantik tuzilishi, uning asosiy tarkibiy qismlari, bilimga asoslangan yondashuvi va bu yondashuv asosida so‘z ma’nosini aniqlash uchun ishlab chiqilgan yangi algoritmik metodlarni tahlil qiladi.

Abstract. In the field of Natural Language Processing (NLP), determining the exact meaning of a word - that is, Word Sense Disambiguation (WSD) - is one of the current and complex tasks. Since a single word can have different meanings in the context of a text, it is necessary to determine which meaning corresponds to this context. WSD plays an important role not only in automatic language understanding, but also in translation, information retrieval, and many other practical areas. Although currently the methods are mainly based on statistical approaches and machine learning algorithms, they have limitations due to the need for large amounts of data and the inability to fully reflect the complex semantic relationships of the language. Therefore, in recent years, the idea of effectively using rich lexical resources such as WordNet, in particular its knowledge-based capabilities, has been put forward. This article analyzes the lexical semantic structure of WordNet, its main components, knowledge-based approach, and new algorithmic methods developed based on this approach for determining the meaning of words.

Абстрактный. В области обработки естественного языка (NLP) определение точного значения слова, то есть устранение неоднозначности слов (WSD), является одной из актуальных и сложных задач. Поскольку одно слово может иметь разные значения в контексте текста, необходимо определить, какое значение соответствует этому контексту. WSD играет важную роль не только в автоматическом понимании языка, но и в переводе, поиске информации и многих других практических областях. Хотя в настоящее время методы в основном основаны на статистических подходах и алгоритмах машинного обучения, они имеют ограничения из-за необходимости больших объемов данных и невозможности в полной мере отразить сложные семантические связи языка. Поэтому в последние годы выдвигается идея эффективного использования богатых лексических ресурсов, таких как WordNet, в частности его возможностей, основанных на знаниях. В данной статье анализируется лексико-семантическая структура WordNet, ее основные компоненты, подход, основанный на знаниях, и новые алгоритмические методы, разработанные на основе этого подхода для определения значения слов.

Kalit so‘zlar: *WordNet, Semantik tahlil, So‘z ma’nosi aniqlash, Leksik-semantik tarmoq, Ontologiya, Synset, NLP, Ma’no munosabatlari.*

Kirish. Tabiiy tilga ishlov berish sohasida so‘z ma’nosini aniqlash masalasi sohaning dolzarb masalalaridan biri. So‘z ma’nosini avtomatik aniqlashda turli yondashuvlardan foydalanish mumkin.

- Bilimga asoslangan yondashuvlar
- Mashinali o‘qitishga asoslangan yondashuvlar
- Gibrid yondashuvlar

Bilimga asoslangan yondashuv – oldindan belgilangan lingvistik qoidalar yordamida so‘zning turli kontekstlardagi ma’nolari aniqlanadi. Ushbu yondashuv ham o‘z navbatida qoidaga asoslangan, *WordNet* yoki *Tezaurusga* asoslangan algoritmlarga bo‘linadi.

Tabiiy tilning qat’iy qoidalari asosida so‘z ma’nosini aniqlash mumkin, ammo bu usuldan foydalanish barcha noaniq so‘zlarni aniqlash imkonini bermaydi. Tilning WordNet tizimi yordamida ham so‘z ma’nosini aniqlash mumkin.

WordNet – tabiiy tildagi so‘zlar va ularning o‘zaro semantik munosabatlarini tizimli shaklda aks ettiruvchi yirik leksik ma’lumotlar bazasidir. Ushbu resurs so‘zlarning ma’nolari, sinonimlik, antonimlik va iyerarxik bog‘lanishlarini aniqlik bilan ifodalaydi. Tabiiy tilning WordNet tizimi tilning semantik strukturasini

chuqurroq o‘rganish, kontekstual tahlilni amalga oshirish va morfologik xususiyatlarni (qo‘shimchalar, so‘z yasash tizimi va boshqalar) qamrab oladi.

Tilning semantik strukturasi so‘zlar o‘rtasidagi murakkab bog‘lanishlarni aniqlaydi va ularni iyerarxik tarzda tasniflaydi. *Kontekstual tahlil* esa so‘zning turli kontekstlarda qanday ma’nolarda qo‘llanilishini o‘rganish. *Morfologik moslashuv* o‘zbek tilining boy qo‘shimcha tizimi va so‘z yasash xususiyatlarini hisobga olgan holda leksik ma’lumotlarni shakllantirish demakdir.

WordNetning tizimi bir qancha qismlarga bo‘linadi. Quyida uning asosiy tarkibiy qismlari keltirilgan:

a) Sinonim to‘plamlari bu *Synsets* deb ham nomlanadi. Har bir *Synset*, ya’ni sinonimlar to‘plami – bitta aniq semantik tushunchani ifodalaydi. Masalan, o‘zbek tilida “go‘zal”, “chiroyli”, “jozibali”, “maftunkor” kabi so‘zlar bitta guruh sifatida ko‘rib chiqilishi mumkin. Synset to‘plamlari yordamida semantic aniqlik hamda kontekstual moslashuvchanlik masalalarini yechish mumkin.

b) Umumiy va xos atamalar (Giperonim va Giponim). Bu munosabatlar so‘zlar o‘rtasidagi iyerarxik bog‘lanishni ifodalaydi:

Giperonim: Kengroq tushunchani ifodalaydi, masalan, “transport vositasi”. Giponim: Ushbu umumiy tushunchaga kiritilgan aniqroq tushunchalarni bildiradi. Masalan, “avtomobil”, “velosiped”, “avtobus”. Demak Giponim so‘zlar giperonimlarni tarkibiy qismi bo‘ladi. Giponim va giperonimlar so‘zlarning semantik munosabatlarini yanada aniqroq aks ettirishga yordam beradi.

c) Qismlar va butunlik munosabatlari (Meronim va Holonim). Bu bog‘lanishlar obyektning qismlari bilan uning butunligi o‘rtasidagi aloqani ifodalaydi:

Meronimlar obyektning alohida qismlarini aks ettiradi. Masalan, “uy” so‘zida “eshik”, “deraza”, “tom”. Holonimlar esa qismlardan butunlikni qayta tiklaydi, ya’ni, qismlar asosida to‘liq obyekt tuzilishini aniqlashga imkon yaratadi.

d) Qarama-qarshi ma’nolar. Antonim– so‘zlarning qarama-qarshi yoki zidd ma’nolarini ifodalaydi. Masalan, “yuqori” va “pastki”, “katta” va “kichik” kabi tushunchalar. Antonim so‘zlar yordamida *semantik chegaralarni belgilash, kontekstual tahlil* qilish kabi masalalarni yechishda yordam berishi mumkin.

WordNetning eng katta afzalliklaridan biri – uning semantik munosabatlarni grafik shaklida aks ettirishi hisoblanadi. Har bir so‘z yoki sinonim to‘plami o‘zaro bog‘lanib, ular orasidagi aloqalar orqali ifodalanadi.

WordNet yordamida so‘z ma’nosini aniqlash. So‘z ma’nosini aniqlash tabiiy tilni tushunish va ishlov berishda muhim bosqichlardan biridir. Bitta so‘zning turli kontekstlarda turlicha ma’nolarga ega bo‘ladi.

1-jadval: Omonimlar va ularning ma’nolari

Omonim so‘zlar	So‘z turkumlari	So‘z ma’nolari
olchin	ot	O‘zbek xalqining etnik sos-taviga kirgan urug‘lardan birining nomi.2 Shu uruqqa mansub kishi.
		Uzunlik o‘lchov birligi; ar-shin (q.) va gaz (q.) qiymatlariga teng.
Ombir	ot	Dastasi uzun katta yog‘och to‘qmoq
		ombur.
		Devor tagidan suv o‘tishi uchun qoldirilgan tuynuk; quvur.

WordNet yordamida so‘z ma’nosini aniqlashda bir nechta algoritmlardan foydalanish mumkin:

Leacock-Chodorow o‘lchovi: WordNetdagi so‘zlar orasidagi masofani va iyerarxik bog‘lanishlarni hisobga olib, ikki so‘z orasidagi masofani aniqlaydi. Ushbu o‘lchov yordamida ularning semantik yaqinligi baholanadi. Ikkita so‘z orasidagi semantik o‘xshashlik quyidagi formula yordamida aniqlanadi.

$$sim(c_1, c_2) = -\log \frac{lenght(c_1, c_2)}{2 \times maxDepth}$$

Bu yerda, $sim(c_1, c_2)$ – c_1 va c_2 so‘zlarning o‘xshashlik koeffitsiyenti, $lenght(c_1, c_2)$– c_1 va c_2 so‘zlar orasidagi masofa, $maxDepth$ – butun ontologik daraxtdagi *eng katta chuqurlik*, ya’ni ildiz (root) tugunidan eng chuqur joylashgan tugungacha bo‘lgan masofa yoki balandlik. Bu parametr normalizatsiya uchun kerak. WordNet’da **tushunchalar,** synsetlar shaklida tashkil qilingan, ularning har birining giperonimi mavjud. **maxDepth**ni aniqlash uchun har bir synset uchun ildizgacha bo‘lgan yo‘l uzunligi hisoblanadi. Semantik o‘xshashlikni aniqlashning yana bir usuli Wu-Palmer o‘lchovidir.

Wu-Palmer o‘lchovi: So‘zlar ierarxiyasidagi eng past umumiy domenanta (common ancestor) ga nisbatan har bir so‘zning chuqurligi hisobga olinadi. Natijada, ikki so‘z orasidagi semantik o‘xshashlik aniqlanib, ularning qanchalik yaqin yoki uzoq ekanligi o‘lchanadi.

Formula:

$$sim(c_1, c_2) = \frac{2 \times Depth(LCS)}{Depth(c_1) + Depth(c_2)}$$

c_1, c_2-taqqoslanayotgan ikkita tushuncha

$Depth(c_1), Depth(c_2)$– c_2 va c_1 tushunchalarning WordNet iyerarxiyasidagi chuqurligi, ya’ni ildizdan boshlab nechinchi qatlamda joylashgani

$Depth(LCS)$–“Lowest Common Subsumer” — c_2 va c_1 tushunchalar uchun umumiy bo‘lgan eng chuqur (yaqin) umumlashtiruvchi tushunchaning chuqurligi.

Agar tabiiy tilning WordNet tizimi mavjud bo‘lsa yoki tildagi so‘zlarning sinset to‘plamlari tuzilgan bo‘lsa, uni quyidagicha aniqlash mumkin. O‘zbek tilidagi so‘zlarning sinset to‘plamlaridan foydalanib aniqlash mumkin. Yuqorida ko‘rsatilgan algoritmlar WordNet bilimga asoslangan yondashuvi yordamida turli so‘zlarning ma’nolarini kontekstga mos ravishda farqlash uchun ishlatiladi. Formulalar biror bir dasturlash tilida kodlash orqali hisoblanadi. Nlp masalalarini hal qilish uchun Python eng qulay til hisoblanadi.

```
From nltk.corpus import synset as wn

def get_wupalmer_similarity (w1, w2, pos=’n’):
    sin1 = wn. Synsets(w1, pos=pos)
    sin2 = wn. Synsets(w2, pos=pos)
    if not sin1 or not sin2:
        return None


    max_sim = 0
    for s1 in sin1:
        for s2 in sin2:
            sim = s1.wup_similarity(s2)
            if sim and sim > max_sim:
                max_sim = sim


    return max_sim
```

Izlanishlar davomida o‘zbek tilidagi leksik omonimlarning semantik to‘plamlari tuzilib json formatga joylandi. Hosil qilingan dataset ustida algoritmlardan foydanildi va 1-jadvaldagi kabi natijalar olindi.

2-jadval: Algoritmlarning aniqlik darajasi

Algoritm	Aniqlik (%)	F1-O‘lchov
Lesk	72.5	68.3
Wu-Palmer	78.9	74.1
Leacock-Chodorow	81.2	79.5

Xulosa. Ushbu maqolada WordNetning leksik ma’lumotlar bazasi, uning asosiy tarkibiy qismlari, bilim grafigi yondashuvi va so‘z ma’nosini ajratib ko‘rsatish (WSD) jarayoni batafsil yoritildi. WordNet bilim grafigi yordamida ishlab chiqilgan yangi algoritmik yondashuv, Leacock-Chodorow, Wu-Palmer va Lesk kabi metodlar asosida so‘zlar orasidagi semantik o‘xshashlikni aniqlash imkonini beradi. Eksperimental natijalar ushbu yondashuvning yuqori aniqligi va samaradorligini tasdiqladi. Jahon miqyosidagi izlanishlar va taqqoslashlar asosida, WordNet bilimga asoslangan yondashuvi istiqbolli yo‘nalish hisoblanadi, chunki u tarjima, axborot qidiruvi va sun’iy intellekt tizimlarida keng qo‘llanilishi mumkin.

Kelajakdagi rivojlanish yo‘nalishlari, jumladan, transformer modellar bilan integratsiya, ko‘p tilli qo‘llab-quvvatlash, katta hajmdagi ma’lumotlar ustidagi eksperimentlar va gibrid tizimlar yaratish, ushbu metodologiyani yanada mustahkamlash va kengaytirishga qaratilgan. Natijada, tilni qayta ishlash tizimlarining samaradorligi oshib, foydalanuvchilarga aniq, ishonchli va mantiqiy xizmatlar taqdim eta oladi. Ushbu maqola WordNet bilimga asoslangan yondashuvi yordamida so‘z ma’nosini ajratib ko‘rsatishning yangi yondashuvini ko‘rsatib, tilshunoslik va amaliy dasturlarda yanada ilg‘or natijalarga erishish uchun mustahkam poydevor yaratadi.

Foydalanilgan adabiyotlar:

1. Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, “IndoWordNet va Word Sense Disambiguation uchun Asboblar Tanishtiruvi”, http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/18_Tools%20for%20IndoWordNet%20&%20WSD.pdf
2. C. Fellbaum, “WordNet: An Electronic Lexical Database”, MIT Press, 1998. Lokesh Nandanwar, Kalyani Mamulkar, “Supervised, Semi-Supervised and Unsupervised WSD Approaches: An Overview”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064
3. Kuhoo Gupta, Manish Shrivastava, Smriti Singh va Pushpak Bhattacharyya, “Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi”, COLING/ACL-2006, Sidney, Avstraliya, iyul, 2006.

https://pdfs.semanticscholar.org/1daa/370563deb1e9f5772d6ada1edfb4ec4cd7b4.pdf

4. Banerjee va Pederson, “An adapted Lesk algorithm for Word Sense Disambiguity using Word Net”, http://www.cs.cmu.edu/~banerjee/Publications/cicling2002.pdf

5. Sinha M. va boshqalar, “Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity”, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.3892&rep=rep1&type=pdf

O‘ZBEK TILI MATNLARIDA SO‘Z BIRIKMASINI AVTOMATIK ANIQLASHDA OMONIMIYA MASALASI

Abdullayeva Oqila Xolmo‘minovna,
Filologiya fanlari falsafa doktori (PhD)
abdullayevaoqila@gmail.com
ToshDO‘TAU doktoranti DSc

Nizomova Fotima Bo‘rixo‘ja qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
nizomovaf44@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlashning zarurligi va uning sintaktik tahlildagi ahamiyati o‘rganib chiqilib xulosa va fikrlar berildi. Bunda o‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlashda omonimiya masalasi haqida fikr yuritildi. Maqolada boshqaruvli so‘z birikmasining nazariyasi o‘rganilib chiqildi va undagi omonimiya masalasi ko‘rsatildi. Boshqaruvli so‘z birikmasining teg va modellari ishlab chiqildi va uning ibora (turg‘un bog‘lanma) bilan omonimligi va farqlari haqida ham fikr va xulosalar berildi.

Annotation. In this article, the necessity of automatic identification of word combinations in Uzbek text and its importance in syntactic analysis were examined, along with conclusions and insights provided. The issue of ambiguity (homonymy) in automatic identification of word combinations in Uzbek texts was discussed. The theoretical framework of governed word combinations was studied, and the issue of homonymy within it was highlighted. Tags and models for governed word combinations were developed, and conclusions and insights were also provided regarding their ambiguity and differences compared to idiomatic expressions (fixed phrases).

Аннотация. В этой статье была рассмотрена необходимость автоматического распознавания словосочетаний в текстах на узбекском языке и его важность для синтаксического анализа, а также приведены выводы и наблюдения. Обсуждался вопрос омонимии при автоматическом распознавании словосочетаний в узбекских текстах. Была изучена теоретическая основа управляемых словосочетаний, и выделена проблема омонимии в их рамках. Были разработаны теги и модели для управляемых словосочетаний, а также приведены выводы и наблюдения касаемо их омонимии и различий по сравнению с идиоматическими выражениями (фиксированными фразами).

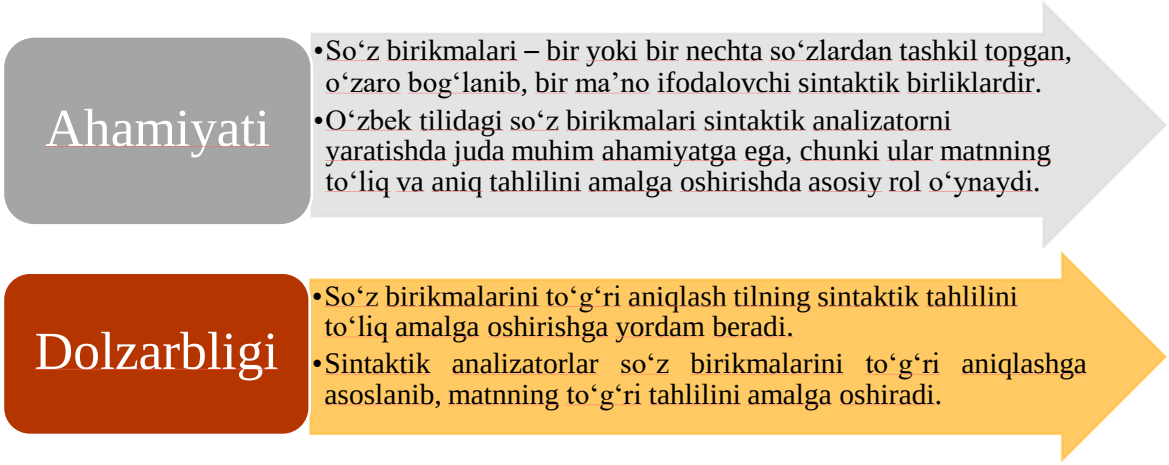
Kalit so‘zlar: So‘z birikmasi, boshqaruvli so‘z birikmasi, turg‘un birikma, erkin birikma, teg, POS teglash, omonimiya.

Kirish. XX asrning ikkinchi yarmida struktur sintaksis, semantik sintaksis yo‘nalishlari paydo bo‘lishi bilan gapning sintaktik tuzilishiga, gap bo‘laklariga nisbatan yondashuv ham o‘zgardi. Formal grammatika talqiniga ko‘ra, faqat kesimgina gapning grammatik yadrosi sifatida e’tirof etildi. Gapning markaziy bo‘lagini belgilashga doir qarashlar dunyo tilshunosligida turlicha bo‘lib kelgan. Fransuz tilshunosi Lyusen Tenyerning verbosentrizm (lotincha “verbum” – “fe’l”, “kesim” ma’nolarini anglatadi, “centrum” – “markaz”) konseptsiyasi fe’lning gap markazi bo‘lgan kesim vazifasida kelishini ifoda etadi.[1]

Sintaksisni tekshirish orfografiyani tekshirishdan ko‘ra murakkabroq. Chunki orfografiyani tekshiruvchi dasturlar yaratilgan, lekin norasmiy matnlarning sintaktik jihatdan avtomatik tahriri masalasi haligacha hal etilgani yo‘q. Rasmiy matnlarning avtomatik tahriri bo‘yicha muhim natijalarga erishilgan [Xolmanova, 2020. 176]

O‘zbek tilida matnlarni avtomatik sintaktik tahlil qilish uchun bugunga qadar kompyuter lingvistikasida amaliy natijalarga erishilgani yo‘q. Ammo, bu sohada sintaktik analizator yaratish uchun dastlabki vazifalar amalga oshirilmoqda. Sintaktik analizator yaratishda so‘z birikmasining tobe va hokim qismlarni bir-biriga bog‘lanish aloqalarini o‘rganib chiqish va ularning teg va modellarini ishlab chiqish oldimizda turgan asosiy vazifalardan biri hisoblanadi.

O‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlashning ahamiyati va dolzarbligi



Boshqaruv aloqasi. Bu aloqada tobe so‘z hokim so‘zning talabi bilan ma’lum forma – shaklga kiradi: tobe so‘z yo biror kelishik qo‘shimchasini oladi, yo biror ko‘makchi bilan birga keladi. Shunga qarab, boshqaruv aloqasi ikki turga ajraladi:

- Kelishikli boshqaruvda tobe so‘z hokim so‘zning talabi bilan tushum, jo‘nalish, o‘rin-payt va chiqish kelishiklaridan birida qo‘llanadi. Bu uning tobeligini ko‘rsatuvchi belgidir: *Iloji bo‘lsa, sizni bu saroydan hovliga ko‘chiramiz* (A. Qodiriy).

- Ko‘makchili boshqaruvda tobe so‘z hokim so‘zning talabiga binoan biror ko‘makchi bilan birga qo‘llanadi. Bu uning tobeligini ko‘rsatuvchi belgidir: 1. *U daraxtlarning qurigan shoxlarini tokqaychi bilan kesdi*. 2. *Chang salomatlik uchun zararlidir*.

Boshqaruv aloqasi orqali obyektli va holli munosabatlar ifodalanadi: bog‘ni tomosha qilmoq (obyektli munosabat), dalaga bormoq (holli munosabat), kitobga chizmoq (obyektli munosabat).

Boshqaruv aloqasida tobe so‘z ot yoki ot o‘rnida qo‘llanadigan so‘zlardan, shuningdek, infinitiv va sifatdoshdan bo‘ladi. Hokim so‘z funksiyasida esa ko‘pincha fe’l qo‘llanadi. Biroq bu vazifada ot, sifat, ravish va boshqa so‘zlar ham kelishi mumkin. Bunda hokim so‘zning o‘zgarishi tobe so‘zga ta’sir qilmaydi.

Misollar: *Oyni etak bilan yopib bo‘lmas* (Maqol).

Yig‘lagani uchun cho‘chidi. (“Mushtum”).

Paxtaga kon O‘zbekiston (Qo‘shiq).

Uyat o‘limdan og‘ir (Maqol).

Qo‘rqqanga qo‘sh ko‘rinar (Maqol).

Ko‘kaldosh ham Alpomishdan kam emas (F.Yo‘ldosh).

O‘zingda yo‘q - olamda yo‘q (Maqol).

Barchin qizim Qorajonga lozimdir (F.Yo‘ldosh). [Abdullayev, 1976. 14].

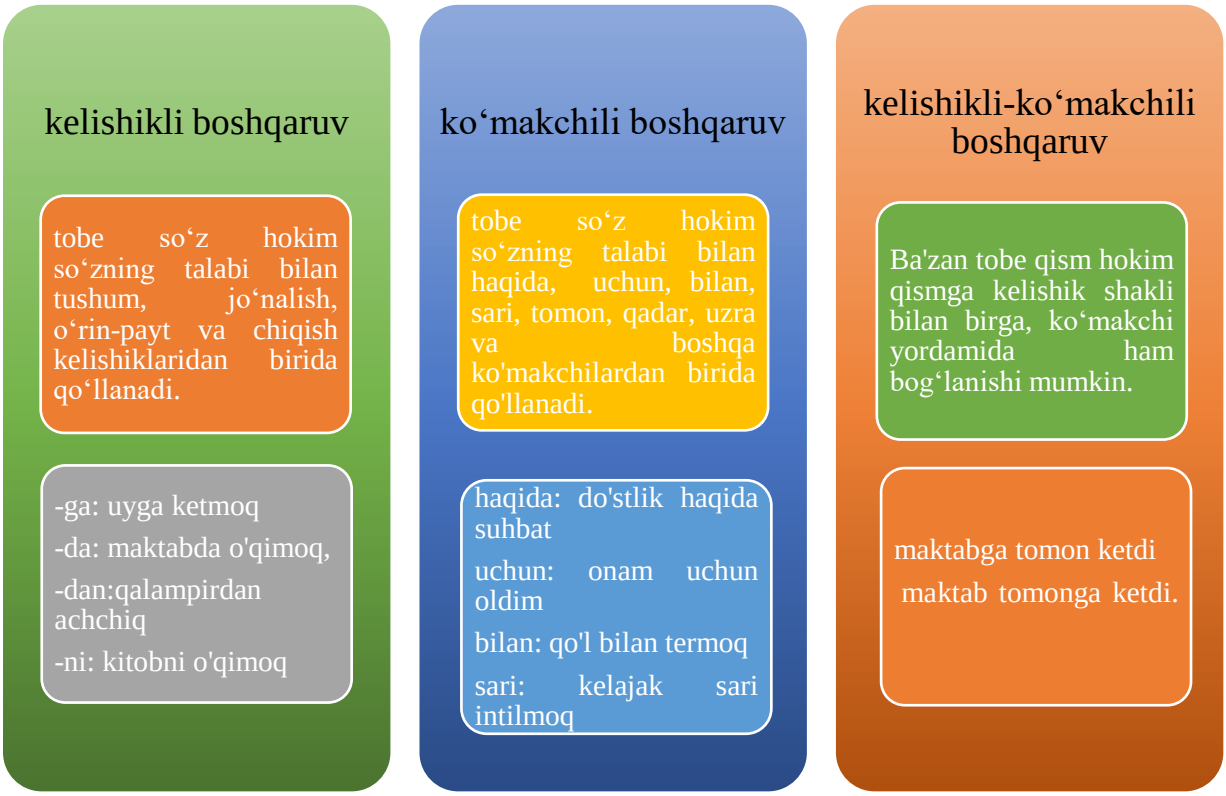
Tobe qismining hokim qismga ko‘makchilar yordamida bog‘lanishi ko‘makchili boshqaruv hisoblanadi: *maktab tomon ketdi, ariq bo‘ylab yurdi*.

Ba’zan tobe qism hokim qismga kelishik shakli bilan birga, ko‘makchi yordamida ham bog‘lanishi mumkin. Bunday ikkita vositali boshqaruv kelishikli-ko‘makchili boshqaruv sanaladi: maktabga tomon ketdi - maktab tomonga ketdi. Shunga ko‘ra, bunday aloqani variantli va variantli bo‘lmagan aloqaga bo‘lish mumkin.

Variantli bo‘lmagan aloqada tobe qism hokim qism talabi bilan faqat bir shaklga ega bo‘ladi: *uyda chiqmoq, Karimni ko‘rmoq*. Variantli aloqada esa tobe qism hokim qismga bir necha shaklda bog‘lanadi: *uzum yemoq – uzumdan yemoq, maktabga ketmoq – maktab tomon ketmoq – maktab tomonga ketmoq* variantlarining har biri o‘zaro umumiylik-xususiylik, aniqlik-noaniqlik ma’nolari bilan farq qiladi. [Mahmudov, Nurmonov, 1995. 14]

Ushbu nazariyalardan xulosa qilib quyidagi sxemani keltirish mumkin.

So‘z birikmasi o‘z ichida 3 turga bo‘linadi:



So‘z birikmalari komponentlarining ma’no mustaqilligi va ularning birikish darajasiga ko‘ra *turg‘un birikma* va *erkin birikmaga* bo‘linadi.

Turg‘un birikma. Tilda bir qolipda, bo‘linmaydigan holda qo‘llanadigan so‘z birikmalari bor. Bunday iboralar so‘zlovchining xotirasida bir butun holatda saqlanadi va nutqda ham xuddi shu holatda qo‘llanadi. So‘zlovchi turg‘un iboralarni so‘zni ishlatgani kabi tanlab ishlatadi. Bunday iboralar gapda bitta so‘roqqa javob bo‘lib, bir bo‘lak vazifasida keladi. Masalan, G‘alaba shodiyonasi millionlar qatori Dildorning ham boshini ko‘klarga ko‘tardi (Y.Sh.) gapida boshini ko‘klarga ko‘tardi – nima qildi? – bitta bo‘lak – kesim vazifasida qo‘llangan. [Abdurahmonov, Sulaymonov, Xoliyorov, Omonturdiyev, 1979. 9]

Turg‘un birikma tarkibiga kiruvchi komponentlar tashqi tomondan grammatik ko‘rsatkichlarga ega bo‘lib, ular ma’lum sintaktik aloqa turlarini hosil qiladilar (boshini ko‘tarmoq — boshqaruvli birikma shakli kabi). Biroq birikma qismlari orasidagi sintaktik aloqa jonli emas, birikma yaxlit holda boshqa biror bo‘lakni boshqarib keladi yoki shu holicha boshqa bo‘lak tomonidan boshqariladi:

Men sevinchimdan terimga sig‘may Xadraga yugurdim kabi turg‘un birikma bir leksik ma’no anglatuvchi bo‘lib qolganligi – leksikalizatsiyalashganligi uchun leksik birlik hisoblanadi. [Abdurahmonov, Sulaymonov, Xoliyorov, Omonturdiyev,1979. 10].

Turg‘un birikmalar, asosan, ikki yirik guruhga bo‘linadi [Abdurahmonov, Sulaymonov, Xoliyorov, Omonturdiyev,1979. 10]*:*

murakkab terminlar:

• Sovet Ittifoqi Kommunistik partiyasi, Termiz Davlat pedagogika instituti, Janubiy Surxon suv ombori, O‘rta Rus pasttekisligi, "O‘qish kitobi", ishlab chiqarish boshqarmasi kabi;

frazeologik iboralar:

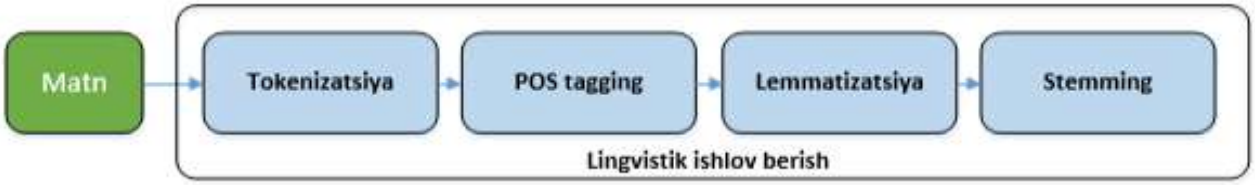
• Og‘zing qani desa, qulog‘ini ko‘rsatmoq; Tegirmonga tushsa butun chiqmoq; Yulduzni beneravon urmoq; Qamchisidan qon tommoq; Orasidan qora mushuk o‘tmoq kabi

So‘z birikmasi va ibora (turg‘un bog‘lanma)ning farqi [Rasulova, 2018. 15]

So‘z birikmasi	*Ibora (turg‘un bog‘lanma)*
• *So‘z birikmasida so‘zlar o‘zaro erkin bog‘langan bo‘ladi*	• *Iboralar turg‘un bog‘langan, so‘zlar ma'noviy butunlik uchun birlashgan-yaxlitlangan bo‘ladi*
• *Har ikki qism alohida tushuncha anglatadi: ko‘z yummoq: ko‘z-a’zo, yummoq - harakat*	• *Ibora bir leksik ma’no anglatadi: ko‘z yummoq-vafot etmoq, ko‘z yummoq-e’tibor bermaslik*
• *So‘z birikmasi lug‘atlarda berilmaydi*	• *Turg‘un bog‘lamaga bir so‘z sifatida qaraladi, lug‘atlarda so‘zlar qatorida beriladi, chunki ular so‘zlar kabi ma'no ifodalaydi: yog‘ tushsa yalagudek-toza*
• *So‘z birikmasidagi so‘zlar erkin bog‘langani uchun almashtirish mumkin: baxtdan kuylab-kuylab baxtdan.*	• *Turg‘un bog‘lamadagi so‘zlarni almashtirib bo‘lmaydi: oq yo‘l. Iboralar bir ma’no ifodalash qonuniyati asosida turg‘unlashgan, yaxlitlangan bo‘ladi*
• *So‘z birikmasi tarkibidagi so‘zlar hokim va tobe qismga ajratiladi*	• *Iboralar bir so‘zga teng, ajralmas holga kelib qolgan, uning tarkibidagi qismlarni hokim va tobe so‘zga ajratish mumkin emas.*
• *So‘z birikmasi tarkibidagi so‘zlarga alohida-alohida savol beriladi: aravani (nimani) olib qochmoq (nima qilmoq)*	• *Bir leksik ma’no, ya’ni bir so‘z (tushuncha)ni ifodalagani uchun iboralarga bitta so‘roq beriladi: aravani quruq olib qochmoq iborasiga nima qilmoq so‘rog‘i beriladi.*

Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo‘linadi: semantik teg, sintaktik teg va grammatik teg. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur.

POS teglash – bu berilgan gapdagi har bir so‘zshaklga uning turkum (ot, fe’l, sifat, son, ravish yoki olmosh)ga mansubligini belgilash (teglash) vazifasidir. POS teglash tabiiy tilni qayta ishlash (Natural Language Processing, NLP)ning asosiy vazifalaridan biri bo‘lib, pipeline konveyerining muhim bosqichi hisoblanadi (1-rasm).



1-rasm. Matnga boshlang‘ich ishlov berish bosqichlari

O‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlashda dastur dastlab matnni tokenizatsiya qiladi. Tokenlarni esa POS tagging jarayonidan o‘tkazadi, lemmasini aniqlaydi va stemlaydi. So‘z birikmasini avtomatik aniqlashda POS tagginglarni quyidagi ro‘yxat asosida amalga oshiriladi.

№	POS teg	Ingliz	Belgilanishi
1.	**Ot**	Noun	N
2.	**Sifat**	Adjective	JJ
3.	**Fe’l**	Verb	VB
4.	**Son**	Number	NUM
5.	**Ravish**	Adverb	RR
6.	**Olmosh**	Pronoun	P
7.	**Bog‘lovchi**	Conjunction	C
8.	**Ko‘makchi**	preposition	II
9.	**Yuklama**	particle	Prt
10.	**Modal so‘z**	modal words	MD
11.	**Undov so‘z**	exclamatory words	UH
12.	**Taqlid so‘zlar**	imitation words	IM

2-rasm. O‘zbek tilidagi POS teglar ro‘yxati (qisqa variant)

So‘z birikmasi va ibora (turg‘un birikma)ning shakliy o‘xshashligi POS tegging jaroyonida omonimlikni keltirib chiqaradi. Bu esa dasturiy ta’minotda murakkablikni keltirib chiqaradi. Tokenizatsiya jarayoni gapni, yoki matnni tahlil qilganda uni qismlarga bo‘lib tashlab, POS teglash jarayoniga tayyorlab beradi. Bunda o‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlashda tokenizatsiya

so‘z birikmasini alohida tokenlarga ajratib berishi, ibora (turg‘un birikma)ni esa yaxlit bir holatda holatda POS teglash jarayoniga o‘tkazib berishi zarur. Buni quyidagi jadval orqali aniq ko‘rishimiz mumkin.

Yigit oqimga qarshi suzib ariqning *boshiga yetdi*-**so‘z birikmasi**

№	So‘z	O‘zak	POS teg	Izoh
1.	Yigit	yigit	Ot	“yigit” alohida leskema, tarkibiy qismlarga bo‘linmaydi
2.	Oqimga	oq	Ot	“oqim” oq+im; -im so‘z yasovchi qo‘shimcha; leksema+grammatik shakl
3.	Qarshi	qarshi	sifat	“qarshi” alohida leskema, tarkibiy qismlarga bo‘linmaydi
4.	Suzib	suz	fe’l	suz+ib lug‘aviy shakl yasovchi; leksema+grammatik shakl
5.	Ariqning	ariq	ot	ariq+ning sintaktik shakl yasovchi; leksema+grammatik shakl
6.	Boshiga	bosh	ot	bosh+i+ga sintaktik shakl yasovchi; leksema+grammatik shakl+grammatik shakl
7.	Yetdi	yet	fe’l	Yet+di sintaktik shakl yasovchi; leksema+grammatik shakl

Qo‘rsligi, be’mani gaplari axiri uning *boshiga yetdi*-**ibora**

№	So‘z	O‘zak	POS teg	Izoh
1.	Qo‘rsligi	Qo‘rs	ot	Qo‘rs+lik+i; so‘z yasovchi+egalik qo‘shimchasi; leksema+grammatik shakl yasovchi+grammatik shakl yasovchi
2.	Bema’ni	ma’ni	sifat	Shakl yasovchi+o‘zak; yasama so‘z
3.	Gaplari	gap	ot	Gap+lar+i; leksema+grammatik shakl yasovchi+grammatik shakl yasovchi
4.	Axiri	axiri	yuklama	Axiri; yuklama
5.	Uning	u	olmosh	U+ning; leksema+grammatik shakl
6.	Boshiga yetdi	Boshiga yetdi	ibora	Bir leksik ma’noga ega bo‘lgan ibora.

Xulosa. O‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlash dasturi kompyuter lingvistikasi sohasida muhim ahamiyat kasb etadi. Ushbu jarayonni amalga oshirishda omonimiya masalasi eng dolzarb muammolardan biri hisoblanadi. So‘z birikmasi va turg‘un iboralarning shakliy o‘xshashligi POS tegging jarayonida murakkablik tug‘diradi. Shuning uchun dasturiy ta’minotni ishlab chiqishda turg‘un iboralarni yaxlit birlik sifatida aniqlash va ularni erkin so‘z birikmalaridan farqlash alohida algoritmik yondashuvni talab qiladi. Boshqaruvli so‘z birikmalari nazariyasini chuqur o‘rganish va ularning teg hamda modellari asosida ishlov berish dasturiy yechimning asosiy poydevorini tashkil etadi. Ushbu yo‘nalishda olib borilgan izlanishlar kelajakda o‘zbek tili matnlarini avtomatik tahlil qilishga mo‘ljallangan ilg‘or tizimlarni yaratish imkonini beradi. Shu bilan birga, o‘zbek tili kabi morfologik jihatdan boy tillarni qayta ishlashga oid tadqiqotlar xalqaro lingvistik va texnologik hamjamiyatga ulkan hissa qo‘shadi.

Foydalanilgan adabiyotlar:

1. Теньер Л. Основы структурного синтаксиса. - М.: Прогресс, 1988.
2. Elov B.B. Xudayberganov N.U. O‘zbek tili korpusi matnlarini pos teglash usullari: file:///C:/Users/User/Downloads/elov-b.b.-xudayberganov-n.u.-ozbek-tili-korpusi-matnlarini-pos-teglash.pdf
3. Xolmanova Z. Kompyuter lingvistikasi. -Toshkent, 2020.
4. Mahmudov N., Nurmonov A. O‘zbek tilining nazariy grammatikasi. – Toshkent, 1995.
5. Abdurahmonov G‘., Sulaymonov A., Xoliyorov X., Omonturdiyev J., Hozirgi o‘zbek adabiy tili; Sintaksis – Toshkent, 1979.
6. Rasulova N. Ona tilidan ma’ruzalar. – Toshkent, 2018.
7. Abdullayev F.A. O‘zbek tili grammatikasi. II jild. – Toshkent, 1976.

SENTIMENT TAHLILIDA SO‘Z TURKUMLARINING AHAMIYATI.

Xusainova Zilola Yuldashevna,
filologiya fanlari falsafa doktori (PhD)
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Xusanova Diyora To‘lqin qizi,
Kompyuter lingvistikasi yo‘nalishi 3-kurs talabasi
tolkinovnadiyora@gmail.com
ToshDO‘TAU

Bekqulova Sevinch Furqat qizi
Kompyuter lingvistikasi yo‘nalishi 3-kurs talabasi
bekqulovasevinchim@gmail.com
ToshDO‘TAU

Annotatsiya. Hozirgi kunda sentiment tahlili NLPning eng faol tadqiq etilayotgan yo‘nalishlaridan biri bo‘lib, matnlardagi hissiy holatni aniqlashda samarali vosita hisoblanadi. Ko‘pgina sentiment tahlil tizimlari so‘zlar semantikasi va kontekstga asoslanadi, biroq so‘z turkumlari (masalan, sifatlar, fe’llar, otlar)ning o‘ziga xos hissiy ulushi yetarlicha hisobga olinmaydi. Ushbu maqolaning asosiy maqsadi turli so‘z turkumlarining sentiment tahlilidagi funksianal rolini aniqlash va baholashdan iborat.

Аннотация. Анализ настроений в настоящее время является одним из наиболее активно исследуемых направлений НЛП и представляет собой эффективный инструмент для определения эмоциональных состояний в текстах. Большинство систем анализа настроений основаны на семантике слов и контексте, но недостаточно учитывают эмоциональное содержание классов слов (например, прилагательных, глаголов, существительных). Основная цель данной статьи — выявить и оценить функциональную роль различных классов слов в анализе тональности.

Abstract. Sentiment analysis is currently one of the most actively researched areas of NLP and is considered an effective tool for identifying emotional states in texts. Most sentiment analysis systems are based on word semantics and context, but the specific emotional content of word classes (e.g., adjectives, verbs, nouns) is not sufficiently taken into account. The main goal of this article is to identify and evaluate the functional role of different word classes in sentiment analysis.

Kalit so‘zlar: *sentiment tahlil, so‘z turkumlari, POS teg, sentiment baholash, sentiment teglash.*

KIRISH

Sentiment tahlil yoki his-tuyg‘ularni tahlil qilish, matn orqali ifodalangan fikrlarning ijobiy, salbiy yoki neytral ekanligini aniqlashga qaratilgan jarayondir. Bugungi kunda sentiment tahlil ko‘plab sohalarda, jumladan, marketing, siyosat, ijtimoiy tarmoqlardagi tahlil va mijozlar fikrlarini o‘rganishda keng qo‘llanilmoqda. Sentiment tahlilining asosiy maqsadi gaplarda ifodalangan his-tuyg‘ularni aniqlashdir. Bunda so‘z turkumlarining ahamiyati juda katta, chunki har bir so‘z turkumi o‘zining grammatik va semantik xususiyatlari bilan his-tuyg‘ularni ifodalashda turlicha rol o‘ynaydi. Ushbu maqolada sentiment tahlilda so‘z turkumlarining ahamiyati ko‘rib chiqiladi. Insonlarning ko‘p qismi ijtimoiy tarmoqdan doimiy foydalanganligi sababli, onlayn sharhlarda ko‘pchilik insonlar o‘zining shaxsiy fikrini erkin bildirishmoqda**.** O‘n minglab fikrlarni tahlil qilish va foydalanuvchilarning turli ruhiy holatda yozilgan izohlarini samarali, tez va sifatli avtomatik ravishda tahlil qilish turli sohalarda jamoatchilik fikrini o‘rganish va shu asosda tez qaror qabul qilishga yordam beradi.

So‘nggi o‘n yil ichida Internetning jadal rivojlanishi juda ko‘p miqdordagi onlayn sharhlarni keltirib chiqardi va izohlarda foydalanuvchilarning kayfiyatlarini tahlil qilish ijtimoiy barqarorlik hamda rivojlanish uchun katta ahamiyat kasb etdi[1]. Ushbu tendensiyaga javoban sentiment tahlil texnologiyasi paydo bo‘ldi. Sentiment tahlil, shuningdek, fikr tahlili deb ham ataladi, tabiiy tilni qayta ishlash (NLP)da muhim o‘rganish sohasi bo‘lib, u matndan hissiyot va turli xil qarashlarni avtomatik ravishda aniqlash va tahlil qilish uchun mo‘ljallangan. Sentiment tahlil sun’iy intellektni rivojlantirish uchun juda muhim hisoblanadi. Uning asosiy o‘rni – matnlardagi fikrlar va emotsiyalarning yo‘nalishini (ijobiy, salbiy yoki neytral) avtomatik tarzda tushunishdir[2].

NLPda sentiment tahlili quyidagicha talqin qilinadi:

- Inson fikrining emotsional tomonini o‘rganadi.
- Matn mazmuni va munosabatini chuqur tahlil qilish imkonini beradi.
- Mijozlar fikri, sharhlar, ijtimoiy media, yangiliklarni tahlil qilishda keng qo‘llaniladi.

Shuningdek, sentiment tahlili NLPning semantik tahlil, tematik tahlil, klassifikatsiya kabi boshqa yo‘nalishlari bilan birgalikda ishlaydi va matnni chuqurroq tahlil qilishga yordam beradi.

So‘z turkumlari – bu so‘zlarning grammatik xususiyatlariga qarab guruhlanishi. Har bir so‘z turkumi matndagi ifodalangan ma’no va his-tuyg‘uni aniqroq tushunishga yordam beradi. Sentiment tahlilida so‘z turkumlarini to‘g‘ri

aniqlash va ularni inobatga olish modelning aniqligini sezilarli darajada oshirishi mumkin.

Mustaqil so‘z turkumlari:

– **Otlar:** Obyektlar, insonlar, joylar yoki tushunchalarni ifodalaydi. Sentiment tahlilda otlar hissiyot va tushunchalarni aniqlashda yordam beradi. Ayniqsa mavhum tushunchalarni bildiruvchi otlar hissiy jarayonlarni aks ettiradi. Masalan: *“quvonch”, “g‘azab”, “mehr”* kabi otlar.

– **Sifatlar** ko‘pincha biror narsa yoki holatning sifatini, ya’ni uning qanday ekanligini ifoda etadi. Sentiment tahlilda sifatlar his-tuyg‘uni bevosita ifodalaydi. Masalan, *“yaxshi”, “ajoyib”, “yomon”, “dahshatli”, “qo‘rqinchli”* kabi sifatlar yordamida matnda qanday hissiyot mavjudligi aniqlanadi.

– **Fe’llar** matnda ifodalanayotgan harakatlar yoki holatlarni bildiradi. Sentiment tahlilda fe’l harakat yoki holat orqali his-tuyg‘uni ifodalaydi. Hissiyotlarni aks ettiruvchi fe’llar, masalan, *“sevmoq”, “nafratlanmoq”, “xursand bo‘lmoq”, “yoqtirmoq”, “baxtiyor bo‘lmoq”, “qo‘rqmoq”* kabi fe’llar sentiment tahlilda muhim rol o‘ynaydi.

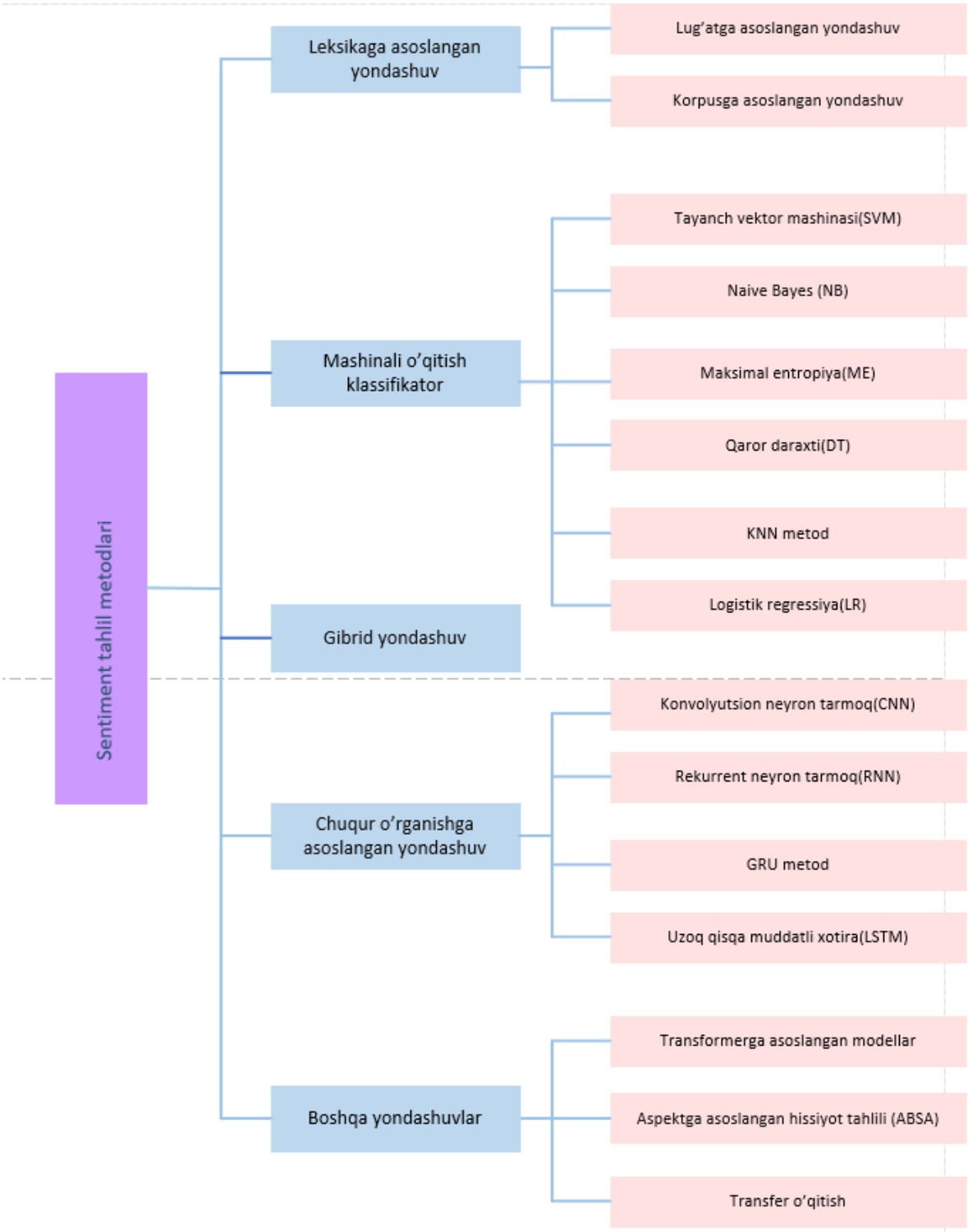
– **Ravish** harakat va holatning belgisini bildiradi. Sentiment tahlilda ravishlar sifatlar yoki fe’llarga nisbatan qo‘shimcha ma’lumot beradi va hissiyotlar darajasini kuchaytirishi, zaiflashtirishi, aniqlik berishi yoki yumshatishi mumkin. Masalan, *“juda”, “nihoyatda”, “ozgina”, “mutlaqo”* kabi so‘zlar hissiyot darajasini aniqlashda qo‘shimcha ma’lumot beradi.

Bu so‘z turkumlari ko‘proq matnning tuzilishini belgilaydi, lekin ular hissiyot ifodalashda kamroq rol o‘ynaydi. Olmosh va son mustaqil so‘z turkumlari sentiment tahlilda matn yoki gapning ma’nosiga ta’sir ko‘rsatmaydi.

Sentiment tahlil metodlari

Sentiment tahlilida so‘z turkumlarini inobatga olish modelning aniqligini oshiradi, chunki har bir turkum o‘zining hissiy ifodalariga ega va matnni ko‘proq tushunish imkonini beradi. **Sifatlar va fe’llar tahlilda asosiy rolni o‘ynaydi**, ammo barcha so‘z turkumlari birgalikda ishlaganda yanada mukammal natija berishi mumkin.

Sentiment tahlil quyidagi 1-rasmda keltirilgan metodlar orqali amalga oshiriladi:



1-rasm. Sentiment tahlil metdolari tasnifi

Ushbu maqolada sentiment tahlil metodlaridan **Leksikaga (leksemaga) asoslangan yondashuv** yoritiladi. Leksikaga asoslangan yondashuv odatda ikkita usulga bo‘linadi: lug‘atga asoslangan va korpusga asoslangan usullar. Quyidagi

1-jadvalda leksikaga asoslangan yondashuvning afzalliklari va kamchiliklari keltirilgan.

1-jadval. Leksikaga asoslangan yondashuvning afzalliklari va kamchiliklari

Yondashuv	Afzalligi	Kamchiligi
Lug‘atga asoslangan	**O‘**qitilgan ma’lumotlarga ehtiyoji y**o‘**q. Muayyan soha lu**g‘**ati uchun ijobiy natijalar beradi. Lu**g‘**atdan s**o‘**z ma’nolarini tez topish mumkin.	Fikrlar muayyan mazmunga y**o‘**naltirilgan b**o‘**ladi. Lu**g‘**atda hali mavjud b**o‘**lmagan soha atamalarini aniqlay olmaydi.
Korpusga asoslangan	Mazmunni yaxshi aniqlay oladi. Soha farqlanganda samarali natijalar beradi.	Lu**g‘**atning kengligi sababli natijalar turlicha b**o‘**ladi. Har bir matn terminini t**o‘**liq qamrab oluvchi matnlar t**o‘**plamini ta’minlash qiyinligi tufayli alohida q**o‘**llab b**o‘**lmaydi.

Taboada va boshq. (2011)[3] matndan hissiyotni olish uchun *Semantic Orientation CALculator (SO-CAL)* deb nomlangan leksikaga asoslangan usulni taqdim etdi. Ushbu usul to‘plangan tokenlarga ballar tayinlaydigan hissiyot leksikasidan foydalanishi kerakligini ko‘rsatadi[3].

So‘z turkumlarini sentiment tahlildagi ahamiyati

Turli sohalarda bir xil so‘z turli xil his-tuyg‘ularga ega bo‘lishi mumkin. Masalan, quyidagi ikkita gap berilgan bo‘lsin:

1. *“Havo keskin **isishi** oqibatida Antarktidada muzlar erishi kuzatilmoqda.”*
2. *“Bahor kelib kunlar **isishi** bilan daraxtlar qiyg‘os gulladi.”*

Birinchi gapdagi *“isishi”* atamasi salbiy, chunki isish haddan tashqari kuchli bo‘lgani uchun salbiy hodisaga olib keladi: muzlar eriyapti, ekologik muammo paydo bo‘lyapti. Shuning uchun *“isishi”* salbiy tarzda qabul qilinadi; ammo, ikkinchi gapda ijobiy, chunki, bu yerda *“isish”* tabiiy jarayon sifatida keltirilgan. Natijada tabiat uyg‘onmoqda, gullar ochilyapti. Lekin bu so‘zning salbiy yoki ijobiy bo‘lib ko‘rinishi gapdagi kontekstga (ya’ni atrofidagi so‘zlar va vaziyatga) bog‘liq.

So‘z turkumlari sentiment tahlilda muhim rol o‘ynaydi. Chunki, so‘z turkumlarining sentiment baholariga ta’sirini to‘g‘ri aniqlash, tilni yanada chuqurroq

tushunishga yordam beradi. Ushbu tahlil quyidagi sohalarda NLPning rivojlanishiga hissa qo‘shadi:

– **Matnning umumiy mazmuni va asosiy mavzusini aniqlashda** so‘z turkumlarining sentimentga ta’sirini hisobga olish muhimdir. Chunki matnning faqat mavzusini emas, balki unga bo‘lgan munosabatni ham tahlil qilish aniq va chuqur tematik tahlil qilish imkonini beradi.

– **Avtomatik tarjima jarayonida** kinoya va his-tuyg‘ularni to‘g‘ri anglash va tarjima qilish uchun so‘z turkumlarining sentimentdagi o‘rni muhim. Bu hodisalarni inobatga olish avtomatik tarjimalarda matn ma’nosining buzilmasligini va tabiiyligini ta’minlaydi.

– **Dialog tizimlari va chat-botlar foydalanuvchi xabarlaridagi** his-tuyg‘ularni (salbiy, ijobiy, neytral) tezda aniqlash va unga mos ravishda javob berish uchun so‘z turkumlari asosidagi sentiment tahlili yordam beradi. Bu esa foydalanuvchi tajribasini sezilarli darajada yaxshilaydi.

So‘z turkumlarining sentiment tahlilidagi ta’sirini o‘rganish, NLP tizimlarining aniqligini va sifatini yaxshilaydi, chunki bu tizimlarga matnlardagi his-tuyg‘ularni chuqurroq va to‘g‘ri aniqlash imkonini beradi. Bu natijalar NLP tizimlarini yanada samarali qiladi, shu bilan birga, ko‘plab sohalarda, masalan, brend monitoringi, mijozlar sharhlarini tahlil qilish va ijtimoiy tarmoqlarda ishlashda katta ahamiyatga ega.

Biz topishimiz mumkin bo‘lgan his-tuyg‘u turlari ijobiy, neytral va salbiyni o‘z ichiga oladi va keyinchalik hayrat, ishonch, kutish, g‘azab, qo‘rquv, qayg‘u, jirkanish, quvonch va boshqalarga bo‘linishi mumkin[4]. Til nuqtai nazaridan his-tuyg‘ularni tahlil qilish tadqiqotlari xitoy[5], ingliz[6], arab[7], nemis[8], portugal[9], rus[10], ispan[11] va boshqalar kabi turli xil tabiiy tillardan foydalanishi mumkin. Sentiment tahlil internet-ijtimoiy platformalar, elektron tijorat, moliya, sog‘liqni saqlash va davlat sektorlarini rivojlantirishda muhim rol o‘ynaydi.

O‘zbek tilida sentiment tahlilda so‘z turkumlarining rolini aniqlash uchun quyidagi manbalardan foydalanib ma’lumotlar bazamizni shakllantirdik:

1. Badiiy asarlarda (masalan, romanlar, she’rlar, dramalar) so‘zlar nafaqat ma’no, balki hissiyotlarni ham ifodalaydi. Fe’l, sifat va otlar orqali muallifning his-tuyg‘ulari, xarakterlar va voqealar orasidagi munosabatlar ochib beriladi, ayniqsa sifat so‘z turkumi hissiy-bo‘yoqdorlikni oshirib beradi. Bu so‘z turkumlari yordamida personajlarning ruhiy holati va voqea rivoji ko‘rsatiladi.
2. Internet saytlaridagi matnlar, maqolalar yoki sharhlar ma’lum bir mavzu yoki muhokama atrofida bo‘ladi. Bu manbalarda ijobiy yoki salbiy fikrlar, savollar, yoki bahslar bo‘ladi. Masalan, ijtimoiy, siyosiy yoki iqtisodiy mavzular haqida yozilgan maqolalarda so‘z turkumlari orqali fikrlarning qanchalik qat’iy yoki ehtimoliy ekanligi ko‘rsatiladi. Fe’llar (“saqlash”, “o‘zgartirish”), sifatlar

(“muammoli”, “mustahkam”) va otlar (“tijorat”, “davlat”) orqali mavzuga nisbatan hissiy munosabatlar ifodalanadi.

3. Ijtimoiy tarmoqlarda ko‘p holatlarda tezkor va hissiy javoblar, sharhlar mavjud. Bu matnlarda asosan sifatlar, shuningdek, fe’llar va yordamchi so‘z turkumlari ko‘p uchraydi. Bu turkumlar ijtimoiy tarmoq foydalanuvchilarining his-tuyg‘ulari va munosabatlarini aniq ko‘rsatadi. Masalan, “ajoyib”, “yomon”, “qiziqarli” kabi sifatlar sentimentni tahlil qilishda muhim rol o‘ynaydi.

Har bir manba o‘zining xususiyatlariga ega va sentiment tahlilini yanada samarali va aniqroq qilishda yordam beradi.

Sentiment tahlili uchun bir nechta yondashuvlar mavjud, ular asosan mashinali o‘qitish modellariga va qoidaga asoslangan yondashuvlarga asoslanadi.

- **Mashinali o‘qitish (Machine Learning) algoritmlari:**

Mashinali o‘qitish usullari, sentimentni aniqlashda avvalgi ma’lumotlarni o‘rganish asosida ishlaydi. Katta hajmdagi sentiment tahlil qilingan matnlar asosida o‘rganib mustaqil ravishda sentiment tahlil qiladi.

- **Qoidaga asoslangan yondashuvlar:**

Qoidaga asoslangan yondashuvlar, ko‘pincha sentiment lug‘atlari va manbalardan olingan so‘zlarning hissiy ma’nolari asosida ishlaydi.

- So‘z turkumlarining ta’sirini o‘lchash usullari:

So‘z turkumlarining sentiment tahliliga ta’siri statistik tahlil, leksemaga asoslangan yondashuvlar va so‘zlarning sentiment baholari bilan lug‘at yordamida o‘lchanadi.

Statistik tahlil:

Statistik tahlil orqali, har bir so‘z turkumi (masalan, ot, sifat, fe’l) matndagi ijobiy yoki salbiy ta’sirni qanday ifodalashini tahlil qilish mumkin. Matndagi so‘z turkumlarining soni, joylashuvi va ulushi, sentimentga qanday ta’sir qilayotganini aniqlash uchun o‘rganiladi. Misol uchun:

- Odatda ot odamlar, joylar, narsalar yoki g‘oyalar bilan bog‘liq bo‘lgan so‘zlar. Bu so‘zlar ko‘pincha neytral yoki salbiy, ijobiy hissiyotlarni ifodalaydi.
- Sifatlar (masalan, yaxshi, yomon, go‘zal) ko‘pincha matnning hissiy ohangini belgilaydi. Ular sentimentni kuchaytiruvchi rol o‘ynaydi.
- Fe’llar matnda harakat yoki holatni ifodalab, sentimentni dinamizatsiya qilishda muhim ahamiyatga ega bo‘lishi mumkin.
- Sentiment tahlilda ravishlar sifatlar yoki fe’llarga nisbatan qo‘shimcha ma’lumot beradi va hissiyotlar darajasini kuchaytirishi, zaiflashtirishi, aniqlik berishi yoki yumshatishi mumkin.

Leksemaga asoslangan yondashuv:

Lug‘at asosida, har bir so‘zning o‘ziga xos sentiment bahosi mavjud. Bu baholar ijobiy, salbiy yoki neytral bo‘lishi mumkin. Matndagi har bir so‘z shu baho asosida tahlil qilinadi va “-5”dan “5”gacha bo‘lgan oraliqda har bir so‘zga sentiment baho qo‘yiladi. Eng yuqori salbiy ma’no anglatgan so‘zga “-5” baho, neytral ma’nodagi so‘zga “0” baho, eng yuqori ijobiy so‘zga “5” sentiment baho qo‘yiladi. Keyin, barcha so‘zlarning baholari yig‘ilib, matnning umumiy sentimenti hisoblanadi.

2-jadval. Matnni sentiment tahlil qilishda so‘z turkumlari

So‘z turkumlari	So‘z	(Ijobiy yoki salbiy)	Sentiment baho	Gap	Sentiment baho
Sifat	qiziqarli	ijobiy	3	Yozda institut jamoasi uchun turli shaharlarda mazmunli va ***qiziqarli*** sayohatlar o‘tkazilishi rejalashtirilgan.	4
	ayanchli	salbiy	-4	Dunyoda hamma narsadan mosuvo bo‘lgan bu ayolning hozirgi ahvoli juda ***ayanchli*** edi.	-5
Fe’l	sevindi	ijobiy	5	Dobshalimning yuborgan insoni unga bu xabarni yetkazganida u nihoyatda ***sevindi***.	5
	achindi	salbiy	-3	"To‘g‘ri, buning gapi ham to‘g‘ri", deb o‘yladi Akobir va o‘z ota-onasining qabri qayerdaligini hatto bilmasligiga ***achindi***.	-4
Ravish	nihoyatda	ijobiy	5	Kutubxonaning kitob va axborot-resurs fondi ***nihoyatda*** boy hamda xilma-xildir.	4
	biroz	salbiy	-2	Tabiiyki, mavzularni o‘zlashtirishda ham ***biroz*** qiyinchilik tug‘diradi.	-1
Ot	baxt	ijobiy	5	Atigi 4 yillik pedagogik tajribaga ega bo‘lgan 26 yoshli muallima uchun bu katta yutuq, yuksak ***baxt***.	5
	g‘am	salbiy	-4	Balki ular ichida boshiga biror tashvish yoki ***g‘am*** tushgan inson ham bordir.	-3

Sentiment tahlilda eng muhim so‘z turkumi fe’l so‘z turkumidir. Chunki gapning markazini ko‘p hollarda fe’l tashkil etadi. Sentiment tahlilda ham gapda asosiy ifodalangan ma’noni fe’l so‘z turkumi ochib beradi. Masalan: “Menga bozorga borish yoqmadi.”. Ushbu gapning sentiment bahosi salbiy. Chunki *“yoqmadi”* fe’lidan salbiylik yaqqol ko‘rinib turibdi. Sifat so‘z turkumi esa gapda ifodalangan his-tuyg‘uni bevosita bildiradi. Masalan: “Ushbu film ajoyib.”. Ushbu gapning sentiment bahosi ijobiyligi *“ajoyib”* so‘zidan yaqqol sezilib turibdi. Ravish so‘z turkumiga mansub so‘zlar fe’lda ifodalangan harakatni zaiflashtirish, kuchaytirish yoki aniqligini oshirish bilan sentiment bahoni oshiradi yoki pasaytiradi. Masalan: “Ular orasida Chexiyadagi Karlov universiteti meni qiziqtirdi.” va “Ular orasida Chexiyadagi Karlov universiteti meni juda qiziqtirdi.” Berilgan ikki gapda sentiment baho ikki xil. Ammo gaplar faqat juda kuchaytiruvchi

ravishi bilan farqlangan. Birinchi gapda ijobiylik bor ammo past. Ya’ni sentiment bahosini “+3” “ijobiy” deb belgilasak bo‘ladi. Ikkinchi gapda esa sentiment baho aniq va kuchaytirilgan. Gapda “juda” kuchaytiruvchi ravishi “qiziqtirdi” fe’lining sentiment bahosini oshirib berganligi tufayli sentiment bahosi “+5” “ijobiy”ga ko‘tariladi. Har bir so‘z turkumi sentiment tahlilda o‘z o‘rniga ega ekanligini ko‘rdik.

Biroq, matnda ifodalangan **kinoya, jargon, sleng** va boshqa tillardan kirib kelgan hayratlanish, haqoratli so‘zlar kabilarga sentiment baho berish muammoligicha qolmoqda. Bu his-tuyg‘ular matnda aniq ifodalanmaganligi sababli, ular yashirin his-tuyg‘ular deb ataladi. Insonlar bu his-tuyg‘ularni ikkita ko‘rsatma orqali qabul qilishadi: ovozli aloqa va matnli kontekst.

Ba’zi so‘zlar tabiiy ravishda obyektivdir, ammo ma’lum bir muhitda yoki kontekstda ishlatilganda subyektivlikka tayanishi mumkin va shuning uchun hissiy jihatdan zaryadlangan bo‘lishi mumkin[12]. Masalan, *“uzoq”* so‘zi hissiyotni ifodalamaydigan neytral sifatdir. Ba’zi hollarda ijobiy (masalan, *“uzoq umr”*) yoki salbiy (masalan, “uzoq navbat” kabi) bo‘lishi mumkin. Ko‘pincha, kontekstda so‘zning qurshovida kelgan so‘zlar bilan sentiment baho aniqlanadi.

Xulosa

Ushbu maqolada so‘z turkumlarining sentiment tahlilidagi o‘rni tahlil qilinib, ularning klassifikatsiyada muhim omil bo‘lishi aniqlangan. Ayniqsa, sifat, fe’l va ravishlar hissiy ohangni belgilashda kuchli semantik ulushga ega bo‘lib, model natijalarini yaxshilashini ta’kidlash lozim. So‘z turkumlarini hisobga olgan holda ishlab chiqilgan sentiment tahlil modellari kelajakda aniqligi yuqori tizimlar yaratish uchun asos bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar:

1. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. *Knowledge-Based Systems*, *226*, 107134.
1. Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. *Mining text data*, 415-463.
2. Mitra, A., & Mohanty, S. (2020). Sentiment analysis using machine learning approaches. *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS*, *2*, 63-68.
3. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational linguistics*, *37*(2), 267-307.
4. Bose, R., Dey, R. K., Roy, S., & Sarddar, D. (2020). Sentiment analysis on online product reviews. In *Information and Communication Technology for*

Sustainable Development: Proceedings of ICT4SD 2018 (pp. 559-569). Springer Singapore.

5. Peng, H., Cambria, E., & Hussain, A. (2017). A review of sentiment analysis research in Chinese language. *Cognitive Computation*, *9*, 423-435.

6. Rodríguez-Ibánez, M., Casánez-Ventura, A., Castejón-Mateos, F., & Cuenca-Jiménez, P. M. (2023). A review on sentiment analysis from social media platforms. *Expert Systems with Applications*, *223*, 119862.

7. Oueslati, O., Cambria, E., HajHmida, M. B., & Ounelli, H. (2020). A review of sentiment analysis research in Arabic language. *Future Generation Computer Systems*, *112*, 408-430.

8. Remus, R., Quasthoff, U., & Heyer, G. (2010, May). SentiWS-A Publicly Available German-language Resource for Sentiment Analysis. In *LREC*.

9. Pereira, D. A. (2021). A survey of sentiment analysis in the Portuguese language. *Artificial Intelligence Review*, *54*(2), 1087-1115.

10. Smetanin, S. (2020). The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. *IEEE Access*, *8*, 110693-110719.

11. Osorio Angel, S., Peña Pérez Negrón, A., & Espinoza-Valdez, A. (2021). Systematic literature review of sentiment analysis in the Spanish language. *Data Technologies and Applications*, *55*(4), 461-479.

12. Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, *64*(1), 17-27.

13. Elov B. and Xudayberganov, N. (2024). O‘zbek tili korpusi matnlarini pos teglash usullari. Computer Linguistics: problems, solutions, prospects, 1(1).

14. Xusainova Z. (2023). O‘zbek tili milliy korpusi qidiruv tizimini optimallashtirishda lemmatizatsiyadan foydalanish. Uzbekistan: Language and Culture, 2(2).

15. Elov B., Alayev R. and Abdullayev A. (2024). Tabiiy tilning statistik modellari. digital transformation and artificial intelligence, 2(6), 178-189.

TABIIY TILGA ISHLOV BERISH SOHASIDA SO‘Z MA’NOSINI ANIQLASH

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Hasanova Gulshoda Ikrom qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
ToshDO‘TAU

Annotatsiya. Tabiiy tilga ishlov berish sohasi matnlarni avtomatik tahlil qilish masalasini yechish bilan shug’ullanadi. Avtomatik tahlil turlaridan biri so’z semantic tahlil bo’lib muhim ahamiyatga ega. Avtomatik semantik tahlil amalga oshirishda uni bir nechta elementlarga ajratgan holda amalga oshiriladi. Ushbu aniqlash, matnni tahlil qilish va so‘zlar orasidagi bog‘lanishni tushunish asosiy vazifasi sanaladi.

Annotation. The field of natural language processing deals with solving the problem of automatic text analysis. One of the types of automatic analysis is word semantic analysis, which is of great importance. When performing automatic semantic analysis, it is carried out by dividing it into several elements. This identification, text analysis, and understanding of the relationships between words are considered the main tasks.

Аннотация. Область обработки естественного языка занимается решением задачи автоматического анализа текстов. Одним из видов автоматического анализа является семантический анализ слов, который имеет важное значение. При выполнении автоматического семантического анализа он осуществляется с разделением на несколько элементов. Это определение, анализ текста и понимание связей между словами считаются основными задачами.

Kalit so‘zlar: *So‘z ma’nosini aniqlash, NLP, noaniqlik, leksema, leksik noainqlik, struktural noaniqlik, referensial noainiqlik*

Kirish

Kompyuter lingvistikasi sohasida matnlarni avtomatik semantik tahlil qilish masalasining yechilishi kunning dolzarb masalalaridan biri. Semantik tahlil matn yoki nutqning ma’nosini, uning tarkibini va mazmunini o‘rganadi. Semantik tahlil orqali so‘zlar, iboralar, jumlalar yoki butun matnning semantik (ya'ni ma’noga oid) tuzilishi tahlil qilinadi. Barcha tabiiy tillarida noaniq so‘zlar mavjud, bu so‘zlar o‘z ma’nosini kontekstdan kelib chiqqan holda o‘zgartiradi. Biror so‘zning to‘g‘ri

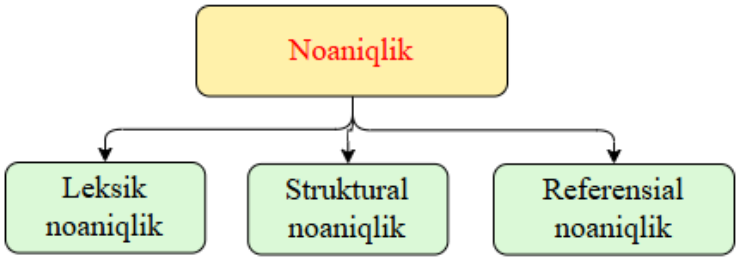
ma’nosini kontekstga qarab aniqlash jarayoni So‘z Ma’nosini Aniqlash (WSD) deb ataladi. Bu jarayon, bir so‘z jumlada ishlatilganda uning maqsad qilingan ma’nosini avtomatik tarzda aniqlaydi. Noaniq so‘zlarning ba’zi misollari quyidagilar:

1-jadva: Noaniq so‘zlar va ularning ma’nolari

t/r	So‘zlar	Ma’nolari
1	ich	1.suyuqlikni ichga yutmoq 2.ichkaridagi joy
2	yara	1. chiqiq, jarohat 2. kerak bo‘lib qolmoq
3	qazi	1. otning to‘sh va peshnop go‘shti shulardan yo‘g‘on ichakka tiqib tayyorlanadigan mahsulot 2. yerni o‘yib, chuqur hosil qilmoq, kovlab olmoq

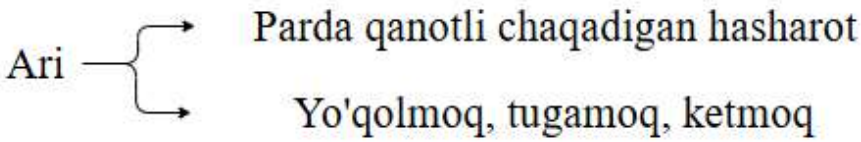
So‘z ma’nosini aniqlash (WSD) turli sohalarda, masalan, mashina tarjimasi (Machine translation, MT), axborotni qayta ishlash (Information reworking, IR), nutqni tanib olish (Automatic Speech Recognition, ASR), matnni qayta ishlash kabi ko‘plab sohalarda qo‘llaniladi. WSD boshqa NPL axborot tizimlarining natijasining aniqligini oshirishda juda muhim rol o‘ynaydi. U so‘zning ma’nosini kontekstga nisbatan saqlaydi.

Noaniqlik (aniqmaslik). Agar bir so‘z, ibora yoki jumla bir nechta ma’noga ega bo‘lsa, u noaniqlik deyiladi. Quyida turli noaniqlik turlari keltirilgan:



5-rasm: Noaniqliklarning turlari

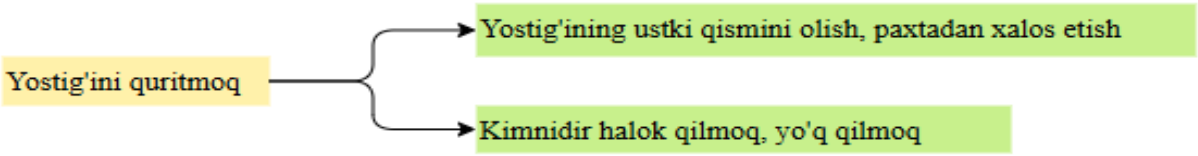
Leksik noaniqlik. Bir nechta ma’noga ega bo‘lgan so‘zlar omonim yoki polisemantik bo‘lishi mumkin. Agar bir so‘zning ikki ma’nosi butunlay farqli bo‘lsa, ammo yozilishi bir xil bo‘lsa, ular omonimlar deb ataladi. Masalan, “ari” so‘zining ikki ma’nosi quyidagicha ishlatilishi mumkin:



2-rasm: Ari so’zning ma’nolari

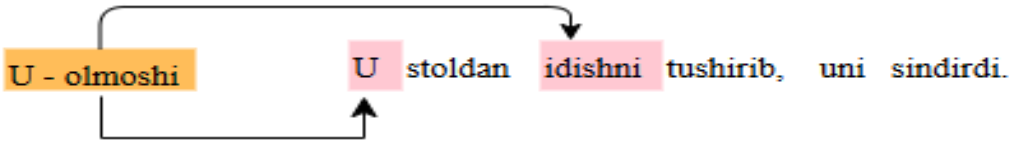
Bir so‘zning ma’nosi bir xil asosiy ma’noning turli xil tuslarini ifodalasa, ya’ni u turli lekin bir-biriga bog‘liq ma’nolarga ega bo‘lsa, u polisemantik deb ataladi. Masalan, “moya” so‘zi polisemantik hisoblanadi, chunki uning ikki ma’nosi – urg‘ochi tuya va hamirturish, achitqi – bir-biriga aloqador bo‘lgan ma’nolardir.

Struktural (Sintaktik) noaniqlik. Agar noaniqlik jumla yoki ibora ko’rinishida bo‘lsa, bu struktural noaniqlik deb ataladi. Bu holat, biror ibora yoki jumla bir nechta asosiy strukturalarga ega bo‘lganda yuzaga keladi. Masalan:



3-rasm: Struktural noaniqlikka misol

Referensial noaniqlik. Agar gapdagi olmoshlar ma’lum bir so‘zlarga ishora qilsa, lekin qaysi so‘zga ishora qilayotganini aniqlash qiyin bo‘lsa, bu referensial noaniqlik deb ataladi. Misol uchun:



4-rasm: Referensial noaniqlikka misol

Bu yerda noaniqlik shundaki, “u” olmoshi qaysi otga ishora qilmoqda: *stolga* yoki *idishga*.

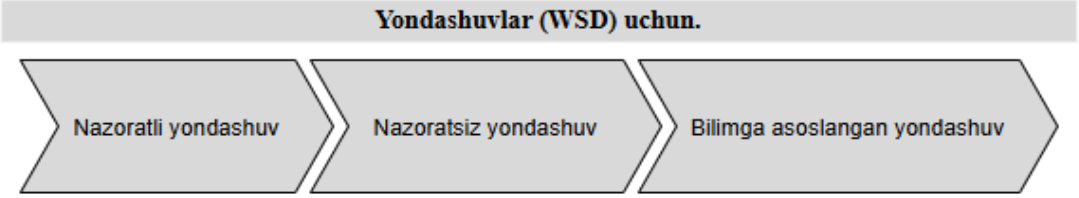
Bunday ishoralar ***anaforik ishoralar*** yoki ***anafora*** deb ataladi.

Ko‘p so‘zli konstruktsiyalar: Ko‘p so‘zli konstruktsiyalar, masalan, idiomalar va frazali fe’llar tufayli yuzaga kelgan noaniqlik. Misol uchun:

t/r	Matn	Ma’nosi
1	Qishloq bug‘doy ekdi.	Qishloq – butun odamlar.
2	Universitet bugun e’lon chiqardi.	Universitet – rahbariyat yoki matbuot xizmati.

5-rasm: Ko'p so‘zli konstruksiyalarga misol

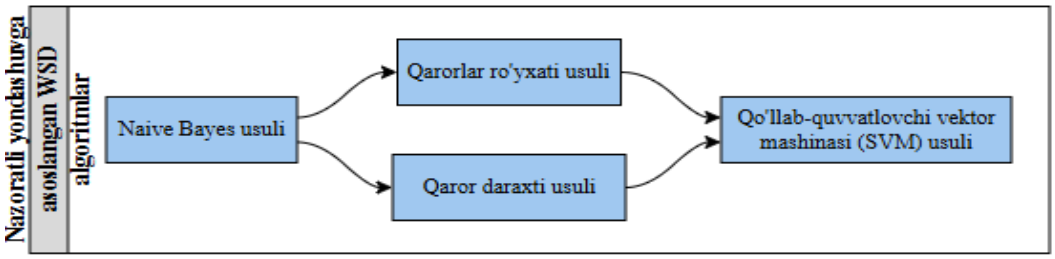
Idiomning haqiqiy ma’nosini uning tarkibiy so‘zlaridan aniqlash qiyin. Bu noaniqliklar semantik tahlilning asosiy masalasidir. Jahon tajribasiga ko’ra noaniqliklarni yechish so’z ma’nosini aniqlash (WSD- Word Sense Disambiguation) deb nomlanadi. Noaniqliklarni aniqlashda turli yondashuvlardan foyalanish mumkin. ***WSD uchun yondashuvlar***. So‘z ma’nosini aniqlash (WSD) yondashuvlari uchta asosiy kategoriya bo‘yicha tasniflanadi:



6-rasm: WSD yondashuvlar turlari

Nazoratli yondashuv. Nazoratli yondashuv semantic teglangan ma’lumotlarga yoki korpusga asoslanadi. Nazoratli yondashuv, tasniflagichlarni qurish uchun o’qitilgan ma’no teglangan korpusga asoslanadi. Dastlab, tasniflagichni qurish uchun teglangan korpus talab qilinadi. Tasniflagich ko’p ma’noli so‘zning ma’nosini kontekstga asoslanib aniqlash uchun ishlatiladi. Ushbu yondashuvning asosiy kamchiligi — katta hajmdagi semantic teglangan korpusni talab qilishi.

Tasniflagichning aniqligi korpusga kiritilgan ma’lumotlar hajmi va xilma-xilligiga bog‘liq. Korpus kattaroq bo‘lsa, natijalar yaxshiroq bo‘ladi. Korpus yaratish o‘zi vaqt va mablag‘ jihatidan muammo bo‘lishi mumkin. Nazoratli yondashuvga asoslangan ba’zi muhim WSD algoritmlari quyidagilardir:



7-rasm: Nazoratli yondashuvga asoslangan WSD algoritmlar

Naïve Bayes usuli – omonimlarni farqlashda va guruhlarga ajratishda foydaniladi.

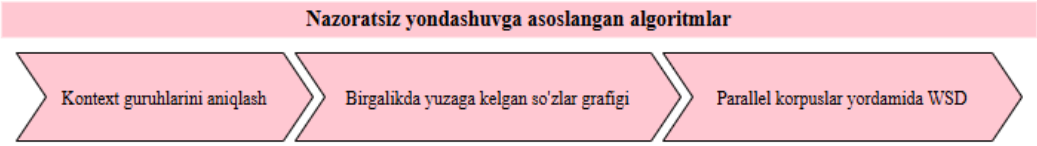
Qarorlar ro‘yxati usuli – so‘zning turli ma’nolarda kelgan matnlarni korpusga to‘playdi va har bir misolni tegishli ma’nosini belgilab oladi. Lekin katta hajmdagi ma’lumotlar va matn murakkabligida ishlamaydi.

Qarorlar daraxti usuli – bu usul so‘zning ma’nosini aniqlash uchun daraxt shaklida tuzilgan bo‘lib, unda har bir tugun matnni tekshiradi, shoxlar natija va burglar esa so‘zning ma’nosini bildiradi.

Qo‘llab-quvvatlovchi vector mashinasi (SVM) usuli – so‘zlarning ma’nolarini o‘rganadi, SVM ma’lumotlarini aniqlash uchun chegara (gipertekislik) hosil qiladi.

Nazoratsiz yondashuv. Nazoratsiz yondashuv teglanmagan korpusga asoslanadi. Bu yondashuv so‘zlarni guruhlashga asoslanadi. U noaniq so‘zning ma’nosi qo‘shni so‘zlardan hosil bo‘lishi mumkin deb hisoblaydi. Ushbu yondashuv

annotatsiya qilinmagan ma’lumotlarga asoslangani uchun, o‘qitish korpusida so‘zning ko‘plab ma’nolari mavjud bo‘lishi juda muhimdir. Ushbu yondashuvga asoslangan ba‘zi algoritmlarni keltiramiz:



8-rasm: Nazoratsiz yondashuvga asoslangan agoritmlar

Kontext guruhlarini aniqlash - yangi tillarni o‘rganadi, so‘zlarni qatnashgan konteksti bo‘yicha guruhlarga ajratadi. Katta hajmdagi matnlarni avtomatik tahlil qilishda aniq tahlil qiladi.

Birgalikda uchragan so‘zlar grafigi - so‘zlarning bir-biri bilan qanday bog‘liqligini grafik shaklda ko‘rsatadi. Matnning muhim so‘zlarini ajratadi.

Parallel korpuslar yordamida WSD - ikki yoki undan ortiq tillarni tarjima qilib, so‘zlarni matnga qarab ma’nolarini aniqlaydi.

Bilimga asoslangan yondashuv. Ushbu yondashuv so‘z ma’nosini aniqlash uchun bilim bazalarini ishlatishga asoslanadi. Foydalanilishi mumkin bo‘lgan bilim bazalari quyidagilar: WordNet, mashinaga o‘qitish mumkin bo‘lgan lug‘atlar, tezauruslar va hokazo. Ushbu yondashuv, ayniqsa, ma’lum bir soha uchun so‘z ma’nosini aniqlashda samarali bo‘ladi. Bilimga asoslangan yondashuvda foydalaniladigan mashhur algoritmlardan biri 1986-yilda Maykl Lesk tomonidan taklif qilingan va ***Lesk algoritmi*** deb nomlangan algoritmdir.

Lesk Algoritmi Maykl Lesk tomonidan 1986-yilda ishlab chiqilgan [5]. Bu polisemantik so‘zlarning ma’nolarini aniqlashda ishlatiladi. U Oxford Advanced Learner’s Dictionary of Current English (OALD) lug‘atidagi so‘z ta’riflarining o‘z va ko’chma ma’nolarini ajratib olgan. Hosil bo’lgan kontekst yordamida so‘z ma’nolarini aniqlagan. O‘zbek tilidagi omonimiyani aniqlashda ham ushbu algoritmda foydalanilgan[15]. ***Banerjee va Pedersen*** WordNetni bilim manbasini Lesk algoritmi uchun dataset sifatida ishlatganlar. Ishlab chiqilgan axborot tizimining umumiy aniqligi 32% deb baholangan [7].

Sinha M. va boshqalar **tomonidan** hind tili matnlari uchun avtomatik WSD tizimini ishlab chiqilgan. Ushbu tizimni ishlab chiqishda ***Hindi WordNeti***dan foydalanilgan [8]. Olimlar so’z ma’nolarini aniqlashda statistik usullardan foydalanganlar. Ishlab chiqilgan axborot tizim matn tarkibidagi faqat ot so’z turkumidagi noaniqlikdan bartaraf qiladi. Ular polisemantik so‘z joylashgan kontekstni WordNetdan olingan kontekstlar bilan solishtirganlar. Ular tizimni Hindistonning Markaziy Til institutidan (CIIL) taqdim etilgan hind korpuslari yordamida baholashgan. Ular algoritmning aniqligini 40% dan 70% gacha bo‘lgan oraliqda aniqlagan.

Shrestha N. va boshqalar[9] Lesk algoritmi asosida Nepali noaniq so‘zlarini noaniqlikdan bartaraf qilish uchun WSD moduli ishlab chiqishgan. Ular bilim resursi sifatida ***synset, gloss***, ***misollar*** va ***hypernym***dan foydalanib, so‘zlarni to‘plashgan. Har bir ma’no uchun maqsadli so‘zning ma’lumotlar bazasida faqat bitta misol mavjud edi. Har bir matndagi so‘zi, tuzilgan ma’lumotlar bazasida mavjud bo‘lgan so‘zlar to‘plamidagi har bir so‘z bilan taqqoslanadi.

Dhungana va Shakya [12] Lesk algoritmi yordamida Nepali tilidagi polisemantik so‘zlarni noaniqlikdan xoli qilish ustida ishlashgan. 348 ta so‘z (shu jumladan, 59 ta polisemantik so‘zni va matnning so‘zlarini o‘z ichiga olgan) bo‘yicha bajarilgan tajribalar va 201 ta Nepali jumlasini o‘z ichiga olgan test ma’lumotlari ularning tizimining aniqligini 88,05% deb ko‘rsatgan.

Richard Singh va K. Ghosh2013-yilda Manipuri tili uchun arxitektura taklif qilishgan. Tizim WSDni ikki bosqichda amalga oshiradi: o‘qitish bosqichi va test bosqichi. Manipuri tilidagi so‘z ma’nosini aniqlash tizimi quyidagi bosqichlardan iborat: (I) oldindan qayta ishlash, (II) xususiyatlarni tanlash va yaratish, (III) o‘qitish, (IV) test qilish va (V) ishlash samaradorligini baholash [14].

Haroon, R.P. Malayalam tilida avtomatik WSD ning birinchi urinishini taqdim etgan. U bilimga asoslangan yondashuvni ishlatgan [10]. ***Rakesh va Ravinder***Punjabi matn hujjatidan noaniqlikni olib tashlash uchun WSD algoritmini taklif qilishgan. Mualliflar WSD uchun ***Modified Lesk Algorithm***ni ishlatishgan [16].

Ayan Das va Sudeshna Sarkar Bengali tili uchun nazoratsiz grafikka asoslangan guruhlash yondashuviga asoslangan WSD tizimini ishlab chiqqanlar, bu tizim Bengali-Hindi mashina tarjimasi uchun qo‘llaniladi[10].

Xulosa. So‘z ma’nosini aniqlash (WSD) hisoblash lingvistikasi sohasida juda faol tadqiqot mavzusi bo‘lib kelmoqda. U tizimlar tomonidan yaratilgan natijalarni yanada aniqroq qilishda muhim rol o‘ynaydi. WSD tadqiqotlarini cheklagan omillardan biri — lingvistik resurslarning mavjud emasligidir. Hozirgi kunda ko‘pgina tadqiqotchilar asosan ***Lesk algoritmi***ga asoslangan ma’no aniqlashning bilimga asoslangan yondashuvini ishlatishmoqda. Korpusning yetishmasligi sababli, tadqiqotchilar noaniqlikni bartaraf etish uchun ***WordNet***dan foydalanmoqdalar. Shuningdek, bilimga asoslangan yondashuvlarning noaniq so‘zlarning ma’nolarini hal qilishda samarali ekanligi o‘rganilgan, ammo tabiiy tilni qayta ishlash vositalarining samaradorligini oshirish uchun leksik resurslarni ishlab chiqish ehtiyoji katta.

Foydalanilgan adabiyotlar:

1. Arindam Chatterjee, Salil Rajeev Joshi, Mitesh M. Khapra, Pushpak Bhattacharyya, “IndoWordNet va Word Sense Disambiguation uchun Asboblar

Tanishtiruvi”, http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/18_Tools%20for%20IndoWordNet%20&%20WSD.pdf

2. C. Fellbaum, “WordNet: An Electronic Lexical Database”, MIT Press, 1998. Lokesh Nandanwar, Kalyani Mamulkar, “Supervised, Semi-Supervised and Unsupervised WSD Approaches: An Overview”, International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064

3. Kuhoo Gupta, Manish Shrivastava, Smriti Singh va Pushpak Bhattacharyya, “Morphological Richness Offsets Resource Poverty- an Experience in Building a POS Tagger for Hindi”, COLING/ACL-2006, Sidney, Avstraliya, iyul, 2006. https://pdfs.semanticscholar.org/1daa/370563deb1e9f5772d6ada1edfb4ec4cd7b4.pdf

4. M. Lesk “Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone”. In Proceedings of the 5th annual international conference on Systems documentation, SIGDOC ’86, pages 24–26, New York, NY, USA. ACM.

5. Udaya Raj va boshqalar, “Word Sense Disambiguation using WSD specific Wordnet of polysemy Words”, https://arxiv.org/ftp/arxiv/papers/1409/1409.3512.pdf

6. Banerjee va Pederson, “An adapted Lesk algorithm for Word Sense Disambiguity using Word Net”, http://www.cs.cmu.edu/~banerjee/Publications/cicling2002.pdf

7. Sinha M. va boshqalar, “Unsupervised Graph-based Word Sense Disambiguation Using Measures of Word Semantic Similarity”, http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.3892&rep=rep1&type=pdf

8. Shrestha N va boshqalar, “Resources for Nepali WSD” in proceedings of IEEE conference on Natural Language Processing and Knowledge Engineering, 2008. NLP-KE '08. 2008.

9. Alok Ranjan Pal va Diganta Saha, “WORD SENSE DISAMBIGUATION: A SURVEY”, International Journal of Control Theory and Computer Modeling (IJCTCM) Vol.5, No.3, iyul 2015

10. Z.Z. Merhbene Laroussi, “Unsupervised system for Lexical Disambiguation of Arabic Language using a vote procedure,” IEEE-ICoAC, 2011

11. Raj Dhungana, Subarna Shakya va boshqalar, “Word Sense Disambiguation Using WSD Specific Wordnet of Polysemy Words”, Udaya International Journal on Natural Language Computing (IJNLC) Vol. 3, No.4, avgust 2014

12. Shallu va boshqalar, “A Survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages”, JOURNAL OF EMERGING TECHNOLOGIES IN WEB INTELLIGENCE, VOL. 5, NO. 4, Noyabr 2013

13. E. B. Boltayevich, A. X. Ilxomovna, P. M. Hakim Qizi and K. N. Uktambay O‘g'li, “Semantic Differentiation of Uzbek Homonyms Using the Lesk Algorithm,” 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 137-140, doi: 10.1109/UBMK59864.2023.10286666.

QAROR RO‘YXATI ALGORITMI YORDAMIDA SO‘Z MA’NOSINI ANIQLASH

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Hayotova Gulmira Hakim qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
hayotovagulmira208@gmail.com
ToshDO‘TAU

Annotatsiya: Ba’zi so‘zlar bir xil tarzda yozilib, turli kontekstda turlicha ma’nolarga ega. Insonlar intellektga ega bo‘lib, ular turli kontekstlarda ko‘p ma’noli so‘zlarning to‘g‘ri ma’nosini osonlik bilan tushunadilar. Buni atrofdagi so‘zlar, ya’ni lokal yoki global kontekst yordamida amalga oshiradilar. Lokal kontekst target so‘zga yaqin bir nechta so‘zlarni anglatadi. Global kontekst esa noaniq so‘zga yaqin bo‘lgan satrlar yoki paragraflarni ifodalaydi. So‘z ma’nosini aniqlashda kontekst muhim ahamiyat kasb etadi. Mashinali o‘qitish yondashuvi asosida so‘z ma’nosini aniqlash kontekstga asoslanadi. Ushbu maqolada qaror ro‘yxatlari algoritmi yordamida so‘z ma’nosini aniqlash haqida batafsil ma’lumot berilgan.

Annotation: Some words are spelled the same but have different meanings in different contexts. Humans are intelligent enough to easily understand the correct meaning of words with multiple meanings in different contexts. They do this by using the surrounding words, i.e., the local or global context. The local context refers to several words close to the target word. Global context refers to lines or paragraphs that are close to the ambiguous word. Context plays an important role in determining the meaning of a word. Machine learning approaches to determining the meaning of a word are based on context. This article provides detailed information about determining the meaning of a word using a decision list algorithm.

Аннотация.: Некоторые слова пишутся одинаково, но имеют разное значение в разных контекстах. Люди разумны и могут легко понимать правильное значение слов, имеющих несколько значений в разных контекстах. Они делают это, используя окружающие слова, то есть локальный или глобальный контекст. Локальный контекст относится к нескольким словам, близким к целевому слову. Глобальный контекст представляет собой строки или абзацы, которые находятся близко к неоднозначному слову. Контекст важен при определении значения слова. Подход машинного обучения к определению значения слова основан на контексте. В данной статье представлена подробная информация об определении значения слова с использованием алгоритма списка решений.

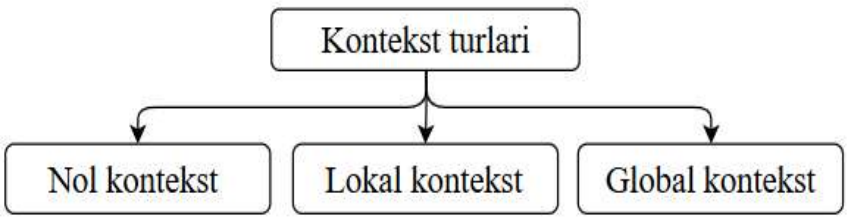
Kalit so‘zlar. *Mashinali o‘qitish yondashuvi, nazorat ostida o’qitish algoritmlar, so‘z ma’nosini aniqlash, kontekst, vazn.*

Kirish

Dunyoning deyarli barcha tabiiy tillarida turli kontekstlarda turlicha ma’noga ega bo‘lgan so‘zlar mavjud. Bunday so‘zlar polisemantik yoki ko‘p ma’noli so‘zlar deb ataladi, ular bir xil yoziladi, lekin turli kontekstlarda alohida ma’noga ega. Ya’ni ularning kontekstdagi ma’nosi noaniq. Noaniq so‘zning to‘g‘ri ma’nosini aniqlash va haqiqiy ma’noni belgilash jarayoni so‘z ma’nosini aniqlash (Word Sense Disambiguation, WSD) deb ataladi. Bu tabiiy tilga ishlov berish (NLP) sohasi uchun muhim yondashuvdir. So‘z ma’nosini aniqlash mashina tarjimasi (MT), semantik xaritalash (SM), ontologiyalarni o‘rganish (OL), semantik annotatsiya (SA) kabi jarayonlarda muhim ahamiyatga ega.

Insonlar intellektga ega bo‘lib, ular turli kontekstlarda ko‘p ma’noli so‘zlarning to‘g‘ri ma’nosini osonlik bilan tushunadilar. Buni atrofdagi so‘zlar, ya’ni lokal yoki global kontekst yordamida amalga oshiradilar. Lekin bu jarayonni tizim orqali amalga oshirilsa, tizimga ba’zi qoida va yo‘riqnomalarni kiritish lozim. Shundan so‘ng tizim kiruvchi kontekstdagi noaniq so‘zning to‘g‘ri ma’nosini aniqlay oladi. So‘zning ma’nosi – bu til jamoasi tomonidan so‘z uchun qabul qilingan ma’no. Noaniq so‘zning to‘g‘ri ma’nosini topish uchun lug‘atga murojaat qilish mumkin. Ya’ni, noaniq so‘zning ma’nosini avtomatik aniqlash uchun resurslar zarur. Masalan ingliz tilida aniqmasliklarni bartaraf etish uchun WORDNET manba sifatida xizmat qiladi [1,2]. O‘zbek tilining WORDNET tizimi ishga tushirilmaganligi sababli, boshqa manba hosil qilishga to‘g‘ri keladi. Garchi o‘zbek tilining WordNet tizimi ishga tushurilmagan bo‘lsada, noaniqliklarni aniqlash bo‘yicha qator izlanish olib borilmoda [7-9]. Ya’ni aniqmasliklarni turli ma’nolarini ifodalovchi gaplar to‘plamini jamlash zarur.

Gaplar shaxs tomonidan ifodalanganligi sababli, faqatgina gapga e’tibor beriladi, his-tuyg‘ularga emas. Inson omili yordamida gaplardagi his-tuyg‘uni aniqlash oson, ammo Mashinali o‘qitish yondashuvlari yordamida faqatgina gapga ishlov berish mumkin va ko‘p ma’noli so‘z uchrasa, uning atrofidagi so‘zlar tahlil qilinadi. Ko‘p ma’noli so‘zning birikuvchilari esa kontekstning o‘lchamiga bog‘liq. Kontekstlar quyidagicha bo‘lishi mumkin:



1-rasm: Kontekstlarning turlari

Nol kontekst – so‘zning o‘zi.

Lokal kontekst (Mahalliy) – ibora va qo‘shma jumla.

Global kontekst – gap, muloqot, mavzu va soha.

Kontekstda noaniq so‘zning to‘g‘ri ma’nosini aniqlash murakkab vazifadir. Kontekst *mahalliy* yoki *global* bo‘lishi mumkin. Mahalliy kontekst noaniq so‘z atrofdagi bir nechta so‘zlarni qamrab oladi. Global kontekst esa bir jumladan ko‘proq so‘zlar demakdir [1].

Kontekst turlari so‘z ma’nosini aniqlashda muhim ahamiyatga ega. So‘z ma’nosini aniqlashda turli yondashuvlardan foydalanish mumkin.

Asosan, ikki xil so‘z ma’nosini aniqlash yondashuvi mavjud:

1) Mashinali o‘qitish yondashuvi.

2) Lug‘atga asoslangan yondashuv.

Mashinali o‘qitish yondashuvi algoritmlari 3 ta turga bo‘linadi:

– nazorat ostida o’qitish algoritmlar,

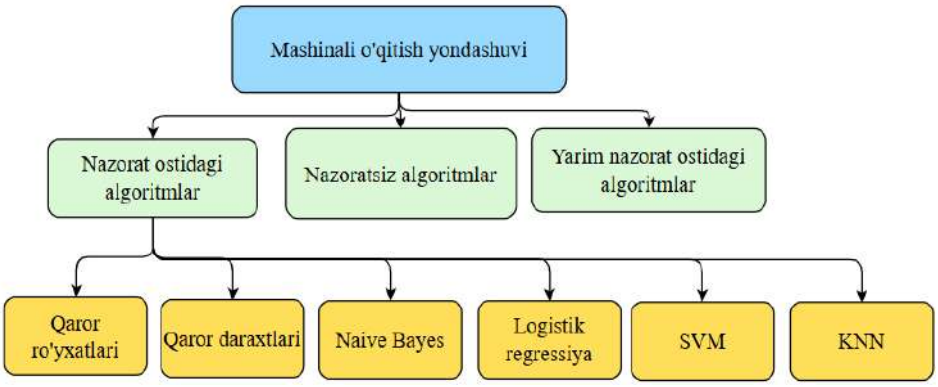
– nazoratsiz o’qitish algoritmlar,

– yarim nazorat ostida o’qitish algoritmlar

Ushbu yondashuvda tizim to‘g‘ri ma’noni aniqlash uchun teglangan dataset o‘qitiladi. Klassifikator xususiyatlarni o‘rganish va yangi kiritilgan gaplardagi noaniq so‘zning ma’nolarni belgilash uchun qo‘llaniladi. Ushbu yondashuvda, noaniq so‘z va noma’lum so‘zning matndagi birikuvchilari tizimga kirish sifatida beriladi.

Jahon tajribasiga ko‘ra WSD tizimilarini ishlab chiqishda asosan mashinali o‘qitish yondashuvidan foydalanilgan. Mashinali o‘qitish yondashuvida inson omili yordamida yaratilgan, ma’nosi annotatsiyalangan ma’lumotlardan foydalaniladi. Klassifikator noaniq so‘zga tegishli misollarni ham o‘z ichiga olgan ma’lumotlar to‘plami asosida o‘qitiladi. Klassifikator joriy noaniq so‘zni kirish sifatida qabul qiladi va to‘g‘ri ma’no uchun qo‘lda yaratilgan ma’no ma’lumotlar to‘plamini qidiradi. O‘qitiladigan ma’lumotlar to‘plami klassifikatorni o‘qitish uchun ishlatiladi, bu ham noaniq so‘zni o‘z ichiga oladi.

So‘z ma’nosini aniqlashda Nazorat ostida o‘qitish algoritmlardan foydalanish.



2-rasm: Nazorat ostida o‘qitish algoritmlar

2-rasmdan ifodalanganidek nazorat ostida o’qitish algoritmlar sinfi ham bir nechta algoritmlarni qamrab oladi. Ushbu algoritmlar yordamida so‘z ma’nosini aniqlashda teglangan ma’lumotlarga zarur bo‘ladi. Oldingi maqolalarimizda Naïve Bayes algoritmi yordamida omonimiyani aniqlash haqida ma’lumot bergan edik [7].

Qaror ro‘yxatlari. Qaror ro‘yxatlari "agar - unda-aks holda" qoidalar to‘plamidir. So‘zning xususiyatlarini olish uchun o‘qitish to‘plami belgilangan ro‘yxatda ishlatiladi. Ushbu qoidalar yordamida natijalar ro‘yxati yaratiladi. Yarowsky ta’riflaganidek “Qaror ro‘yxatlari (Decision Lists, DL) — noaniqlik muammolarini hal qilish uchun ishlatiladigan sodda vositalardan biridir “[4]. Ushbu yondashuv urg’uni tiklash, so‘z ma’nosini aniqlash va omofonlarni ajratish kabi vazifalarda muvaffaqiyatli qo‘llanilgan. Bu usul Senseval so‘z ma’nosini aniqlash eng muvaffaqiyatli tizimlardan biri bo‘lgan [5].

O‘qitish ma’lumotlari xususiyatlarni (features) ajratib olish uchun qayta ishlanadi, bu xususiyatlar logarifmik ehtimollik (log-likelihood) o‘lchovi yordamida baholanadi. Logarifmik ehtimollik qiymatlariga ko‘ra tartiblangan barcha xususiyatlarning ro‘yxati qaror ro‘yxatini hosil qiladi. Olimlar formulani ikkitadan ortiq noaniqlik holatlarini hisobga olish imkonini berish uchun moslashtirganlar.

$$weight(s_i|f_k) = \log(\frac{P(s_i|f_k)}{\sum_{j\neq i} P(s_j|f_k)} \quad (1)$$

Bu yerda, s_i– kontekstdagi noaniq so‘zning i- ma’nosi,
f_k– k-xususiyat,
$weight(s_i|f_k)$- s_i so‘z ma’nosining f_k xususiyat bo‘yicha vazni.

Qaror ro‘yxatlari algoritmida so‘z ma’nosining turli xususiyatlari bo‘yicha vaznlarini hisoblash orqali so‘z ma’nosini aniqlaydi. Bunda biror xususiyat bo‘yicha so‘zning har bir ma’nosi vaznlanadi va yuqori vaznga ega bo‘lgan so‘z ma’nosi natijasi sifatida qabul qilinadi. Xususiyatni aniqlashni turli usullari mavjud. Eng ko‘p foydalaniladigan xususiyat turlaridan biri *Bag-of-Words (BOW)* lug’atidir.

Bunda teglangan datasetdan joriy ko‘p ma’noli so‘zning atrofidagi so‘zlar chastotasi aniqlanadi. BOW ni hosil qilishda odatda joriy so‘z atrofidagi 5 ta so‘z olinadi.

Xususiyatlarni quyidagicha tanlab olish mumkin:

– Bigrams/trigrams – datasetdagi joriy so‘z ishtirokidagi bigram/ trigrammalar asosida;

– Lemmalar asosida xususiyatlar-so‘zning lug’aviy shakllari olinadi;

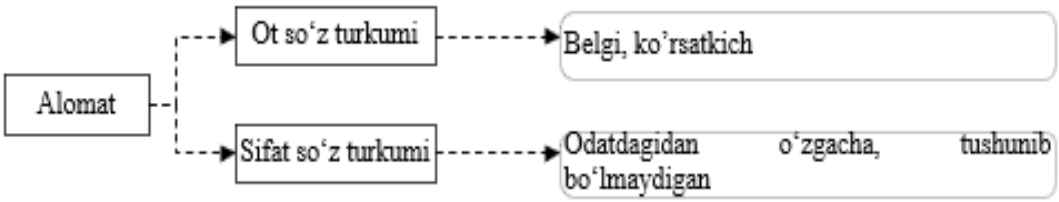
– Positsion xususiyatlar-so‘zning oldin yoki keyin uchraydigan so‘zlar;

– Semantik maydon asosida-tabiiy tilning WordNet tizimi

Kabi xususiyatlarni tanlab olish mumkin.

Osmondagi qora bulutlar shiddatli yomg’ir alomati demakdir.

Gapidagi *alomat* so‘zi omonim so‘z bo‘lib, u turli kontekstlarda quyidagicha ma’nolarni anglatishi mumkin (3-rasm).



3-rasm: Alomat so‘zining ma’nolari

Demak, ushbu aniqmaslikni qaror ro‘yxatlari algoritmi yordamida aniqlashni ko‘rib chiqamiz. (1) formuladag parametrlarning qiymatlarini aniqlab olamiz.

$$s = \{Belgi, ko'rsatkich; Odatdagidan\ o'zgacha, tushunib\ bo'lmaydigan\}$$

$$f = \{so'z, bigram, trigram, lemma, position, semantik\ maydon, win, morfologik\}$$

Yuqoridagi gapda keltirilgan *alomat* so‘zi xususiyatlar to‘plami bo‘yicha qiymatlari 1-jadvalda keltirilgan.

1-jadval: Gapdagi alomat so‘zining xususiyatlar bo‘yicha qiymatlari

Xususiyat turi	Namuna
So‘z	Yomg’ir / demakdir
Bigram	Yomg’ir alomati, alomati demakdir
Trigram	Yomg’ir alomati demakdi
Lemma	Lemma(shiddatli)=shiddat
position	So‘zdan oldingi so‘z=yomg’ir
Semantik maydon	Xususiyatlar maydoniga tegishli
Morfologik	Alomat so‘zi ot sifatida foydalanilgan.

Alomat so‘zining s_i ma’nosi bo‘yicha jamlangan gaplar to‘plamiga namuna keltiramiz.

2-jadval: Teglangan datasetdan namunalar

1-ma’nosi	2-ma’nosi
Bu kasallikning ilk alomati – tana haroratining ko‘tarilishidir.	**Bu alomat ko‘z qarashlari bilan u nimadir demoqchi edi.**
Osmonda qora bulutlar to‘plangani yomg’irning yaqin alomati edi.	**Alomat jimlik xonadagi muhitni yanada g’alati qildi.**
Donishmandlar aytadiki, odamning so‘zlari va xatti-harakati uning ichki dunyosining alomatidir.	**O‘sha kechadagi alomat tush hali-hanuz yodimdan chiqmaydi.**
Uning ko‘zidagi mung va yuzidagi charchoq yashirin g’amning alomatlari edi.	**Ular alomat harakatlari bilan atrofdagilarni bezovta qilishardi.**
Til o‘rganishda muvaffaqiyatli bo‘lishning eng yaxshi alomati – muntazam mashq qilishdir.	**Alomat so‘zlar bilan boshlangan suhbat kutilmagan tomonga burildi**
…	**…**

Xuddi shu kabi *alomat* so‘zining har bir ma’nosi bo‘yicha 100 tadan gap jamlandi. Alomat so‘zining har bir ma’nosi bo‘yicha vaznini hisoblashda biz *position* xususiyatidan foydalandik. *Position* xususiyati bo‘yicha statistik ko‘rsatkichlar oldik. Dastlab alomat so‘zidan oldin yomg’ir so‘zi uchrash ehtimolligini hisobladi. Quyidagi natijalar olindi.

$$P(s_1|yomg'ir) = \frac{C(s_1|yomg'ir)}{C(s_1)} = \frac{12}{100}$$

$$\sum_{j \neq i} P(s_j|f_k) = \frac{13}{200}$$

$$weight(s_1|yomg'ir) = \log\left(\frac{0.12}{0.065}\right) = \log(1.85) = 0.889$$

Ko‘rinib turganidek, *alomat* so‘zidan oldin *yomg’ir* so‘zi uchraganda uning s_1 ma’noni anglatishdagi vazni 0.889 ga teng. Endi s_2 ma’nosi bo‘yicha vaznini hisoblaymiz.

$$P(s_2|yomg'ir) = \frac{C(s_2|yomg'ir)}{C(s_2)} = \frac{1}{100}$$

$$\sum_{j \neq i} P(s_j|f_k) = \frac{13}{200}$$

$$weight(s_2|yomg'ir) = \log\left(\frac{0.01}{0.065}\right) = \log(0.15) = -0.81$$

s_1 ma’nosi bo‘yicha vazni -0.81 ga teng ekanligini ko‘rish mumkin. Ma’lumki noaniq so‘zning vazni 0 yoki manfiy qiymatga teng bo‘lsa, demak ushbu xususiyat bo‘yicha so‘zning joriy ma’nosi ahamiyatga ega emas.

Osmondagi qora bulutlar shiddatli yomg’ir alomati demakdir.

Gapidagi *alomat* omonim so‘zining ma’nosini aniqlashda Qaror ro‘yxatlari algoritmidan foydalanilganda quyidagicha vaznlar aniqlandi.

$$weight(belgi, xususiyat|yomg'ir) = 0.889$$

$$weight(Odatdagidan\ o'zgacha|yomg'ir) = -0.81$$

Aniqlangan vaznlardan quyidagicha qaror qabul qilish mumkin, berilgan gapdagi noaniq so‘z “$belgi, xususiyat$” ma’nosini anglatgan.

Qaror ro‘yxatlari (Decision lists) algoritmidan foydalanish vazn hisoblash xususiyatlarining ko‘pligi, uni tanlashning ixtiyoriyligi yoki mavjud barcha xususiyatlardan foydalangan holda ham vaznlashtirish mumkinligi bilan ushbu masalani yechishda qulaylik yaratib beradi. Agar omonim so‘z turli so‘z turkumlari doirasida turli ma’nolarga ega bo‘lsa, u holda *morfologik* xususiyatdan foydalanish maqsadga muvofiq. Chunki omonim so‘z mumkin bo‘lgan har bir so‘z turkumi doirasida faqat bitta ma’noga ega bo‘ladi. Teglangan datasetda ham bu aks ettiriladi albatta, u holda statistik ko‘rsatkichlar nisbatan kichikroq dataset ustida hisoblanadi hamda aniqroq natijalar olinadi.

Xulosa o‘rnida shuni aytish mumkinki, Nazorat ostida o‘qitish algoritmlar inson omili tomonidan teglangan, tekshirilgan ma’lumotlar ustida ish olib boradi. Qisqacha qilib aytganda namunalar asosida o‘rganadi. Ushbu o‘qitish yondashuvi namunalardan klassifikatsiya modelini ishlab chiqadi. Model har safar misol kelganda xususiyatlar fazosida nuqtalarni saqlaydi, bu yondashuv yangi misol va xususiyatlar fazosidagi ma’lumotlar o‘rtasidagi masofani hisoblaydi. Oldindan saqlangan ma’lum bir sinfga yaqinroq bo‘lsa, o‘sha sinfga qabul qilish imkoniyati yuqori bo‘ladi. Ushbu maqolada so‘z ma’nosini aniqlash uchun qaror ro‘yxatlari algoritmidan foydalanishning batafsil tahlili keltirildi. Asosan, mashinali o‘qitish usulida bu jarayonni amalga oshirish uchun katta korpusdan foydalaniladi va ushbu usulda aniqlash jarayoni uchun tizimni o‘qitish zarur. Ushbu mashinali o‘qitish algoritmlari yaxshi natijalar ko‘rsatadi, ammo asosiy kamchiliklari shundaki, ular katta korpusga ehtiyoj sezadi.

Foydalanilgan adabiyotlar:

1. Kilgarriff, A., Evans, M. (eds): Special issue on SENSEVAL. Computer and the Humanities, 34(1-2) (2000)
2. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: Five Papers on WordNet. Special Issue of International Journal of Lexicography, 3(4), (1990)

3. S.Ramana, M.V.Ramana Murthy, & N.Bhaskar. (2017). Ensuring data integrity in cloud storage using ECC technique, International Journal of Advanced Research in Science and Engineering, BVC NS CS 2017, 06(01), 170–174. ISSN Number: 2319-8346

4. Yarowsky, D. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French’, in Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pp. 88--95. 1994.

5. Kilgarriff, A. and M. Palmer. (eds). Special issue on SENSEVAL. Computer and the Humanities, 34 (1-2). 2000

6. R Alugubelli. (2016). Exploratory Study of Artificial Intelligence in Healthcare. International Journal of Innovations in Engineering Research and Technology, 3(1), 1–10.

7. B.Elov, X.Axmedova. Naïve Bayes klassifikatori yordamida omonimiyani aniqlash. Muhammad al-Xorazmiy avlodlari ilmiy-amaliy va axborot tahliliy jurnal 4(26)/2023 ISSN -2181-9211, 194-201-b.

8. E. B. Boltayevich and A. X. Ilxomovna, "Business Process Modeling That Distinguishes Homonymy Within Three Parts of Speechs in The UzbekLanguage," 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 278-283, doi:10.1109/UBMK55850.2022.9919453

9. Axmedova X. I. Turli so‘z turkumlari orasidagiomonimiyani aniqlovchi matematik modellar// Scienceand innovation international scientific journal volume 1issue 7 uif-2022: 8.2 | issn: 2181-3337.https://doi.org/10.5281/zenodo.7238546[12]

10. Axmedova X.I. Chastotali usul yordamida omonimiyani aniqlash// “O‘zbek amaliyfilologiyasi istiqbollari” Respublika ilmiy-amaliy konferensiyasi Toshkent: 2022.-164-170 b.

LEKSIK DARAJADAGI PEREFRAZ TAKSONOMIYASINI AMALGA OSHIRISH BOSQICHLARI

Axmedova Xolisxon Ilxomovna,
Texnika fanlari falsafa doktori PhD
a.xolisa@navoiy-uni.uz
ToshDO‘TAU

Ko‘charova Iroda Shavkat qizi,
Kompyuter lingvistikasi yo‘nalishi talabasi
kocharovairoda516@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada, tabiiy tilga ishlov berish (Natural Language Processing - NLP) sohasidagi perefraz qilish (paraphrasing) algoritmlariga oid muhim tushunchalar keltirib o‘tiladi. Perefraz qilishning leksik darajadagi taksonomiyasi, sinonimlari, antonimlari va morfologik o‘zgarishlari kabi turlari orqali gaplarni semantik jihatdan ekvivalent tarzda ifodalash jarayoni o‘rganiladi. Glove algoritmi so‘z vektorlarining o‘rtacha qiymati va Skip-gram usullari yordamida gaplarning semantik o‘xshashligini baholashga yordam beradi.

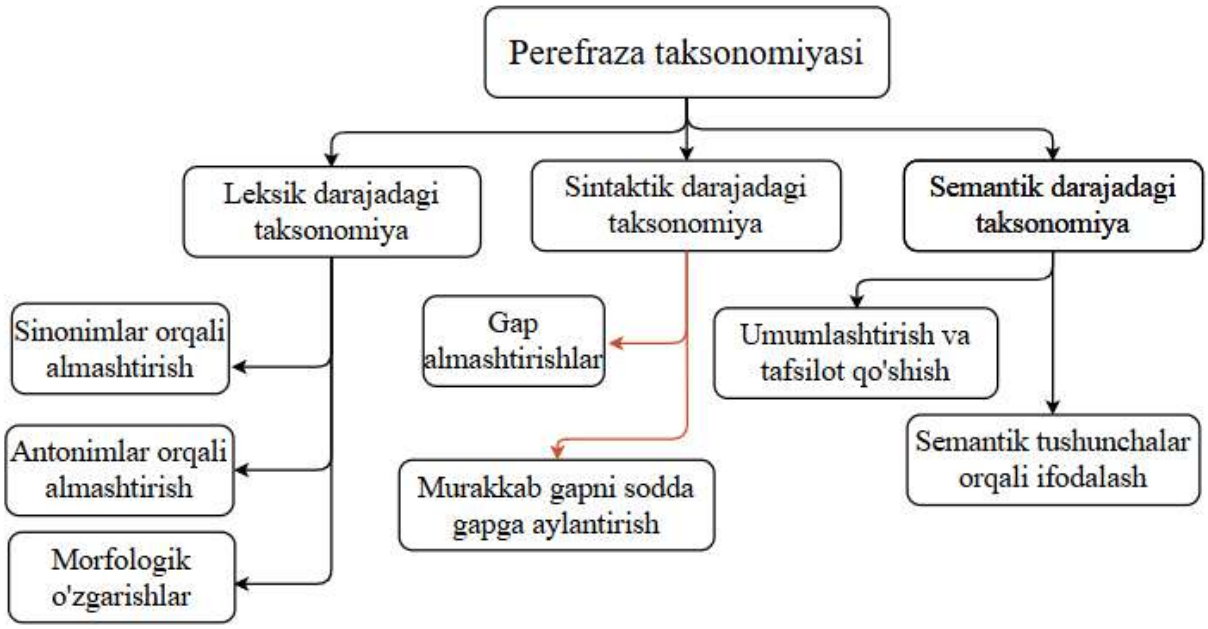
Abstract: This article presents important concepts related to paraphrasing algorithms in the field of Natural Language Processing (NLP). It explores the process of expressing sentences in a semantically equivalent manner through types such as lexical-level taxonomy, synonyms, antonyms, and morphological variations. The GloVe algorithm and Skip-gram methods help evaluate the semantic similarity of sentences based on the average value of word vectors.

Аннотация. В данной статье представлены важные понятия, относящиеся к алгоритмам перефразирования в области обработки естественного языка (Natural Language Processing — NLP). Рассматривается процесс выражения предложений в семантически эквивалентной форме через такие виды, как таксономия на лексическом уровне, синонимы, антонимы и морфологические изменения. Алгоритм GloVe и метод Skip-gram помогают оценить семантическое сходство предложений на основе среднего значения векторов слов**.**

Kalit so‘zlar: *Perafraza taksonomiyasi, NLP, Skip-gram, leksik darajadagi taksonomiya, sinonim, antonym, morfologik o‘zgarishlar, GloVe.*

Tabiiy tilga ishlov berish texnologiyalari sun’iy intellektning muhim yo‘nalishlaridan biri bo‘lib, tilni tushunish va ishlov berishda asosiy o‘rinni egallaydi. Ushbu maqola so‘zlarni sonli ifodalashning turli yondashuvlari va chuqur o‘rganish modellarining asoslari, afzallik va qiyinchiliklarini o‘rganish uchun tezkor ma’lumot bo‘lib xizmat qilishi mumkin. Perafraza - bu bir xil ma’noni ifodalovchi turli xil so‘z yoki iboralar bilan ifodalash usuli. Masalan, “o‘qituvchi” va “muallim”

so‘zlari bir xil ma’noni bildiradi, lekin turli kontekstlarda qo‘llanilishi mumkin. Tabiiy tilga ishlov berish sohasida gaplarni perefraz qilish, gap ekvivalentlarni yaratish, gaplarni semantik generatsiya qilish tushunchalari o‘zaro sinonim tushunchalardir. Gapning perefrazi uning ma’nosini saqlab qoladi, biroq asl shakldan farqli so‘z va iboralarni tanlaydi. Perefraz yaratish (Paraphrase Generation-PG) turli tabiiy tilga ishlov berish ilovalari ishlab chiqishda muhim masala bo‘lib, u tabiiy til birliklarini kengaytiradi. Perefraz yaratish *savol-javob tizimlari, mashina tarjimasi va axborot izlash* kabi ko‘plab vazifalarni bajarishda uchun muhim rol o‘ynaydi [1]. Savol - javob tizimlarida PG semantik jihatdan ekvivalent savollarni hamda ularga turli ko‘rinishdagi javoblarni yaratish uchun ishlatilishi mumkin. Bu javoblarni qidirishda so‘zlar mos kelmasligi muammolarini hal qilishga yordam beradi [2]. Perafraza taksonomiyasi esa bunday o‘xshashliklarni turkumlash va tasniflash uchun tuzilgan sxema(1-rasm).



1-rasm: Perefraz qilish taksonomiyasi

O‘zbek tilidagi gaplarni perefraz qilishni taksonomiya bo‘yicha amalga oshirish jarayonini ko‘rib chiqamiz.

Leksik darajadagi taksonomiya - bu so‘zlar, ularning ma’nolari va bir-biriga bo‘lgan munosabatlari asosida so‘zlarning turkumlarini tasniflash yoki tartibga solish tizimi. Odatda, bu jarayon leksik birikmalarni o‘rganish va tasniflash orqali amalga oshiriladi. Leksik darajadagi taksonomiyaning asosiy yo‘nalishlari 1-rasmda ham aks etgan bo‘lib, bulardan birinchi bosqich sinonimlar orqali almashtirishdir. *Sinonimlar orqali almashtirish* - bu so‘zlarni yoki iboralarni ma’no o‘zgarishisiz, lekin turlicha shaklda ifodalash uchun ishlatiladigan usul. Masalan, quyidagi gapni ko‘rib chiqamiz.

Asl gap: *“Odobli insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi”*

Sinonimi orqali almashtirilgan gaplar: bunda gapdagi har bir so‘zning sinonimlari ro‘yxatini shakllantirish kerak bo‘ladi. Ya’ni

Odobli -> boadab, axloqli, tarbiyali.

1. Boadab insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi.

2. Axloqli insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi.

3. Tarbiyali insonlar nafaqat boshqalarga hurmat ko‘rsatadi, balki o‘zini ham tinch va baxtli his qiladi.

Bu yerda bir so‘z o‘rnida uning sinonimi ishtirok etgan bo‘lsada, gapning ma’nosiga ta’sir ko‘rsatmayabdi. Sinonimlar orqali almashtirish, tilni yanada boyitishga yordam beradi va gaplarni turli shakllarda ifoda etishda samarali usul hisoblanadi. Bu orqali matnni yangilash va gapning ma’no samaradorligini oshirish mumkin.

Antonimlar orqali almashtirish - bu o‘zaro qarama-qarshi bo‘lgan so‘zlar yoki iboralar yordamida matndagi so‘zlar yoki gaplarni almashtirish usuli. Masalan:

Asl gap: *“Uning so‘zlari aniq va mantiqiy edi, lekin fikrlashi tor”*.

Antonimi orqali almashtirilgan gap: *“Uning so‘zlari aniq va mantiqiy edi, lekin fikrlashi keng emas”*

Asl gap: *“Izohli lug‘at kartotekasi hali ham kambag‘al”*

Antonimi orqali almashtirilgan gap: *“Izohli lug‘at kartotekasi hali ham boy emas”*

Yuqoridagi misolda ko‘rganimizdek, asl gapga nisbatan antonimi orqali almashtirilgan gapda ma’no boyoqdoligi yuqoriroqdir. Ushbu yo‘nalish, ayniqsa, *badiiy asarlarda, insholar va tahliliy matnlarda* juda foydalidir. O‘qishga qiziqarli va boy ma’no kiritish uchun antonimlardan foydalanish matnni yangi nuqtai nazardan ko‘rishga yordam beradi.

Leksik darajadagi taksonomiyaning oxirgi yo‘nalishi bu - morfologik o‘zgarish. Ushbu yo‘nalish so‘zlarga qo‘shiladigan shakllarni o‘zgartirish orqali ifodalanadi. Misol sifatida quyidagilarni keltirish mumkin.

Asl gap: *“U mamlakatda ekonomika fanlari akademiyasini barpo qilish masalasini muhokama qilish kerak edi deb hisoblaydi”*.

Perefraz qilingan gap: *“U mamlakatda ekonomika fanlari akademiyasini barpo qilish masalasini muhokama qilish kerak edi deb hisoblamaydi emas”.*

Asl gap: *“Uning gaplari bolalarcha sodda, samimiy va aniq edi”.*

Perefraz qilingan gap: *“Uning gaplari bolalarcha sodda, samimiy va noaniq emas edi”.*

So‘zlarning morfologik shaklini o‘zgartirish gap yoki matn ma’nosiga ta’sir qilmaydi. Gap uchun to‘g‘ri tanlangan morfologik shakllar so‘zlarning ma’nosini aniqroq ifodalashga yordam beradi, matnni boyitadi, soddalashtiradi yoki yangilaydi.

Leksik avtomatik perefrazni amalga oshirishda bir algoritmlardan foydalanish mumkin. Shunday algoritmlardan biri GloVe algoritmi bo‘lib, uning asosiy vazifasi so‘zning semantik o‘xshashlik koeffitsiyentlari aniqlashdan iborat. O‘xshashlik koeffitsiyentlarini aniqlashda dastlab berilgan dataset asosida matritsa tuziladi. Bu matritsadan gaplarning xususiyatlarini hisoblash uchun ishlatiladi. Ushbu matritsaning qatorlari korpus gaplarini, ustunlari esa xususiyatlarni ifodalaydi. Xususiyatlar quyidagilarni o‘z ichiga olishi mumkin:

Unigrams (bitta so‘z)	“Men”, “yangi”, “kitob”, “o‘qidim”
Bigrams (ketma-ket keluvchi so‘z juftliklari)	“Men yangi”, “yangi kitob”, “kitob o‘qidim”
Dependency pairs (so‘zlar orasidagi munosabat)	(“Men”, “o‘qidim”) - “Men” subyekt, “o‘qidim” fe’l bilan bog‘langan. (“yangi”, “kitob”) - “yangi” sifat, “kitob” ot bilan bog‘langan. (“kitob”, “o‘qidim”) - “kitob” obyekt, “o‘qidim” fe’l bilan bog‘langan.

2-rasm: Xususiyatlarni taqqoslash jadvali

Matritsa tuzilganidan so‘ng, so‘z chastotasining Kullback-Leibler farqlanishi (TF-KLD) xususiyat uchun vazn sxemasi hisoblanadi. Ushbu vazn berishning maqsadi muhimroq xususiyatlarga yuqori vazn berishdir. Yangi vazn berish sxemasi barcha xususiyatlarning vaznini hisoblash uchun quyidagi faktorlarga asoslanadi:

Xususiyatning parafrazlangan gaplarda mavjud bo‘lish ehtimoli	U juda yaxshi raqsga tushishni biladi.
Xususiyatning parafrazlanmagan gaplarda mavjud bo‘lish ehtimoli	U yaxshi raqsga tushadi.

3-rasm: Perefrazlashda xususiyatlar: Mavjudlik ehtimollari

Matritsa faktorizatsiyasi usuli yordamida gaplar vektorlarini ajratish uchun ishlatiladi. Ushbu vektorlar yordamida gaplar o‘rtasidagi o‘xshashlikni o‘lchash mumkin.

So‘z vektorlarining o‘rtacha qiymati

Ushbu usulda gapni ifodalovchi vektor so‘zlarning o‘rtacha vektorlari ushbu gapga tegishli deb hisoblanadi. Oldindan o‘qitilgan *qo‘lqop* so‘z vektorlaridan foydalangan[12]. *Qo‘lqop* semantik maydonidan gapdagi har bir so‘z uchun mos keladigan vektor tushuniladi va gaplarni ifodalash uchun ushbu vektorlar orasidagi kosinus o‘xshashligi gaplarning semantik o‘xshashligini baholash uchun ishlatiladi.

O‘qitishga asoslangan vektorlar. Bundan tashqari, so‘zlarni joylashtirish (word embedding) yordamida yaratilgan so‘z vektorlari asosida gap o‘xshashligini aniqlash modelini Mueller taklif qilgan[19]. Uzoq va qisqa muddatli xotira (LSTM) modelida oddiy bir o‘zgartirish taklif qilinadi va treningda qo‘llaniladi. Kirish ma’lumotlarida har bir gap so‘z vektorlaridan iborat bo‘lib, bu vektorlar bir xil o‘lchamga ega. Bundan tashqari, har bir gap juftligi uchun kirish ma’lumotlarida inson tomonidan baholangan o‘xshashlikni ifodalovchi belgi beriladi. Taklif qilingan Manhattan LSTM (MaLSTM) modelini kirish gapi vektorlarini o‘qish uchun ishlatadi va oxirgi yashirin qatlamni kirish gaplari uchun taqdimot sifatida ishlatadi. Gaplar uchun yangi yaratilgan taqdimot semantik o‘xshashlik o‘lchovini o‘rganish uchun qo‘llaniladi.

Skip-Gramm vektorlari. Ushbu usul nazoratsiz o‘quv modeliga asoslangan gaplar uchun vektorli tasvirni oladi. Ushbu modelga kiritilgan senentsiyalar vektorlar to‘plami sifatida ifodalanadi. Bu vektorlar, gaplarning so‘zlariga mos keladigan, oldindan o‘qitilgan word2vec fazosidan tanlanadi. Skip-thought modeli, berilgan so‘zga asoslanib atrofdagi so‘zlarni bashorat qiladigan Skip-gram modeliga o‘xshashdir. Biroq, Skip-thought modeli fokuslangan gapga asoslanib atrofdagi gaplarni bashorat qiladi. Gaplar o‘xshashligini hisoblashda, avvalo gaplar oldindan o‘qitilgan modelga yuborilib, vektor taqdimoti olinadi. So‘ngra, o‘xshashlik o‘rganish vazifasiga asoslanib hisoblanadi.

Xususiyat	Skip-gram	Skip-thought
Maqsad	So‘zlarni konteksdagi boshqa so‘zlar yordamida bashorat qilish	Gap yoki matnni vektorga aylantirish
Yondashuv	So‘zlar o‘rtasidagi semantic o‘xshashlikni o‘rganish	Gaplar o‘rtasidagi semantic o‘xshashlikni o‘rganish
Natija	So‘z vektorlarini yaratadi	Gap yoki matn vektorlarini yaratadi
Turi	Word2Vec modelida ishlatiladi	Reccurrent Neural Network (RNN) yordamida ishlaydi
Misol	“Men kitobni o‘qidim” - [o‘qish, kitob]	“Men kitobni o‘qidim” va “Men sport bilan shug‘ullandim” orasidagi o‘xshashlik

4-rasm: Skip-gram va Skip-thought model xususiyatlari

Vektor asosidagi o‘xshashlik yondashuvi, ayniqsa chuqur o‘rganish yondashuvini qo‘llaganidan so‘ng yaxshi natijalar ko‘rsatgan bo‘lsa-da, gaplarining asosiy xususiyatlarini aniqlashda hali ko‘p yaxshilanishlar uchun imkoniyatlar mavjud.

Umuman olganda, ko‘plab yondashuvlar tushuntirilgan texnikalarning ikkita yoki undan ortiq birlashtirilishini qo‘llaydi. 1-jadvalda muhokama qilingan ba’zi yondashuvlar va har bir taklif qilingan tizimda ishlatilgan texnikalar ko‘rsatilgan. Bundan tashqari, har bir taklif qilingan tizimni baholash uchun ishlatilgan ma’lumotlar to‘plami va erishilgan natijalar keltirilgan. 1-jadvalda har bir yondashuvning tafsilotlarini ko‘rsatadi va zamonaviy texnologiyalarni chizadi. 1-jadvalda ko‘rsatilganidek, gap o‘xshashligini o‘lchash uchun aytib o‘tilgan metodlar aralash tarzda qo‘llanilgan.

Bundan tashqari, gaplar o‘rtasidagi o‘xshashlikni o‘lchash uchun turli yondashuvlarni birlashtirish yaxshi natijalar ko‘rsatdi, chunki bu muammoni turli nuqtai nazarlardan ko‘rib chiqadi. Masalan, so‘zga asoslangan yondashuvni va tuzilishga asoslangan yondashuvni birlashtirish, yaxshi tuzilgan gaplar uchun ham, yaxshi tuzilmagan gaplar uchun ham samarali ishlaydigan tizimni yaratadi. Shuningdek, o‘xshashlikni o‘lchashda strukturaviy ma’lumotlardan foydalanish keng tarqalgan emas. Bundan tashqari, gap o‘xshashligini hisoblashda gap tuzilmasidan foydalanishni o‘rganish uchun ko‘plab imkoniyatlar mavjud.

Foydalanilgan adabiyotlar:

1. J Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. Parabank: Monolingual bi text generation and sentential paraphrasing via lexically constrained neural machine translation. arXiv preprint, 2019.
2. Lin, D., Pantel, P.: Discovery of inference rules for question-answering. Natural Language Engineering 7(4), 343–360 (2001)
3. Mengqiu Vang, Smit na, Mitamura T. jeop- ardy modeli nima? Qa uchun kvazi-sinxron grammatika, yilda: EMNLP-CoNLL; 2007, 7. p. 22-32.
4. Mamdouh Faruk, Mitsuru Ishizuka, Danushka Bollegala. Grafik taalukli asoslangan semantik qidiruvi. Metadata va semantikani tadqiq qilish bo‘yicha 12-xalqaro konferentsiya materiallari, Kipr; 2018 yil. https://doi. org/10.1007/ 978-3-030 -14401-2_8.
5. Aliguyev RM. Matnni avtomatik umumlashtirish uchun yangi gap o‘xshashlik o‘lchovi va gapga asoslangan ekstraktiv texnika, ilovalar bilan ekspert tizimlari, 2009; 36:7764-72. https://doi.org/10.1016/j.eswa.2008.11.022.
6. De Boni M, Manandhar S. savolga javob berish uchun gap o‘xshashligini semantik dolzarblik metrikasi sifatida ishlatish. Savollarga javob berishda yangi yo‘nalishlar bo‘yicha AAAI simpoziumining materiallari; 2003.
7. Alzahrani S, Salim N, AbrahamA. Plagia - rism lingvistik naqshlarini, matn xususiyatlarini va aniqlash usullarini, tizimlar, odam va kibernetika bo‘yicha IEEE operatsiyalarini tushunish. 2012; 42(2): 133-49. https://doi.org/10.1109 / TSMCC.2011.2134847.
8. Jaro MA. 1985 yilda qo‘llanilgan rekord bog‘lanish metodologiyasidagi yutuqlar Tampa Florida aholini ro‘yxatga olish, Amerika statistika jamiyati jurnali. 1989; 84(406): 414-20. https: / / doi.org/10.1080/01621459.1989.10478785.
9. Vinkler biz. Fellegi-Sunter modelidagi yozuvlarni bog‘lash modelidagi taqqoslagich o‘lchovlari va Kengaytirilgan qaror qoidalari. Tadqiqotni tadqiq qilish usullari bo‘limining materiallari, Amerika statistika assotsiatsiyasi; 1990 yil. p. 354-59.
10. Mikolov t, Chen K, Corrado G, Dean J. vektor kosmosda so‘z vakolatxonalari samarali estima - yozildi, arXiv preprint arXiv: 1301.3781. 2013.
11. Miller GA. So‘z tarmog‘i: ingliz tili uchun leksik ma’lumotlar bazasi, ACM aloqalari. 1995; 38(11):39-41. https: / / doi.org/10.1145/219717.219748.
12. Vu Z, Palmer M. fe’l semantikasi va leksik tanlov. Davomlari32ndannualmeetingoftheassociation Uchun Hisoblash Tilshunosligi, Las Cruces, Nyu-Meksiko; 1994. https: / /doi.org/10.3115/981732.981751.
13. Vijaymeena MK, Kavitha K. matn qazib olish, mashinani o‘rganish va ilovalardagi o‘xshashlik mea - sures bo‘yicha so‘rov: xalqaro jurnal. 2016; 3(1):19-28. https://doi. org/10.5121 / mlaij.2016.3103.

14. Guessoum D, Miraoui M, Tadj C. keng tarqalgan hisoblashda semantik o‘xshashlik choralarini o‘rganish, aqlli sezgi va aqlli tizimlar bo‘yicha xalqaro jurnal. 2015; 8(1): 125-58. https://doi.org/10.21307/ijssis-2017-752.

15. Gomaa, fahmi AA. Matn o‘xshashligi yondashuvlarini o‘rganish, kompyuter ilovalarining InternationalJournal. 2013; 68(13):13-18. https://doi.org/10.5120/11638-7118.

16. 16 Pradhan N, Gyanchandani M, Vadhvani R. irda ishlatiladigan matn o‘xshashligi texnikasi va uni qo‘llash bo‘yicha sharh, xalqaro kompyuter dasturlari jurnali. 2015; 120(9). https://doi.org/10.5120/21257-4109.

17. Xorxe Martines-Gil, Mario Pich. Semantic yozishmalar kashfiyotini qo‘llab-quvvatlash uchun inson adabiyotida so‘zlarning birgalikda paydo bo‘lishini tahlil qilish. Bilim texnologiyalari va ma’lumotlarga asoslangan biznes bo‘yicha 14-xalqaro konferentsiya materiallari, Graz, Avstriya; 2014 yil. https: / /doi.org/10.1145/2637748.2638422.

18. Terra E, Clarke CLA. Statistik so‘z o‘xshashligi o‘lchovlari uchun chastotani baholash, NAACL ' 03 inson Tili texnologiyasi bo‘yicha hisoblash tilshunosligi Assotsiatsiyasining Shimoliy Amerika bobining 2003-yilgi konferentsiyasining materiallari-1-jild, Edmonton, Kanada; 2003 yil. p. 165-72. https://doi.org/10.3115/1073445.1073477.

19. Jonas Myuller, Thyagarajan A. Siam takroriy arxi - tectures gap o‘xshashligini o‘rganish uchun. Sun’iy intellekt bo‘yicha o‘ttizinchi AAAI konferentsiyasi materiallari, Feniks, Arizona; 2016. p. 2786-92. https://dl.acm.org / iqtibos.cfm?id=3016291.

IX SHO‘BA. KOMPYUTER LINGVISTIKASIDA TERMINOLOGIYA VA TERMINOGRAFIYA

UDK. 1751

KOGNITIV LINGVISTIKA NUQTAI NAZARIDAN POLIKODLI MATNLARNING SEMANTIK VA SINTAKTIK STRUKTURALARI

Odilova Kamola Taxirovna,
Tayanch doktoranti
kt.odilova@uzswlu.uz
O‘zDJTU

Annotatsiya. Ushbu maqolada polikodli matnlarning semantik va sintaktik strukturalari kognitiv lingvistika nuqtai nazaridan tahlil qilinadi. Polikodli matnlarning tarkibiy qismlari, ularning o‘zaro aloqalari va mazmun shakllanishidagi roli o‘rganiladi. Kognitiv yondashuv asosida polikodli matnlarning semantik va sintaktik xususiyatlari, ularning inson ongida qayta ishlanishi va tushunilishi jarayonlari tadqiq etiladi. Maqolada polikodli matnlarning turli kognitiv modellar asosida tahlili, ularning kontseptual tuzilishi va semantik maydonlari o‘rganiladi.

Abstract. This article analyzes the semantic and syntactic structures of multimodal texts from the perspective of cognitive linguistics. The components of multimodal texts, their interrelationships, and their role in content formation are examined. Based on a cognitive approach, the semantic and syntactic features of multimodal texts, their processing, and comprehension in human consciousness are studied. The article explores the analysis of multimodal texts based on various cognitive models, their conceptual structure, and semantic fields.

Аннотация. В данной статье анализируются семантические и синтаксические структуры поликодовых текстов с точки зрения когнитивной лингвистики. Исследуются компоненты поликодовых текстов, их взаимосвязи и роль в формировании содержания. На основе когнитивного подхода изучаются семантические и синтаксические особенности поликодовых текстов, процессы их обработки и понимания в человеческом сознании. В статье рассматривается анализ поликодовых текстов на основе различных когнитивных моделей, их концептуальная структура и семантические поля.

Kalit so‘zlar: *polikodli matnlar, kognitiv lingvistika, semantik struktura, sintaktik struktura, kontseptual tahlil, multimodallik.*

Zamonaviy lingvistikada polikodli matnlar tadqiqi tobora dolzarb ahamiyat kasb etmoqda. Polikodli matnlar – bu turli semiotik kodlar (verbal, vizual, audio va boshqalar) kombinatsiyasidan tashkil topgan murakkab tuzilmalardir. Ularning

semantik va sintaktik strukturalarini o‘rganish nafaqat nazariy, balki amaliy ahamiyatga ham ega, chunki bu kabi matnlar zamonaviy kommunikatsiyaning ajralmas qismiga aylanib bormoqda.

Kognitiv lingvistika nuqtai nazaridan polikodli matnlarni o‘rganish, ularning inson ongida qayta ishlanishi va tushunilishi jarayonlarini ochib berish imkonini beradi. Bu yondashuv polikodli matnlarning semantik va sintaktik strukturalarini yanada chuqurroq tushunishga yordam beradi.

Polikodli matnlarning semantik strukturasi ularning mazmuniy tashkil etilishini aks ettiradi. Bunda turli semiotik kodlar o‘rtasidagi ma’noviy munosabatlar, ularning o‘zaro ta’siri va umumiy mazmun shakllanishidagi roli muhim ahamiyat kasb etadi. Kognitiv lingvistika nuqtai nazaridan, polikodli matnlarning semantik strukturasi quyidagi jihatlarga ko‘ra tahlil qilinishi mumkin:

1. Kontseptual tuzilish: Polikodli matnlarda aks ettirilgan asosiy tushunchalar va ularning o‘zaro aloqalari.
2. Semantik maydonlar: Turli semiotik kodlar orqali ifodalangan ma’nolar va ularning o‘zaro bog‘liqligi.
3. Metaforik va metonimik ko‘chimlar: Polikodli matnlarda qo‘llanilgan kognitiv metafora va metonimiyalar.
4. Freymlar va skriptlar: Matnda aks ettirilgan bilim strukturalari va stereotip vaziyatlar.
5. Kontseptual integratsiya: Turli semiotik kodlardan kelgan ma’lumotlarning birlashishi va yangi ma’nolarning shakllanishi.

Polikodli matnlarning sintaktik strukturasi esa ularning tashqi tuzilishini, ya’ni turli semiotik kodlarning o‘zaro joylashuvi va aloqalarini aks ettiradi. Kognitiv lingvistika nuqtai nazaridan, bu struktura quyidagi jihatlarga ko‘ra tahlil qilinishi mumkin:

1. Vizual-verbal sintagmatika:

Vizual-verbal sintagmatika polikodli matnlarda verbal va vizual elementlarning o‘zaro joylashuvi va aloqalarini o‘rganadi. Bu tushuncha matnning lingvistik va vizual komponentlari qanday birlashib, yaxlit ma’no hosil qilishini tushuntirishga yordam beradi. Masalan, gazeta sahifasidagi sarlavha, matn va rasmlarning joylashuvi, yoki reklama plakatidagi so‘zlar va tasvirlarning o‘zaro munosabati vizual-verbal sintagmatikaning namunalaridir. Bu elementlarning o‘zaro joylashuvi va aloqalari ma’noning shakllanishida muhim rol o‘ynaydi va kitobxon e’tiborini boshqaradi.

2. Multimodal kogerentlik:

Multimodal kogerentlik turli semiotik kodlar (masalan, matn, rasm, ovoz, video) o‘rtasidagi bog‘liqlik va uyg‘unlikni anglatadi. Bu tushuncha polikodli

matnning turli qismlari qanday qilib bir-birini to‘ldirishi va yaxlit ma’no hosil qilishini tushuntirishga yordam beradi. Masalan, ilmiy maqoladagi matn va diagrammalar, yoki o‘quv videosidagi ovoz va animatsiyalar o‘rtasidagi uyg‘unlik multimodal kogerentlikning namunasidir. Kogerentlik darajasi yuqori bo‘lgan polikodli matnlar odatda kitobxon tomonidan osonroq tushuniladi va eslab qolinadi.

3. Informatsion ierarxiya:

Informatsion ierarxiya polikodli matndagi ma’lumotlarning ahamiyatiga ko‘ra taqsimlanishini anglatadi. Bu tushuncha matnning qaysi qismlari eng muhim ma’lumotlarni o‘z ichiga olishini va qanday qilib kitobxon e’tiborini jalb qilishini tushuntirishga yordam beradi. Masalan, gazetadagi asosiy sarlavha, kichik sarlavhalar, asosiy matn va rasmlar o‘rtasidagi ierarxiya, yoki veb-sahifadagi turli elementlarning joylashuvi informatsion ierarxiyaning namunasidir. To‘g‘ri tashkil etilgan informatsion ierarxiya kitobxonga ma’lumotlarni samarali idrok etish va tushunishga yordam beradi.

4. Kognitiv xaritalar:

Kognitiv xaritalar tushunchasi polikodli matn elementlarining fazoviy joylashuvi va ularning idrok etilishini o‘rganishga yordam beradi. Bu yondashuv kitobxonning matn elementlarini qanday “xaritalashtirib” olishini va bu xaritalar asosida ma’noni qanday shakllantirishini tushuntirishga imkon beradi. Masalan, infografikadagi turli elementlarning joylashuvi, yoki darslikdagi matn va rasmlarning taqsimlanishi kognitiv xaritalarning shakllanishiga ta’sir qiladi. Samarali tashkil etilgan kognitiv xaritalar ma’lumotni eslab qolish va qayta ishlatishni osonlashtiradi.

5. Gestalt prinsiplari:

Gestalt prinsiplari polikodli matn elementlarining yaxlit idrok etilishini ta’minlovchi tamoyillardir. Bu prinsiplar vizual idrokni tushuntirishga yordam beradi va polikodli matnlarni loyihalashda keng qo‘llaniladi. Asosiy Gestalt prinsiplari quyidagilarni o‘z ichiga oladi:

a) Yaqinlik prinsipi: Bir-biriga yaqin joylashgan elementlar guruh sifatida idrok etiladi.

b) o‘xshashlik prinsipi: o‘xshash elementlar bir guruhga mansub deb qabul qilinadi.

c) Yopiqlik prinsipi: Inson miyasi to‘liq bo‘lmagan shakllarni to‘ldirib idrok etishga moyil.

d) Davomiylik prinsipi: Uzluksiz chiziqlar yoki shakllar yaxlit idrok etiladi.

e) Figura-fon prinsipi: Ob'ektlar old plandagi figura va orqa plandagi fon sifatida ajratiladi.

Bu prinsiplar polikodli matnlarning tuzilishini tushunish va samarali loyihalash uchun muhimdir. Masalan, veb-sahifa dizayni, infografikalar yoki reklama materiallari yaratishda Gestalt prinsiplaridan foydalanish ma’lumotlarni yanada samarali va jozibali tarzda taqdim etishga yordam beradi.

Polikodli matnlarning semantik va sintaktik strukturalarini kognitiv lingvistika nuqtai nazaridan o‘rganish quyidagi ahamiyatga ega:

1. Kommunikativ samaradorlik: Bu yondashuv polikodli matnlarning qanday qilib eng samarali tarzda ma’no uzatishini tushunishga yordam beradi. Bu bilim turli sohalarda, masalan, ta’lim, marketing va ommaviy axborot vositalarida qo‘llanilishi mumkin.
2. Kognitiv jarayonlarni tushunish: Polikodli matnlarni o‘rganish orqali biz insonning ma’lumotni qabul qilish, qayta ishlash va saqlash jarayonlari haqida ko‘proq bilib olamiz. Bu psixolingvistika va kognitiv psixologiya sohalarida muhim ahamiyatga ega.
3. Sun’iy intellekt va mashina o‘rganishi: Polikodli matnlarning strukturasini tushunish kompyuter tizimlari uchun murakkab vizual-verbal ma’lumotlarni qayta ishlash algoritmlarini yaratishda yordam beradi.
4. Madaniyatlararo muloqot: Turli madaniyatlarda polikodli matnlarning qanday idrok etilishi va talqin qilinishini o‘rganish madaniyatlararo muloqotni yaxshilashga yordam beradi.
5. Ta’lim metodikasini takomillashtirish: Polikodli matnlarning strukturasini tushunish o‘quv materiallarini yanada samarali va tushunarli qilish imkonini beradi.
6. Tarjima va lokalizatsiya: Polikodli matnlarning lingvistik va vizual jihatlarini tushunish tarjima va lokalizatsiya jarayonlarini takomillashtirishga yordam beradi.

1-jadval. Kognitiv struktural jadvali

Kognitiv model	**Tavsif**	**Qo‘llanilish sohalari**
Sxema nazariyasi	Bilimlarni tashkil etish va saqlash uchun mavjud mental strukturalar	Ta’lim, xotira tadqiqotlari, sun’iy intellekt
Mental modellar	Voqelik haqidagi ichki tasavvurlar va tushunchalar	Muammolarni hal qilish, qaror qabul qilish, tizimli fikrlash
Prototip nazariyasi	Kategoriyalarning eng xarakterli vakillariga asoslangan tushunchalarni shakllantirishni tushuntiradi	Tilshunoslik, kognitiv psixologiya, marketing
Freym nazariyasi	Ma’lumotlarni tashkil etish uchun ierarxik strukturalar	Sun’iy intellekt, kompyuter lingvistikasi, bilimlarni taqdim etish

Tarmoq modellari	Tushunchalar va ular orasidagi aloqalarni aks ettiruvchi tarmoqlar	Semantik xotira, assotsiativ o‘rganish, tilni qayta ishlash
Paralel taqsimlangan qayta ishlash	Miyaning parallel ravishda ko‘p sonli sodda birliklardan foydalanib ma’lumotlarni qayta ishlashini tushuntiradi	Neyron tarmoqlari, sun'iy intellekt, kognitiv neyrofan
Dual-kod nazariyasi	Verbal va noverbal ma’lumotlarni alohida, ammo o‘zaro bog‘liq tizimlar orqali qayta ishlash g‘oyasi	Multimediali ta’lim, xotira tadqiqotlari, vizual kommunikatsiya
Kognitiv yuk nazariyasi	Ishchi xotiraning cheklangan sig‘imini hisobga olgan holda o‘qitish va o‘rganishni optimallashtirish	Ta’lim dizayni, interfeys dizayni, axborotni taqdim etish
Situativ bilish	Bilish jarayonlarining kontekst va atrof-muhit bilan o‘zaro ta’sirini ta’kidlaydi	Ta’lim, sun’iy intellekt, ijtimoiy psixologiya
Metakognitiv model	o‘z bilish jarayonlarini kuzatish va boshqarish qobiliyati	Ta’lim strategiyalari, o‘z-o‘zini boshqarish, muammolarni hal qilish

Polikodli matnlarning semantik va sintaktik strukturalarini kognitiv lingvistika nuqtai nazaridan o‘rganish, ularning inson ongida qayta ishlanishi va tushunilishi jarayonlarini yanada chuqurroq tushunishga imkon beradi (*1-jadval*). Bu esa o‘z navbatida, samarali kommunikatsiya strategiyalarini ishlab chiqish, ta’lim jarayonlarini takomillashtirish va sun'iy intellekt tizimlarini rivojlantirishda muhim ahamiyat kasb etadi.

“Polikodli matnlarning semantik va sintaktik strukturalari ularning kontseptual mazmunini shakllantirishda muhim rol o‘ynaydi”[2]. Bu strukturalar nafaqat matnning tashqi ko‘rinishini, balki uning ichki ma’nosini ham belgilaydi. Shu bois, polikodli matnlarni tahlil qilishda ularning har ikki jihatini ham inobatga olish zarur.

Kognitiv lingvistika nuqtai nazaridan, polikodli matnlarning semantik strukturasi kontseptual metaforalar va metonimiyalar orqali ham shakllanishi mumkin. Masalan, reklama plakatlarida tez-tez uchraydigan “vaqt – puldir” metaforasi verbal va vizual kodlar kombinatsiyasi orqali ifodalanishi mumkin. Bu esa o‘z navbatida, matnning umumiy ma’nosini yanada kuchaytiradi va uning ta’sir kuchini oshiradi.

Polikodli matnlarning sintaktik strukturasi esa ularning vizual kompozitsiyasi bilan chambarchas bog‘liq. Gestalt psixologiyasi prinsiplari asosida tashkil etilgan polikodli matnlar, odatda, yaxlit va uyg‘un idrok etiladi. Masalan,

yaqinlik, o‘xshashlik va davomiylik prinsiplari asosida joylashtirilgan matn elementlari, ularning oson va tez tushunilishini ta’minlaydi[3].

Polikodli matnlarning semantik va sintaktik strukturalarini o‘rganishda multimodal diskurs tahlili usullaridan ham foydalanish mumkin. Bu usul turli semiotik resurslarning o‘zaro ta’sirini va ularning umumiy ma’no shakllanishidagi rolini o‘rganishga imkon beradi. Masalan, Kress va van Leeuwen tomonidan ishlab chiqilgan vizual grammatika nazariyasi[4], polikodli matnlarning vizual komponentlarini tahlil qilishda keng qo‘llaniladi.

Polikodli matnlarning kognitiv-lingvistik tahlili quyidagi amaliy ahamiyatga ega:

1. *Ta’lim sohasida*: Polikodli matnlarning strukturasini tushunish, o‘quv materiallarini yanada samarali tashkil etish va taqdim etish imkonini beradi.

2. *Marketing va reklama*: Polikodli matnlarning ta’sir mexanizmlarini tushunish, yanada samarali reklama va marketing materiallarini yaratishga yordam beradi.

3. *Axborot texnologiyalari*: Polikodli matnlarning strukturasini tushunish, foydalanuvchi interfeyslarini optimallashtirishga va axborotni yanada samarali taqdim etishga imkon beradi.

4. *Tarjima va lokalizatsiya*: Polikodli matnlarning lingvomadaniy xususiyatlarini tushunish, ularni boshqa tillarga va madaniyatlarga moslashtirishda muhim ahamiyatga ega.

5. *Ijtimoiy media*: Polikodli matnlarning strukturasini tushunish, ijtimoiy media kontentini yanada samarali yaratish va tarqatish imkonini beradi.

Polikodli matnlarning kognitiv tahlili bo‘yicha yanada chuqurroq tadqiqotlar o‘tkazish maqsadga muvofiq. Bu borada Forceville (2020) ning fikri e’tiborga loyiq: “Multimodal diskurs tahlili sohasida kognitiv yondashuvlarni qo‘llash orqali biz matnlarning murakkab tuzilishini yanada yaxshiroq tushunishimiz mumkin” [2]. Maqolada keltirilgan kognitiv modellarni amaliy misollar bilan boyitish tavsiya etiladi. Kress va van Leeuwen (2006) ta’kidlaganidek, “Nazariy modellarni real hayotdagi multimodal matnlarga qo‘llash orqali biz ularning amaliy ahamiyatini ko‘rsata olamiz” [4]. Polikodli matnlarning turli madaniyatlarda qabul qilinishi va talqin etilishi masalasini ham o‘rganish foydali bo‘lardi. Jewitt (2014) ning fikriga ko‘ra, “Multimodal kommunikatsiyaning madaniy jihatlari global kontekstda

muhim ahamiyat kasb etadi” [3]. Maqolada keltirilgan nazariy yondashuvlarni zamonaviy raqamli texnologiyalar kontekstida ko‘rib chiqish maqsadga muvofiq. Bateman (2017) ta’kidlaganidek, “Raqamli muhitda polikodli matnlarning yaratilishi va qabul qilinishi yangi tadqiqot yo‘nalishlarini ochib beradi” [1]. Polikodli matnlarning ta’lim jarayonidagi rolini ham o‘rganish tavsiya etiladi. Mayer (2009) ning multimedia ta’lim nazariyasiga ko‘ra, “Matnlar va tasvirlarning birgalikda qo‘llanilishi o‘quvchilarning materialni yaxshiroq tushunishi va eslab qolishiga yordam beradi” [5].

Xulosa qilib aytganda, polikodli matnlarning semantik va sintaktik strukturalarini kognitiv lingvistika nuqtai nazaridan o‘rganish, zamonaviy kommunikatsiya jarayonlarini yanada chuqurroq tushunish va takomillashtirish imkonini beradi. Bu sohada olib boriladigan kelajakdagi tadqiqotlar, turli fanlar kesishmasida, jumladan lingvistika, psixologiya, semiotika va axborot texnologiyalari sohasida amalga oshirilishi maqsadga muvofiq. Bunday kompleks yondashuv polikodli matnlarning tabiatini yanada to‘liqroq ochib berish va ulardan amaliy faoliyatda samarali foydalanish imkonini beradi.

Foydalanilgan adabiyotlar:

1. Bateman, J. A. (2017). Multimodality and genre: A foundation for the systematic analysis of multimodal documents. Springer.
2. Forceville, C. (2009). Non-verbal and multimodal metaphor in a cognitivist framework: Agendas for research. In C. Forceville & E. Urios-Aparisi (Eds.),
3. Jewitt, C. (2014). The Routledge handbook of multimodal analysis. Routledge.
4. Kress, G., & van Leeuwen, T. (2006). Reading Images: The Grammar of Visual Design (2nd ed.). London: Routledge.
5. Mayer, R. E. (2009). Multimedia learning (2nd ed.). Cambridge University Press.

X SHO‘BA. KOMPYUTER LINGVISTIKASIDA SUN’IY INTELLEKT

UDK: 808.852

KO‘P TILLI AVTOMATIK NUTQNI TANISH TIZIMLARINI YARATISH MUAMMOLAR VA YECHIMLARI

Ochilov Mannon Musinovich,
Texnika fanlari bo‘yicha falsafa doktori, PhD
ochilov.mannon@mail.ru
TATU

Narzullaev Oybek Otabek o‘g‘li
Magistrant
oybeknarzullaev99@gmail.com
TATU

Annotatsiya. Ushbu maqolada ko‘p tilli avtomatik nutqni tanib olish (Multilanguage ASR) texnologiyasidagi muammolar va ularning yechimlari haqida so‘z yuritiladi. Maqolada shuningdek, kam resursli tillar uchun ma’lumotlar yetishmasligi, tillarning o‘zaro aralashuvi, latentsiya va dialektlar kabi muammolarni hal qilishning turli usullari va yondashuvlarining afzalliklari hamda kamchiliklari ko‘rib chiqiladi.

Abstract. This article discusses the problems and solutions related to multilanguage Automatic Speech Recognition (ASR) technology. The article also examines the advantages and disadvantages of various methods and approaches to address issues such as data scarcity for low-resource languages, code-switching, latency, and dialects.

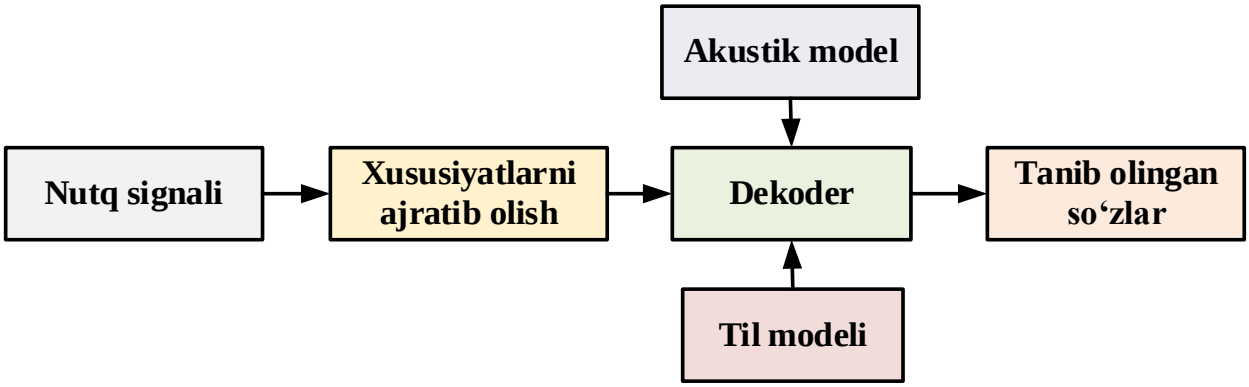
Абстракт. В этой статье рассматриваются проблемы и решения, связанные с многоязычным автоматическим распознаванием речи (multilanguage ASR). Также в статье анализируются преимущества и недостатки различных методов и подходов к решению таких проблем, как нехватка данных для языков с низкими ресурсами, переключение кодов, задержки и диалекты.

Kalit so‘zlar: *Ko‘p tilli ASR, nutqni tanib olish, kam resursli tillar, code-switching, end-to-end modellar, neyron tarmoqlar, SpeechLM, dialektlar.*

So‘nggi o‘n yilliklarda sun’iy intellekt va mashinaviy o‘qitish texnologiyalarining jadal rivojlanishi inson-nutq interfeyslari sohasida katta yutuqlarni keltirib chiqardi. Avtomatik nutqni tanib olish (Automatic Speech Recognition, ASR) tizimlari bugungi kunda virtual yordamchilardan tortib, real vaqtda tarjimaga qadar keng qo‘llanilmoqda[1,2]. Masalan, OpenAI tomonidan

ishlab chiqilgan Whisper modeli yoki Meta kompaniyasining SeamlessM4T tizimi ko‘p tilli real vaqtda tarjima imkoniyatlari orqali global muloqotni osonlashtirmoqda[17,18]. Biroq, global miqyosda ko‘p tilli ANT (multilanguage ASR) tizimlarini joriy etish va ularning samaradorligini oshirish hali ham murakkab va dolzarb muammo sifatida qolmoqda.

Dunyoda 7000 dan ortiq til mavjud bo‘lib, ularning har biri o‘ziga xos fonetik, grammatik va leksik xususiyatlarga ega. Shu bilan birga, zamonaviy jamiyatda tillarning o‘zaro ta’siri, globalizatsiya va ko‘p tilli muhitda muloqot qilish zarurati ko‘p tilli ANT texnologiyalariga bo‘lgan talabni yanada oshirmoqda.



1-rasm. ANT ning tizim arxitekturasi

Ko‘p tilli ANT ning asosiy maqsadi – turli tillarda nutqni aniq tanib olish, uni matnga aylantirish va foydalanuvchilar uchun tabiiy, samarali muloqot imkonini yaratishdir. Bu texnologiya nafaqat kundalik hayotda, balki ta’lim, tibbiyot, biznes va xalqaro aloqalar kabi sohalarda ham muhim ahamiyatga ega. Masalan, ko‘p tilli ANT tizimlari o‘qituvchilarga turli tillarda dars o‘tishda yordam berishi, shifokorlarga bemorlar bilan til to‘siqlarisiz muloqot qilish imkonini berishi yoki xalqaro konferensiyalarda real vaqtda tarjimani ta’minlashi mumkin. Biroq, ushbu imkoniyatlarni to‘liq ro‘yobga chiqarish uchun bir qator jiddiy muammolarni hal qilish zarur.

Hozirgi kunda ANT texnologiyalari asosan yuqori resursli tillar (masalan, ingliz, xitoy, ispan) uchun yaxshi optimallashtirilgan bo‘lib, Google Speech-to-Text, Amazon Transcribe yoki OpenAI Whisper kabi tizimlar ushbu tillarda yuqori aniqlik ko‘rsatadi. Biroq, dunyo tillarining aksariyati kam resursli hisoblanadi, ya’ni ular uchun yetarli audio va matn ma’lumotlari mavjud emas. Bundan tashqari, foydalanuvchilarning bir nechta tilni aralashtirib gapirishi (code-switching), dialektlarning xilma-xilligi, real vaqtda ishlashdagi kechikishlar va nutqning paralingvistik jihatlarini yo‘qotish kabi muammolar ko‘p tilli ANT tizimlari ning keng qo‘llanilishiga to‘sqinlik qilmoqda[8]. Masalan, o‘zbek tilida gapiruvchi

foydalanuvchi ruscha so‘zlarni aralashtirsa yoki xorazm shevasida nutq so‘zlasa, mavjud tizimlar ko‘pincha noto‘g‘ri yoki umuman javob bera olmaydi.

Ushbu muammolar nafaqat texnik jihatdan, balki ijtimoiy va madaniy nuqtai- nazardan ham katta ahamiyatga ega. Agar ko‘p tilli ANT tizimlari faqat bir nechta dominant tillarga xizmat qilsa, bu kam resursli tillarning raqamli dunyoda yo‘qolib borishiga olib kelishi mumkin. Shu sababli, ushbu texnologiyani rivojlantirish nafaqat foydalanuvchi tajribasini yaxshilash, balki lingvistik xilma-xillikni saqlash va global miqyosda teng huquqli muloqotni ta’minlash uchun ham muhimdir.

Mazkur maqola ko‘p tilli ANT texnologiyasidagi asosiy muammolarni chuqur tahlil qilishga va ularga qarshi innovatsion yechimlarni taklif qilishga qaratilgan. Tadqiqotda mavjud tizimlarning cheklovlari o‘rganiladi, ularning samaradorligi sinov ma’lumotlari (masalan, o‘zbek, turk va ingliz tillari aralashmasi) asosida baholanadi va yangi yondashuvlarning afzalliklari isbotlanadi.

Tadqiqotning asosiy maqsadi ko‘p tilli ANT tizimlarini yanada aniq, tezkor va tabiiy qilish yo‘llarini aniqlash hamda kam resursli tillar uchun qamrovni kengaytirishga hissa qo‘shishdir. Ushbu ish nafaqat texnologik jihatdan, balki global muloqotning kelajagini shakllantirish nuqtai-nazaridan ham muhim natijalarga olib kelishi kutilmoqda.

Ko‘p tilli ANTning asosiy muammolari. Ma’lumotlarning yetishmasligi, tillarning aralashuvi, texnik cheklovlar va nutqning tabiiy jihatlarini yo‘qotish kabi omillar foydalanuvchi tajribasini pasaytiradi va tizimlarning samaradorligini cheklaydi bu esa ko‘p tilli ANT tizimlari uchun bir qancha muammolarni keltirib chiqaradi. Quyida ko‘p tilli ANT tizimlarining global miqyosda keng qo‘llanilishiga to‘sqinlik qiladigan asosiy muammolar yoritib beriladi.

1. **Ma’lumotlarning cheklanganligi.** Ko‘p tilli ANT tizimlari yuqori sifatli audio va matn ma’lumotlariga tayanadi. Biroq, dunyodagi 7000 dan ortiq tillarning aksariyati, ayniqsa kam resursli tillar (masalan, o‘zbek, qirg‘iz, suahili) uchun yetarli ma’lumotlar mavjud emas. Bu modellar o‘qitilishi uchun zarur bo‘lgan katta hajmdagi datasetlarning yo‘qligi bilan bog‘liq. Shuning uchun ko‘p tillarda yozma matnlar raqamlashtirilmagan, audio yozuvlar esa professional tarzda to‘planmagan. Masalan, o‘zbek tilida nutqni tanib olish uchun millionlab soatlik audio ma’lumotlar kerak bo‘lsa-da, mavjud resurslar minglab soatlar bilan cheklangan. Ma’lumotlarning kamligi tufayli modellar umumlashtirish qobiliyatini yo‘qotadi va faqat ma’lum bir kontekst yoki dialektda yaxshi ishlaydi. Masalan, toshkent shevasida o‘qitilgan model namangan shevasini tushunmasligi mumkin.

2. **Tillarning o‘zaro aralashuvi (Code-Switching).** Ko‘p tilli muhitda foydalanuvchilar bir gap ichida bir nechta tilni aralashtirishi odatiy holat. Masalan, “*Men bugun meetingga boraman*” (o‘zbek va ingliz tillari aralashmasi). ANT

tizimlari bunday holatlarda qaysi tilga ustunlik berishni bilmay qoladi. Chunki modellar odatda bitta til uchun optimallashtiriladi va til chegaralarini aniqlashda qiynaladi. Bundan tashqari, o‘qitish ma’lumotlari ko‘pincha “sof” tillarga asoslanadi, aralash nutq esa kam uchraydi. Natijada tizim noto‘g‘ri transkripsiya qiladi yoki umuman javob bera olmaydi. Masalan, “*Salom, how are you?*” degan gapni tanib olishda model “*Salom*”ni o‘zbekcha, qolgan qismini inglizcha deb farqlay olmay, xato qiladi. Yana boshqa misol, Hindistonda hind va ingliz tillarini aralashtirib gapiradigan foydalanuvchilar uchun ANT tizimlari ko‘pincha chalkashib qoladi, chunki til o‘tishlari oldindan bashorat qilinmagan.

3. Yuqori latentsiya va xatolarning to‘planishi. An’anaviy ko‘p tilli ANT tizimlari bir nechta bosqichlardan iborat: nutqni tanib olish (ANT), matnni qayta ishlash (Large language model, LLM), va nutqqa aylantirish (Text to speech, TTS). Har bir bosqichda vaqt sarflanadi va xatolar yig‘iladi. Sababi har bir modul alohida ishlaydi va ma’lumotlar uzatilishi jarayonida kechikishlar yuzaga keladi. Masalan, nutqni matnga aylantirishda kichik xato keyingi bosqichda katta muammoga aylanadi. Real vaqtda suhbat qilish imkoni cheklanadi. Masalan, bir foydalanuvchi o‘zbek tilida savol bersa, tizim uni ingliz tilida noto‘g‘ri talqin qilib, javobni boshqa kontekstda berishi mumkin. Holbuki bu latentsiya esa suhbatning tabiiyligini yo‘qotilishiga olib keladi.

4. Paralingvistik ma’lumotlarning yo‘qolishi. Nutqni matnga aylantirish jarayonida ohang, intonatsiya, emotsiya va urg‘u kabi paralingvistik elementlar hisobga olinmaydi. Bu esa ko‘p tilli ANT tizimlarini “sovuq” va tabiiy bo‘lmagan qiladi. Chunki ko‘p tizimlar faqat fonetik transkripsiyaga e’tibor beradi, lekin nutqning hissiy yoki kontekstual jihatlari uchun maxsus o‘qitilmagan bo‘ladi. Masalan, “*Yaxshi*” so‘zi istehzoli yoki samimiy ohangda aytilganini farqlash qiyin. Natijada foydalanuvchi niyatini noto‘g‘ri tushunish yoki javobning noadekvat bo‘lishi kelib chiqadi. Masalan, “*Nima bo‘ldi?*” degan savol g‘azab bilan yoki qiziqish bilan aytilishi mumkin, lekin tizim buni ajrata olmaydi. Misol ingliz tilida “*Sure*” so‘zi turli intonatsiyalar bilan ijobiy yoki salbiy ma’no bildirishi mumkin, ammo ANT tizimlari buni matnda aks ettira olmaydi.

5. Dialektlar va aksentlarning turli-tumanligi. Har bir til ichida dialektlar va aksentlar mavjud bo‘lib, ular nutqni tanib olishni qiyinlashtiradi. Masalan, o‘zbek tilida Xorazm shevasi Toshkent shevasidan sezilarli darajada farq qiladi. Chunki Modellar ko‘pincha standartlashtirilgan til versiyalarida o‘qitiladi, lekin real hayotda odamlar o‘ziga xos talaffuz va so‘z birikmalaridan foydalanadi. Buning natijasida dialekt yoki aksentni tushunolmagan tizim noto‘g‘ri transkripsiya qiladi yoki umuman ishlamaydi. Misol tariqasida O‘zbek tilida “kel” so‘zi Xorazmda esa “gal” deb talaffuz qilinishi mumkin, bu esa tizimni chalkashtiradi.

Ko‘p tilli ANT muammolari uchun yechimlar. Ko‘p tilli modellar, end-to-end yondashuvlar, paralingvistik tahlil va dialektga moslashuv kabi usullar tizimlarning aniqligini oshiradi. Latentsiyani kamaytiradi va foydalanuvchi tajribasini yaxshilaydi. Quyida ko‘p tilli ANTning asosiy muammolarini hal qilishga qaratilgan asosiy yondashuv va yechimlar keltiriladi.

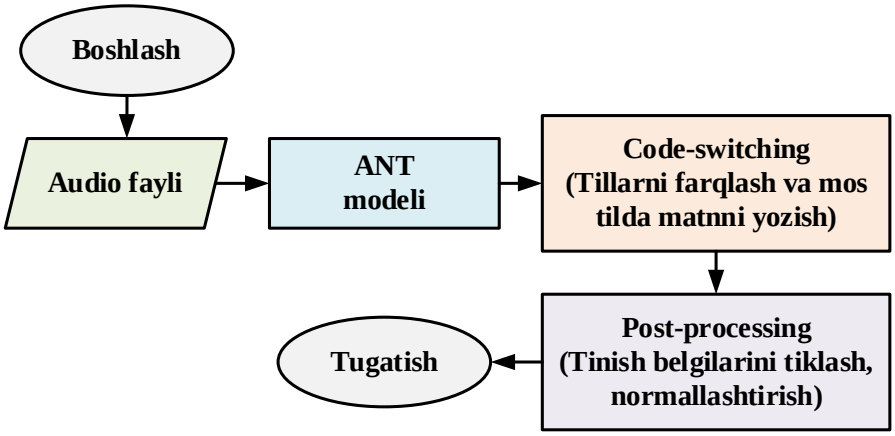
1. Ko‘p tilli modellar va sintetik ma’lumotlar yaratish. Kam resursli tillar uchun ma’lumotlar yetishmasligini bartaraf etish uchun ko‘p tilli modellar birgalikda o‘qitiladi va sintetik ma’lumotlar generatsiyasi qo‘llaniladi. Masalan, J. Devlin va boshqalar o‘z tadqiqotida ko‘p tilli pre-training usulidan foydalanib, kam resursli tillarda matn tahlilining aniqligini 15-20% ga oshirgan[3]. O‘zbek tili uchun ham bundan foydalanilsa, ingliz tilidagi katta ma’lumotlardan transfer learning orqali o‘zbekcha audio tahlilini yaxshilash mumkin. A. Baevski va boshqalar o‘z-o‘zini nazorat qiluvchi o‘qitish usuli bilan oz miqdordagi belgilangan ma’lumotlardan foydalanib, WERni 30% dan 15% ga tushirgan[4]. Bu bilan o‘zbek tili uchun, mavjud 1500 soatlik audio bilan modelni samarali o‘qitish imkoniyati paydo bo‘ladi. G. Hinton va boshqalar sintetik nutq generatsiyasi orqali ma’lumotlar hajmini 50% ga oshirib, nutq sintezini tabiiyligini yaxshilagan[7]. Xuddi shunday o‘zbek tili uchun, sintetik o‘zbekcha audio yaratib, ma’lumotlar bazasini kengaytirish mumkin. Quyida eng mashhur ko‘p tilli modellarga misollar keltirilgan:

1-jadval. Eng keng tarqalgan ko‘p tilli modellar

Model nomi	Ishlab chiqqan kompaniya	Qo‘llab-quvvatlaydigan tillar soni
mBERT (Multilingual BERT)	Google	~104 ta til
XLM-R (XLM-RoBERTa)	Facebook AI (Meta)	100+ til
mT5 (Multilingual T5)	Google	101 ta til
BLOOM	BigScience	46+ til
mGPT (Multilingual GPT)	Sber AI	40+ til
ByT5	Google	Har qanday til
NLLB-200 (No Language Left Behind)	Meta AI	200+ til
SeamlessM4T	Meta AI	Matn va nutq uchun
LaBSE (Language-agnostic BERT Sentence Embedding)	Google	109 til
GPT-4 Turbo (ko‘p tilli imkoniyatlar)	OpenAI	50+ til

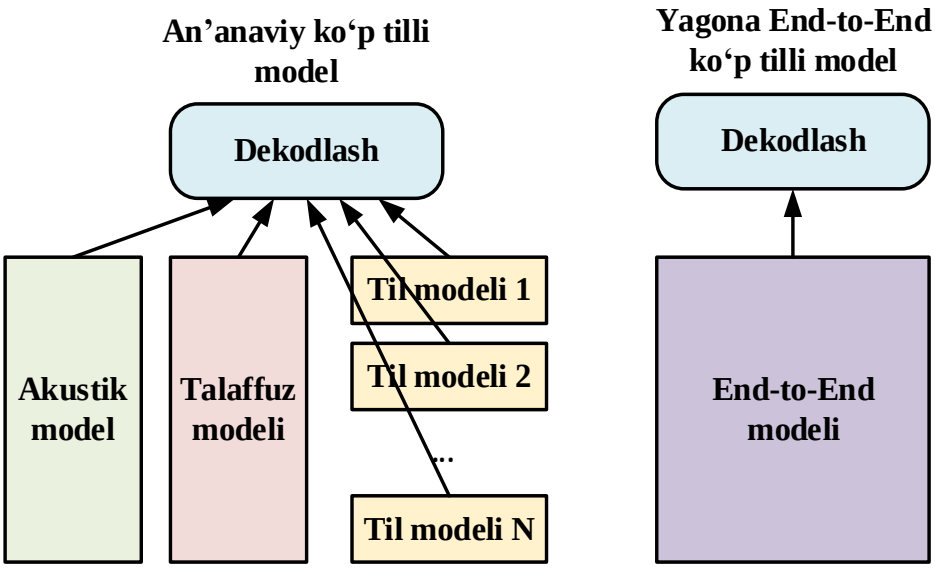
2. Code-Switching uchun maxsus modellar. Tillarni aralashtirib gapirishni aniqlash uchun maxsus ko‘p tilli modellar ishlab chiqiladi. Bunga misol qilib D. Serdyuk va boshqalar end-to-end modellar yordamida hind-ingliz code-switching holatlarida WERni 25% ga kamaytirgan[11]. Bu usuldan foydalanish orqali, “Men meetingga boraman” kabi aralash gaplarni aniq tanib olinishi mumkin. G. Winata va boshqalar ikki tilli so‘z vektorlaridan foydalanib, til o‘tishlarini

aniqlashda 20% aniqlik oshirgan[16]. Bu esa o‘zbek va ingliz tillari aralashmasini tahlil qilishda samaradorlikni oshirish mumkin. T. Schultz va boshqalar ko‘p tilli akustik modellar bilan aralash nutqni tanib olishda 15% yaxshilanishga erishgan[10]. O‘zbek tili uchun bunday foydalansak, o‘zbek-rus yoki o‘zbek-ingliz aralashmalarini yaxshiroq boshqarish imkoni bo‘ladi.



2-rasm. Code-switchingga asoslangan ko‘p tilli ANT ning umumiy arxitekturasi

3. End-to-End Speech Language Models (SpeechLMs). Nutqdan nutqqa to‘g‘ridan-to‘g‘ri o‘tadigan modellar latentsiyani kamaytiradi va xatolarni minimallashtiradi. Chunonchi A. Gulati va boshqalar Conformer modeli yordamida latentsiyani 20-30% ga qisqartirib, WERni 10% ga kamaytirgan[6]. Bu orqali real vaqtda o‘zbekcha suhbatlarda ham kechikishni kamaytirish mumkin. Y. Wang va boshqalar SpeechLM bilan nutqdan nutqqa o‘tishda xatolarni 15% ga qisqartirgan[13]. Bu O‘zbekcha savol-javob tizimlarini tabiiyroq va tezkor qilish imkoni paydo qiladi. S. Watanabe va boshqalar end-to-end ANT bilan ko‘p tilli tizimlarda latentsiyani 25% ga kamaytirgan[14]. O‘zbek tili uchun bundan foydalanilsa, o‘zbek tilidagi tarjimani real vaqtda yaxshilash imkoni yaratiladi.



3–rasm. An’anaviy va End-to-End ko‘p tilli modellarining umumiy tuzilishi.

4. Emotsiya va intonatsiyani tanib oluvchi modellar. Nutqning ohang va emotsiyasini saqlab qolish uchun paralingvistik tahlil qo‘shiladi. F. Weninger o‘z tadqiqotida deep learning yordamida nutqdan emotsiyani aniqlashda 85% aniqlikka erishgan[15]. O‘zbek tili uchun bunday foydalansak, “Yaxshi” so‘zining istehzoli yoki samimiy ohangini farqlash mumkin. Xususan B. McFee o‘z tadqiqotida audio signalning pitch va energy xususiyatlarini tahlil qilib, paralingvistik ma’lumotlarni 90% aniqlikda saqlagan[9]. O‘zbek tili uchun bu nutqning intonatsiyasini matnga o‘tkazish imkoni berishi mumkin. E. Cambria va b. emotsional tahlil bilan nutqni tushunishda 20% yaxshilanishga erishgan[5].

5. Moslashuvchan modellar. Dialekt va aksentlarni tanib olish uchun maxsus datasetlar bilan moslashuvchan modellar o‘qitiladi. Masalan D. Snyder va b. X-Vector texnologiyasi bilan aksentlarni aniqlashda 95% aniqlikka erishgan[12]. Bu yondashuv bizga toshkent va xorazm shevalarini farqlash imkonini berishi mumkin. M. Karafiat va b. bottleneck xususiyatlari bilan dialekt tanib olishda 30% yaxshilanishga erishgan. Bu orqali turli shevalarni tahlil qilishda aniqlik oshadi. deep neural networks bilan aksentlarga moslashishda 25% samaradorlikka erishgan[8].

Yuqoridagi yechimlar yordamida o‘zbek tili uchun ushbu usullarni qo‘llash orqali ko‘p tilli ANT tizimlarini ishlab chiqish va samaradorligini oshirish, kam resursli til sifatidagi cheklovlarni bartaraf etish va foydalanuvchi tajribasini yaxshilash mumkin.

Foydalanilgan adabiyotlar:

1. Mukhamadiyev, A., Mukhiddinov, M., Khujayarov, I., Ochilov, M., & Cho, J. (2023). Development of Language Models for Continuous Uzbek Speech Recognition System. *Sensors*, *23*(3), 1145. https://doi.org/10.3390/ s23031145

2. Musaev, M., Khujayorov, I., Ochilov, M. (2021). Automatic Recognition of Uzbek Speech Based on Integrated Neural Networks. In: Aliev, R.A., Yusupbekov, N.R., Kacprzyk, J., Pedrycz, W., Sadikoglu, F.M. (eds) 11th World Conference “Intelligent System for Industrial Automation” (WCIS-2020). WCIS 2020. Advances in Intelligent Systems and Computing, vol 1323. Springer, Cham. https://doi.org/10.1007/978-3-030-68004-6_28

3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv. https://arxiv.org/abs/1810.04805

4. Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv. https://arxiv.org/abs/2006.11477

5. Cambria, E., Schuller, B., Xia, Y., & Havasi, C. (2017). Sentic computing: A common-sense-based framework for concept-level sentiment analysis. IEEE Computational Intelligence Magazine, 12(2), 14–23. https://doi.org/10.1109/MCI.2017.2670544

6. Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ... & Pang, R. (2020). Conformer: Convolution-augmented Transformer for speech recognition. arXiv. https://arxiv.org/abs/2005.08100

7. Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 29(6), 82–97.

8. Karafiát, M., Burget, L., Cernocký, J., & Grezl, F. (2016). Multilingual bottleneck features for language and dialect recognition. Computer Speech & Language, 46, 252–267. https://doi.org/10.1016/j.csl.2017.03.002

9. McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and music signal analysis in Python. Proceedings of the 14th Python in Science Conference, 8, 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

10. Schultz, T., Waibel, A., & others. (2006). Multilingual and crosslingual speech recognition. In Handbook of Multilingualism and Multilingual Communication (pp. 281–298). De Gruyter Mouton.

11. Serdyuk, D., Wang, Y., Bengio, Y., & others. (2018). Towards end-to-end spoken language understanding. arXiv. https://arxiv.org/abs/1802.08395

12. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375

13. Wang, Y., Huang, P. S., Gu, J., & others. (2022). SpeechLM: Enhanced speech pre-training for spoken language understanding. arXiv. https://arxiv.org/abs/2207.04672

14. Watanabe, S., Hori, T., Kim, S., Hershey, J. R., & Hayashi, T. (2018). Hybrid CTC/attention architecture for end-to-end speech recognition. IEEE Journal of Selected Topics in Signal Processing, 11(8), 1240–1253. https://doi.org/10.1109/JSTSP.2017.2763455

15. Weninger, F., Eyben, F., Schuller, B., Mortillaro, M., & Scherer, K. (2013). On the acoustics of emotion in audio: What speech signals reveal about our feelings. Frontiers in Psychology, 4, 738. https://doi.org/10.3389/fpsyg. 2013.00738

16. Winata, G. I., Cahyawijaya, S., Lin, Z., Liu, Z., & Fung, P. (2019). Code-switching language modeling with bilingual word embeddings. arXiv. https://arxiv.org/abs/1909.03460

17. Radford, A., Gao, L., Behrmann, E., Brockman, G., & Sutskever, I. (2022). *Whisper: Robust speech recognition via large-scale weak supervision*. OpenAI. https://openai.com/research/whisper

18. Zhang, X., Fan, A., Siddhant, A., Koehn, P., & Chaudhary, V. (2023). *SeamlessM4T: Massively multilingual & multimodal machine translation*. Meta AI Research. https://arxiv.org/abs/2308.04620

BERT MODELI YORDAMIDA MATNLARNI TASNIFLASH

Alayev Ruhillo Habibovich,
Texnika fanlari bo‘yicha falfasa doktori PhD, Dotsent
mr.ruhillo@gmail.com
O‘zMU

Berdiyev Jahongir Botir o‘g‘li,
Kompyuter lingvistikasi mutaxassisligi magistranti
berdiyevjahongir94@gmail.com
ToshDO‘TAU

Annotatsiya. Hozirgi kunda tabiiy tilni qayta ishlash (NLP) sohasi sun’iy intellektning eng jadal rivojlanayotgan yo‘nalishlaridan biri hisoblanadi. Bu sohada erishilayotgan yutuqlar turli tildagi matnlarni avtomatik tarzda tahlil qilish, tushunish va tasniflash imkonini bermoqda. Ayniqsa, chuqur o‘rganishga asoslangan modellar, xususan, BERT (Bidirectional Encoder Representations from Transformers) modeli ushbu jarayonlarda samaradorlikni sezilarli darajada oshirdi. Google kompaniyasi tomonidan ishlab chiqilgan ushbu model til birliklarini kontekstual tarzda, ya’ni ikki tomonlama yo‘nalishda tahlil qilish orqali yuqori aniqlikka erishadi. Mazkur maqolada BERT modelidan foydalanilgan holda matnlarni muayyan kategoriyalar bo‘yicha tasniflash masalasi ko‘rib chiqiladi. Tadqiqotda modelni o‘qitish, sinovdan o‘tkazish va baholash bosqichlari amalga oshiriladi. Natijalar asosida BERT modelining matn tasnifi sohasidagi samaradorligi tahlil qilinadi va tegishli xulosalar chiqariladi.

Abstract. Currently, Natural Language Processing (NLP) is one of the most rapidly developing fields within artificial intelligence. Advances in this domain enable automatic analysis, comprehension, and classification of texts in various languages. Deep learning models, particularly BERT (Bidirectional Encoder Representations from Transformers), have demonstrated high effectiveness in this regard. Developed by Google, BERT analyzes linguistic units in a contextual manner, processing information bidirectionally, which contributes to its high accuracy. This research focuses on the task of classifying texts into specific categories using the BERT model. The study involves the implementation of model training, testing, and evaluation stages. Based on the obtained results, the effectiveness of BERT in text classification tasks is analyzed, and relevant conclusions are drawn.

Аннотация. В настоящее время обработка естественного языка (NLP) является одним из наиболее динамично развивающихся направлений искусственного интеллекта. Достижения в этой области позволяют автоматически анализировать, понимать и классифицировать тексты на

различных языках. Особенно эффективно в этом контексте показали себя модели глубокого обучения, в частности, модель BERT (Bidirectional Encoder Representations from Transformers). Разработанная компанией Google, данная модель анализирует языковые единицы в контекстуальном виде, то есть в двух направлениях, что обеспечивает высокую точность. В данной научной работе рассматривается задача классификации текстов по определённым категориям с использованием модели BERT. В ходе исследования реализуются этапы обучения модели, тестирования и оценки. На основе полученных результатов проводится анализ эффективности модели BERT в задачах текстовой классификации и формулируются соответствующие выводы.

Katil so‘zlar. *Tasniflash, tokenizatsiya, matnlarni tozalash, tokenizator, fine tuning.*

Bilamizki, ko‘p resursli tillarda, jumladan, ingliz tilida til modellarida bajariladigan vazifalar samarali ancha samarali natijalar bermoqda. Bu esa bir tomondan tillarning ko‘p resursliligi bo‘lsa, boshqa tomondan ularning morfologik qurilishi bilan bog‘liq. Kam resursli tillarni modellashtirish esa yetarlicha aniq natija bermaydi. Kam resursli til bo‘lishiga qaramay, hozirgi kunga kelib o‘zbek tili ham modellashtirishda yaxshi natijalar bermoqda. Ammo shunga qaramay ma’lumotlarning kamligi hamda o‘zbek tilining morfoligik qurulish jihatdan agglyutinativ til ekanligi o‘zbek tilini modellashtirishda ba’zi muammolarni keltirib chiqaradi. Chunki ingliz tili kabi flektiv tillardan farqli ravishda o‘zbek tili kabi agglyutinativ tillarning so‘z tartibida aniqlikning yo‘qligi, qo‘shimchalarning qo‘shilish tartibi hamda so‘zga qo‘shimcha qo‘shish natijasida asos va qo‘shimchada tovush o‘zgarish hodisalari mavjud, bu esa o‘zbek tilini modellashtirishda qiyinchiliklarni keltirib chiqaradi. O‘zbek tilining boshqa tillardan farqini quyidagicha farqlash mumkin:

1. **Agglutinativlik:** O‘zbek tili so‘z va so‘z shakli yasashda qo‘shimchalar orqali yangi ma’nolar hosil qiladi. Masalan, “kitob” so‘zidan “kitobxon”, “kitoblar” kabi shakllar hosil bo‘ladi. Bu holat esa matnlarni tokenlash jarayonida ayrim muammoli vaziyatni keltirib chiqaradi. Chunki tokenlash jarayonida har bir so‘zni yoki belgini bitta token sifatida belgilash juda katta xotira talab qiladi. Buning oldini olish uchun tokenlarga ajratishda ko‘p qo‘llaniladigan belgilar va so‘zlar tokenlashtiriladi. Shundan kelib chiqib, “kitob”, “kitobxon”, “kitoblar” so‘zlarining har birini emas, balki faqat “kitob” so‘zini token sifatida ajratish samarali hisoblanadi.

2. Lug‘at tarkibi: O‘zbek tilida arabcha, forscha, ruscha va boshqa tillardan olingan so‘zlar ko‘p. Bu esa sof o‘zbek tilidagi so‘zlarni ajratishni qiyinlashtirsa, ikkinchi tomondan gap qurilishi va so‘z yasalishi, ularga qo‘shimcha qo‘shilishdagi

ba’zi istisnolarni keltirib chiqaradi. Bu esa o‘z navbatida o‘zbek tilidagi matnlarni qayta ishlashdagi jarayonini murakkablashtiradi.

3. Sintaksis: Gap tuzilishi ba’zan boshqa tillardan sezilarli darajada farq qiladi.

Agar BERT modeli O‘zbek tiliga moslashtirilsa, u o‘zbek tilidagi matnlarni yaxshiroq tahlil qilishi, tarjima qilish, savol-javob tizimlarini yaxshilash va boshqa ko‘plab sohalarda qo‘llanilishi mumkin bo‘ladi.

BERT modelini O‘zbek tiliga moslashtirish jarayoni bir necha qadamdan iborat. Quyida ular batafsil tushuntiriladi.

Ma’lumotlarni to‘plash. Ma’lumotlarni to‘plash va katta hajmli ma’lumotlar bazasi NLP muammolarini yechishning samarali yo‘lidir. Har qanday til modellari beradigan natijalarining aniqligi ma’lumotlar bazasining hajiga bog‘liq. Jumladan, BERT Modelini o‘zbek tiliga o‘qitish uchun ham katta hajmdagi ma’lumotlar to‘plami kerak. Bu ma’lumotlar O‘zbek tilidagi kitoblar, maqolalar, veb-saytlar, ijtimoiy tarmoqlar postlari va boshqa matn manbalaridan olinadi. Biz ma’lumotlar to‘plami sifatida *daryo.uz* veb saytidagi maqolalarni kategoriyalar bo‘yicha ajratib to‘pladik va ushbu maqolalardan ma’lumotlar to‘plami sifatida foydalandik. Ma’lumotlar to‘plamini tayyorlashda e’tibor qilish kerak bo‘lgan jihati shundaki, matnlarni tasniflash uchun yig‘ilgan ma’lumotlar bazasi tarkibidagi kategoriyalar hajmi jihatidan bir biriga teng bo‘lishi kerak. Agar kategoriyalar orasidagi muvozanat buzilsa, dastur bir tomonga og‘ib ketishi va dastur xato ishlay boshlashi mumkin.

Matnlarni tozalash. Bu ma’lumotlar dasturlar to‘g‘ri va sifatli ishlashi uchun, avvalo, uni tozalab olish kerak. Matnlarni tozalashning bir qancha usullari mavjud. Matnlarni avtomatik tozalash uchun maxsus dasturlardan foydalanish mumkin, yoki inson qo‘l mehnati bilan tozalash mumkin. Qo‘l mehnati bilan tozalash katta hajmli matnlar uchun juda ko‘p vaqt va xarajat talab qiladi. Ammo sifatli tozalash imkonini beradi.

Biz matnlarni tozalash uchun o‘zbek tilidagi matnlarni tozalovchi dasturdan foydalandik. Tozalash jarayonida HTML kodlari, reklama matnlari, turli xil emojilar olib tashlandi va takroriy jumlalarning faqat bittasi olib qolindi.

Tasniflash jarayoni. Keyingi bosqich matnlarni tasniflash jarayonidir. Matnlarni tasniflashning bir qancha turi bor. Matnni tasniflash ijtimoiy media postlarining kayfiyatini tahlil qilishda, ayniqsa, nafrat nutqi kabi salbiy his-tuyg‘ularni aniqlashda ham qimmatlidir. Mashinani o‘rganish modelidan foydalanib, matnni haqoratli til va nafrat nutqi uchun tasniflash va kuzatish mumkin[1]. Biz matnlarni turli kategoriyalarga ajratishni ko‘rib chiqamiz. Buning uchun to‘plangan matnlarni kategoriyar bo‘yicha ajratib olish kerak. *Daryo.uz* veb saytidan to‘plangan

maqolalar quyidagi 15ta kategoriyaga ajratildi va tasniflash uchun tayyorlandi: “Madaniyat”, “Qonunchilik”, “Sport”, “Salomatlik”, “Dunyo”, “O‘zbekiston”, “Jamiyat”, “Foto”, “Ayollar”, “Texnologiya”, “Siyosat”, “Iqtisodiyot”, “Pazandachilik”, “Jinoyat”, “Avto”.

Matnlarni tasniflash uchun ishlatiladigan matnlar to‘plamining hajmi juda katta bo‘ladi. Bu esa matnlarga ishlov berishni qiyinlashtiradi. Bundan tashqari, BERT modeli o‘ta murakkab struktura bo‘lgani uchun katta hisoblash resurslariga muhtoj. Shuning uchun matnlarni tasniflashda GPU (Grafik protsessor)dan foydalanildi. Quyidagi rasmda buni ko‘rish mumkin(2.2.1-rasm):

```
import tensorflow as tf

gpus = tf.config.experimental.list_physical_devices('GPU')
if gpus:
    try:
        for gpu in gpus:
        tf.config.experimental.set_memory_growth(gpu, True)
    print("GPU aniqlangan va ishlatilmoqda.")
 except RuntimeError as e:
     print(e)
else:
    print("GPU topilmadi. CPU ishlatilmoqda.")
```

1-rasm. GPU (grafik protsessor)ning mavjudligini tekshiruvchi va undan foydalanishga ruxsat beruvchi kod

Ma’lumotlarni BERT modeli uchun mos formatga keltirish uchun oldindan qayta ishlash funksiyasini (preprocess_data) aniqlaymiz. Bu funksiya matnlarni tokenlashtiradi, padding qo‘shadi va yorliqlarni kodlaydi.

BERT modeli orqali tokenlash uchun quyidagi kodni ko‘rib chiqish mumkin:

```
from transformers import AutoTokenizer
tokenizer = AutoTokenizer.from_pretrained("bert-base-multilingual-cased")
```

Bu yerda AutoTokenizer BERT modeliga mos keladigan mos keladigan tokenizatorni avtomatik yuklash uchun ishlatiladi. AutoTokenizer transformer modellari uchun maxsus ishlab chiqilgan bo‘lib, u turli modellarga mos ravishda **tokenizatsiya** jarayonini amalga oshiradi. Tokenizatsiya bu matnni kichik qismlarga ajratish jarayonidir. Transformer modellar so‘zlarni emas, balki **tokenlar** bilan ishlaydi, shuning uchun har bir matn modelga tushunarlilik darajasida kodlanishi kerak. AutoTokenizer orqali tokenizatsiya qilingan matn maxsus **token**

identifikatorlari orqali ifodalanadi. U turli xil tokenizatorlarni qo‘llab-quvvatlaydi va ishlayotgan modelga mos ravishda eng samarali usulni tanlaydi.

tokenizer = AutoTokenizer.from_pretrained(“bert-base-multilingual-cased”) – bu kod ko‘p tilli BERT modeli uchun oldindan o‘qitilgan tokenizatorni yuklaydi (“**bert-base-multilingual-cased”** modeli **100dan ortiq tilni,** jumladan, **o‘zbek tilini** ham qo‘llab-quvvatlaydi). Ushbu tokenizator matnlarni qismlarga bo‘lib, IDlarga o‘giradi. “**Cased” versiya** bo‘lgani uchun **“bert-base-multilingual-cased”** katta-kichik harflarni farqlaydi.

Tokenizatsiyada **WordPiece** yoki **SentencePiece tokenizatsiyalaridan foydalaniladi. Lekin bu yerda** bert-base-uncased modeli ishlatilganligi uchun WordPiece tokenizatsiyasi ishlatilgan. Chunki **bert-base-multilingual-cased modeli WordPiece tokenizatsiyasidan foydalanadi. Bundan tashqari, o‘zbek tilidagi matnlarni tasniflashda WordPiece tokenizatsiyasi samarali hisoblanadi.**

O‘zbek tilidagi matnlarni tasniflashda WordPiece tokenizatsiyasi samarali bo‘lishi mumkin, lekin uning natijaviyligi matnning o‘ziga xos tuzilishi va so‘zlarning murakkabligi kabi omillarga bog‘liq bo‘ladi. Ba’zi hollarda SentencePiece ham samarali bo‘lishi mumkin, ayniqsa, o‘zbek tilining agglutinativ xususiyatlarini hisobga olsak.

Shuni ham ta’kidlash kerakki, o‘zbek tili agglutinativ til bo‘lgani uchun so‘zlarni qismlarga ajratish juda muhim. Chunki ishlatilayotgan model har bir so‘z shaklini alohida lug‘at birligi sifatida saqlasa, juda ko‘p xotira talab qiladi. Bu esa nafaqat joy, balki mablag‘ sarfini ham oshiradi. Shu sababli modellar, asosan, lug‘at sifatida so‘zlarning asosiy qismini saqlaydi. Tokenizatsiya jarayonida esa model lug‘atida mavjud bo‘lmagan, ammo unga o‘xshash so‘zlarni alohida tokenlarga ajratib hisoblaydi. Masalan, “kitoblarimizdan” so‘zi “kitob”, “lar”, “imiz” va “dan” qismlariga ajratiladi. Model lug‘atida esa faqat “kitob” so‘zi saqlanadi.

Shundan kelib chiqib, o‘zbek tili uchun WordPiece tokenizatsiyasidan foydalanish samarali ekani ta’kidlanadi. Ushbu tokenizatsiya so‘zlarning asos qismini lug‘at sifatida ajratib, qo‘shimchalarni alohida token sifatida tahlil qiladi.

Odatda, til modellari kirish ketma-ketligini bir yo‘nalishda o‘qiydi: chapdan o‘ngga yoki o‘ngdan chapga. Maqsad keyingi so‘zni bashorat qilish/yaratish bo‘lsa, bunday bir yo‘nalishli trening yaxshi ishlaydi. Ammo til kontekstini chuqurroq anglash uchun BERT ikki tomonlama treningdan foydalanadi. Ba’zan u “yo‘nalishsiz” deb ham ataladi. Shunday qilib, u bir vaqtning o‘zida oldingi va keyingi tokenlarni hisobga oladi. BERT Transformatorning ikki tomonlama treningini tilni modellashtirishda qo‘llaydi, matn tasvirlarini o‘rganadi[2]. BERT modelini maxsus vazifalar uchun moslashtirishda ikkita yondashuv mavjud: modelni noldan oldindan o‘qitish yoki oldindan o‘qitilgan modeldan foydalanib, uni fine-tuning qilish. Modelni noldan oldindan o‘qitish katta hajmdagi ma’lumotlar to‘plami va yuqori hisoblash

quvvatiga ega qurilmalarni talab qiladi. Ushbu jarayon vaqt va resurs jihatidan ancha talabchan bo‘lgani uchun, odatda, maxsus vazifalar, jumladan, matnlarni tasniflash kabi masalalar uchun oldindan o‘qitilgan BERT modellaridan foydalaniladi. Keyin esa ushbu model fine-tuning orqali yangi ma’lumotlarga moslashtiriladi, ya’ni modelning barcha yoki ba’zi qatlamlari qayta o‘qitilib, tasniflash vazifasi uchun optimallashtiriladi.

Maxsus vazifalar uchun moslashtirish(fine-tuning). BERT kirish matnini yuqori o‘lchamli bo‘shliqda aks ettirish uchun ko‘p qatlamli ikki tomonlama transformator kodlovchisidan foydalanadi. Ya’ni, u jumladagi har bir so‘zning butun kontekstini hisobga olishi mumkin, bu esa matnning ma’nosini yaxshiroq tushunishga yordam beradi[3]. Oldindan o‘qitilgan **BERT modeli**ni o‘zimizga kerakli vazifaga moslashtirish uchun **fine-tuning** amalga oshiriladi. Quyidagi kodda “**bert-base-multilingual-cased**” modeli yuklanib, **ko‘p tegli klassifikatsiya (multi-label classification)** muammosi uchun moslashtiriladi. Model **yorliqlar soniga (num_labels)** qarab sozlanadi va har bir yorliqning **id bilan mosligi (id2label, label2id)** belgilanadi:

```
from transformers import AutoModelForSequenceClassification
model = AutoModelForSequenceClassification.from_pretrained(
    "bert-base-multilingual-cased",
    problem_type="multi_label_classification",
    num_labels=len(labels),
    id2label=id2label,
    label2id=label2id
)
```

Bu jarayon **oldindan o‘qitilgan modelni maxsus vazifaga moslashtirishga** yordam beradi va **modelni fine-tuning qilishga tayyorlaydi.** trainer.train() **kodi orqali fine-tuning qilish boshlanadi va o‘quv ma’lumotlari asosida sozlanadi.**

Maxsus vazifalar uchun moslashtirish (fine-tuning) jarayoni tugagandan so‘ng, modelning natijalari baholanadi. Bunda o‘qitilgan modelning samaradorligini tekshirish uchun **sinov (test) ma’lumotlari** ustida baholash amalga oshiriladi. Baholash jarayonida **F1-score, ROC-AUC va aniqlik (accuracy)** kabi metrikalar qo‘llanilib, modelning tasniflash sifati aniqlanadi. Agar model natijalari kutilgan darajada bo‘lmasa, **giperparametrlarni sozlash (hyperparameter tuning) yoki ma’lumotlarni kengaytirish (data augmentation)** kabi usullar bilan natijalarni yaxshilash mumkin.

XULOSA

Xulosa qilib shuni aytish mumkinki, BERT modelining yaratilishi tabiiy tilni qayta ishlashda katta inqilob bo‘ldi. Shuningdek, BERT modelining oldindan o‘qitilganligi turli vazifalar uchun moslashtirishni osonlashtiradi. Shuningdek, BERT modeli matnlarni tasniflash, his-tuyg‘ularni tahlil qilish, xabarlarning rost yoki yolg‘onligini aniqlash kabi vazifalarda samarali vosita hisoblanadi. Bundan

tashqari, BERT yordamida tokenizatsiya qilish uchun turli tillarning grmmatik qurilishi uchun mos ravishda tokenizatorlar mavjudligi bilan ham bu model qulay hisoblanadi.

Foydalanilgan adabiyotlar:

1. Khang Pham (2023). Text classification with BERT
2. Shre Varsheni (2024). Why and how to use BERT for NLP Text Classification? *Analytics Vidhya*
3. https://medium.com/@khang.pham.exxact/text-classification-with-bert-7afaacc5e49b

BERTOPIC: NEYRON TEMATIK MODELLASHTIRISH USULI

Narzillo Aloyev Raxmatilloyevich,
Tayanch doktorant
vip.alayev@gmail.com
ToshDO‘TAU

Annotatsiya. Tematik modellashtirish – bu katta matn to‘plamlarida yashirin tematik tuzilmalarni aniqlashga yordam beradigan tabiiy tilni qayta ishlash usulidir. Yashirin Dirixle taqsimoti (LDA) va manfiy bo‘lmagan matritsa faktorizatsiyasi (NMF) kabi an'anaviy yondashuvlar ko‘p yillar davomida tematik modellashtirish sohasida asosiy o‘rinlarni egallab kelishgan, ammo ular so‘zlar orasidagi semantik munosabatlarni e'tiborsiz qoldiradigan Bag-of-Words ga tayanganligi sababli turxli xil cheklovlarga duch kelishmoqda. BERTopic o‘z navbatida transformatorlarga asoslangan holda til modellarining semantikasini birlashtirish orqali klasterlash usullari va maxsus ishlab chiqilgan sinfga asoslangan TF-IDF protsedurasi bilan birlashtirilgan holda, tematik modellashtirishga yangi yondashuvni taqdim etadi. Ushbu kombinatsiya moslashuvchan bo‘lib adaptatsiyalashgan tizimni saqlagan holda, yanada interpretatsiya qilinadigan mavzularni yaratishga imkon beradi.

Annotation. Topic modeling is a natural language processing technique that helps to identify hidden thematic structures in large text collections all over the world and in any language. Traditional approaches such as LDA and NMF have been the mainstays of thematic modeling for many years, but they suffer from various limitations due to their reliance on “Bag-of-Words”, which ignores semantic relationships between words in texts. BERTopic, in turn, provides a new approach to thematic modeling by combining the semantics of language models based on transformers and a specially designed class-based TF-IDF procedure. This combination allows for the creation of more interpretable topics and understand of many texts in any language.

Аннотация. Тематическое моделирование – это метод обработки естественного языка, который помогает выявлять скрытые тематические структуры в больших текстовых коллекциях. Традиционные подходы, такие как скрытое распределение Дирихле (LDA) и неотрицательная матричная факторизация (NMF), на протяжении многих лет были основой тематического моделирования, но они сталкиваются с различными ограничениями из-за своей зависимости от «Bag-of-Words», который игнорирует семантические связи между словами. BERTopic, в свою очередь, представляет новый подход к тематическому моделированию, объединяя семантику языковых моделей на основе трансформаторов и специально разработанной процедурой TF-IDF на

основе классов. Такое сочетание позволяет создавать более интерпретируемые темы, сохраняя при этом гибкую и адаптируемую систему.

Kalit so‘zlar: BERTopic, BERT, TF-IDF, LDA, NMF, neyron tematik modellashtirish

Mavzuni modellashtirishning an’anaviy usullari tadqiqotchilarga katta hujjatlar to‘plamini tahlil qilishda yordam berishda asosiy rol o‘ynaydi, ammo ular sezilarli cheklovlarga ega:

1) **Bag-of-Words cheklovi**: LDA kabi modellar hujjatlarga kontekstual va semantik munosabatlarni e'tiborsiz qoldirib, tartibsiz so‘zlar to‘plami sifatida qaraydi.
2) **Semantik tushunchaning yetishmasligi:** An'anaviy modellar nozik ma'nolarni va so‘zlar orasidagi munosabatlarni tu’hunishda qiyinchiliklarga duch keladi.
3) **Chekli lug‘at:** Ko‘pgina an’anaviy usullar oldindan belgilangan lug‘atni talab qiladi, bu esa o‘z navbatida ularning moslashuvchanligini cheklab qo‘yadi.

Tabiiy tilni qayta ishlash sohasidagi so‘nggi yutuqlar, xususan, BERT kabi transformatorga asoslangan til modellari hajmga ega vektor bo‘shliqlaridan kontekstual ma'nolarni ajratib olish orqali matn tasvirini yangicha tarzda inqilob qildi. Tematik modellashtirishning bir nechta usullari ushbu qo‘shimchalardan foydalanishga harakat qildi, ammo ko‘pchiliklari mavzularning izchil tasvirlarini yaratishda hali ham qiyinchiliklarga duch kelmoqda. BERTopic hujjatni joylashtirish, klasterlash va mavzuni ko‘rsatish bosqichlarini ajratuvchi modulli tizimni amalga oshirish orqali ushbu muammolarni hal qiladi va uni turli til modellari va foydalanish holatlariga moslashtiradi.

BERTopic metodologiyasi uchta asosiy komponentlardan iborat:

1. **Hujjatlarning vektor tasviri**

Model hujjatlarni hajmga ega vektorli tasvirlarga aylantirish uchun Sentence-BERT (SBERT) dan foydalanadi. SBERT semantik o‘xshashlik vazifalari uchun nozik sozlangan oldindan o‘rgatilgan til modellaridan foydalanadi, bu o‘xshash ma’noga ega bo‘lgan hujjatlarni vektor fazosida bir-biriga yaqin joylashtirish imkonini beradi. Ushbu usulning moslashuvchanligi, MiniLM kabi kichikroq va MPNET kabi kattaroq va kuchliroq modellarga qadar oldindan o‘rgatilgan turli xil til modellaridan foydalanishga imkon beradi.

2. **Hujjatlarni klasterslash**

Hujjatlarning vektor tasviri yaratilganidan so‘ng BERT quyidagilarni amalga oshiradi:

a) PCA yoki t-SNE kabi alternativalarga qaraganda yuqori o‘lchamli ma'lumotlarning mahalliy va global xususiyatlarini yaxshi saqlaydigan UMAP (Uniform Manifold Approximation and Projection) yordamida o‘lchamlarni kamaytirish.

b) HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) yordamida klasterlash ya’ni, har xil zichlikdagi klasterlarni topadi. Ushbu yondashuv bir-biriga bog‘liq bo‘lmagan hujjatlarni har qanday klasterga biriktirilishining oldi olinadi hamda, mavzularning taqdim etilishi sifat jihatidan ancha yaxshilanadi.

UMAP uchun matematik formulani minimallashtirish sifatida quyidagicha ko‘rsatish mumkin:

$$\sum(i,j)\,[\,v_{ij} * log(w_{ij}) + (1 - v_{ij}) * log(1 - w_{ij})\,]$$

Bu yerda v_ij katta o‘lchamli o‘xshashlikni va w_ij kichik o‘lchamli o‘xshashlikni ifodalaydi.

3. Matnning ko‘rinishi

Aynan shu qismda BERTopic o‘zining eng innovatsion yo‘nalishini taqdim etadi: sinflarga asoslangan (class-based) TF-IDF yondashuvi.

An’anaviy TF-IDF hujjatdagi so‘zning hujjatlar to‘plamiga nisbatan ahamiyatini ifodalaydi. BERTopic ushbu kontseptsiyani hujjat darajasida emas, balki klaster darajasida ishlashga moslashtiradi. Sinflarga asoslangan TF-IDF quyidagi tartiblarni o‘z ichiga oladi:

a. **Hujjatlarni yig‘ish**: Klasterdagi barcha hujjatlar bitta “hujjat” ko‘rinishiga keltiriladi

b. **Sinflarga asoslangan atama chastotasi**: atamalarning har bir klasterda uchratilishining chastotasini hisoblash

c. **Hujjatning o‘zgartirilgan teskari chastotasi**: sinfga asoslangan holdagi ko‘rinishini sozlash

c-TF-IDF matematik formulasi quyida keltiriladi:

$$c\text{-}TF\text{-}IDF(t,c,C) = tf(t,c) * idf(t,C)$$

Bu yerda:

- tf(t,c) — c sinfdagi t terminning chastotasi
- idf(t,C) = log(1 + |C|/(1 + df(t,C)))
- |C| - sinflar soni
- df(t,C) - t atamasi mavjud bo‘lgan sinflar soni

Ushbu yondashuv bir qator afzalliklarni taqdim etadi:

- Mavzularni taqdim etishda barcha hujjatlar klasteri hisobga olinadi
- Muayyan mavzuga xos bo‘lgan so‘zlarga ustunlik beradi
- U ma’noli va so‘zlarga boy mavzular ko‘rinishini yaratadi.

c-TF-IDF tasvirlar generatsiya qilinganidan so‘ng, BERTopic o‘z navbatida har bir mavzu uchun eng yuqori ko‘rsatkichga ega n ta so‘zni tanlab oladi.

BERTopic o‘z doirasini dinamik mavzularni modellashtirish uchun kengaytiradi va mavzular vaqt o‘tishi bilan qanday rivojlanishini tahlil qilish imkonini beradi. Quyida yondashuvlar berilgan:

1) Mavzularning global tasvirini yaratish uchun BERTopic butun korpusga qo‘llaniladi
2) Global mavzu strukturasidan foydalangan holda har bir vaqt davri uchun mahalliy ko‘rinishni yaratadi
3) Mavzu evolyutsiyasini modellashtirish uchun har bir bosqichda yaratilgan c-TF-IDF matritsalaridan foydalanadi

Bu bizga har bir vaqt davri uchun butun bir modelni qayta sozlashni talab qilmasdan, vaqt o‘tishi bilan mavzuning tarqalishini va tarkib jihatidan o‘zgarishini kuzatish imkonini beradi.

BERTopic bir nechta ma’lumotlar to‘plamida, jumladan 20 NewsGroups, BBC News va Topicmodel.uz da baholanib ko‘rildi. Asosiy urg‘ular quyidagilarga mos keladi:

1) **Mavzuning ma’nolik darajasi**: BERTopic odatda turli ma’lumotlar to‘plamlari bo‘yicha yuqori ko‘rsatkichga ega, bu esa u yaratgan mavzular semantik jihatdan mazmunli ekanligini ko‘rsatadi.

2) **Mavzu xilma-xilligi**: BERTopic mavzular xilma-xilligida raqobatbardosh bo‘lsa-da, CTM (Kontekstlashtirilgan mavzu modeli) doimiy ravishda ustunlikka ega bo‘lib kelmoqda.

3) **Til modellari orasida barqarorlik belgisi**: BERTopicning ishlash mahsuldorligi SBERT til modellariga nisbatan barqarorligicha qolmoqda, bu uning mustahkamligi va moslashuvchanligini ko‘rsatadi.

4) **Dinamik mavzularni modellashtirishning samaradorligi**: BERTopic vaqt o‘tishi bilan mavzular evolyutsiyasini kuzatishda o‘zini yaxshi tomondan ko‘rsatadi.

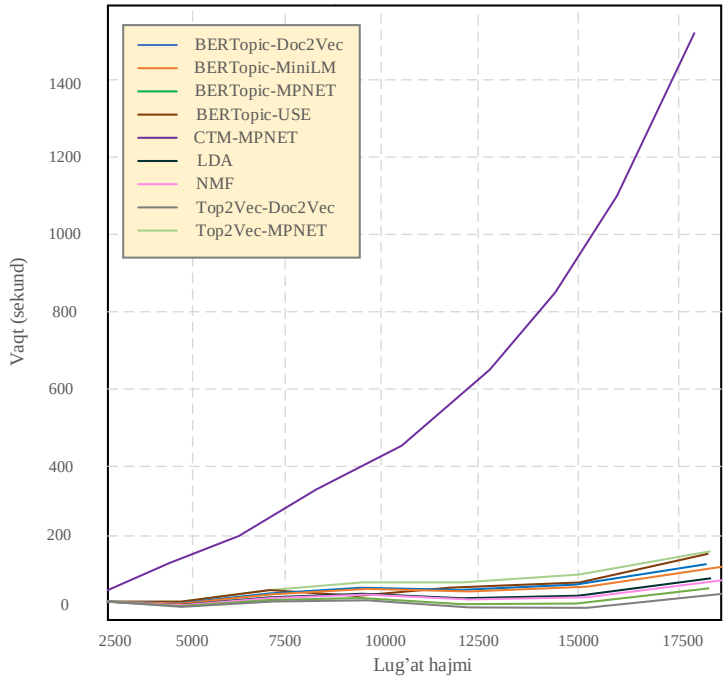
BERTopic tomonidan natija sifatida mavzular misoli quyidagicha ko‘rinishi mumkin:

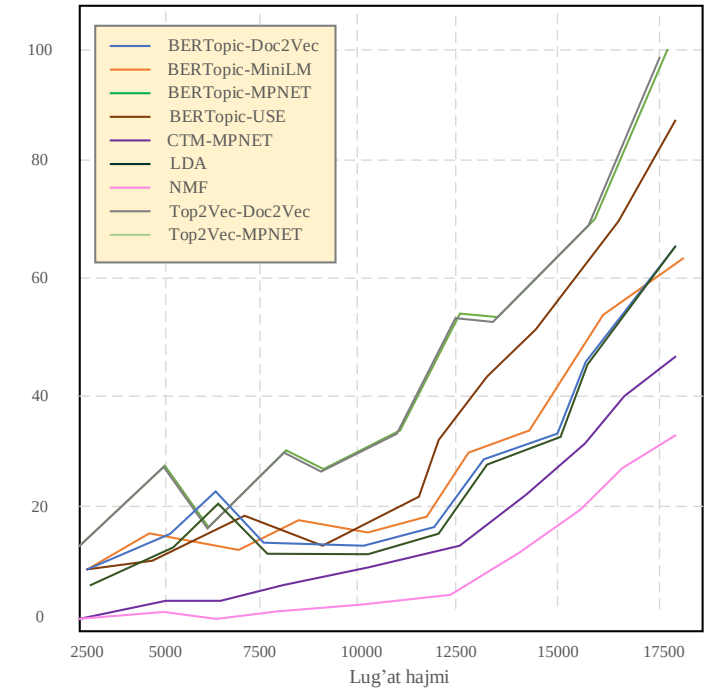
1-Mavzu: ['fazo‘, 'nasa', 'shatl', 'yer', 'start', 'missiya', 'orbita']

2-Mavzu: ['o‘yin', 'jamoa', 'o‘yinchi', 'sezon', 'xokkey', 'liga']

3-Mavzu: ['shifrlash', 'kalit', 'xavfsizlik', maxfiylik', 'davlat']

Ushbu mavzu ko‘rinishlari odatda an’anaviy mavzu modellari tomonidan ishlab chiqarilganidan ko‘ra ko‘proq interpretatsiyaga moyildir.





1-rasm: LDA va NMF kabi an'anaviy yondashuvlar bilan solishtirganda BERTopic ning hisoblash samaradorligini taqqoslash. Chapdagi grafik to‘liq bajarilish vaqtini taqqoslash, o‘ng grafik esa bajarilish vaqti shkalasining quyi qismiga qaratilgan.

1-rasmda ko‘rsatilganidek, BERTopic ishlash va hisoblash samaradorligi jihatidan yaxshi ko‘rsatkichlarni taklif qiladi. Til modelini tanlash o‘z navbatida jarayon vaqtiga sezilarli darajada ta’sir qiladi:

a) MiniLM kabi kichikroq modellardan foydalanish BERTopic-ni tezlik bo‘yicha odatiy usullar bilan raqobatbardosh qiladi.

b) MPNET kabi kattaroq modellar yaxshiroq semantik ma’no jihatidan yaxshiroq ko‘rsatkichlarni taqdim qiladi, lekin o‘ziga yarasha ko‘proq hisoblash resurslarini talab qiladi.

c) BERTopic odatda CTM dan tezroq, lekin odatiy LDA yoki NMF usullariga qaraganda sekinroq.

Hisoblash murakkabligi lug‘at hajmi bilan ortadi, ammo o‘sish tezligi ba'zi raqobatdosh yondashuvlarga nisbatan mutanosib ravishda boshqaruvga egadir.

BERTopicning sohaga qo‘shgan hissasi quyidagilardan iborat:

1) **Mavzular uyg‘unligi yaxshilandi**: Semantikaning vektor ko‘rinishi va sinfga asoslangan TF-IDF yondashuvidan foydalangan holda, BERTopic yanada izchil va izohlanadigan mavzularni ishlab chiqaradi.
2) **Moslashuvchanlik**: Modulli dizayn turli til modellari bilan oson integratsiya qilish va muayyan foydalanish holatlari uchun sozlash imkonini beradi.
3) **Natijaga erishish**: HDBSCAN-dan foydalanish BERTopic-ga hech qanday mavzuga tegishli bo‘lmagan hujjatlarni aniqlash imkonini beradi, bu esa mavzularning umumiy sifatini yaxshilaydi.
4) **Dinamik mavzu tahlili**: BERTopicning mavzu evolyutsiyasini kuzatish qobiliyati vaqtinchalik tahlil uchun yangi imkoniyatlarni ochadi.

Hayotiy qo‘llanmalar:

- Yangiliklar va ijtimoiy tarmoqlarda kontent tahlili
- Mijozlarning sharhlarini turkumlash
- Akademik tadqiqot yo‘nalishlarini tahlil qilish
- Ilmiy adabiyotlarni o‘rganish
- Siyosiy va huquqiy hujjatlarni tashkil etish

Xulosa o‘rnida aytish joizki, ushbu usul oqilona hisoblash samaradorligini saqlab qolgan holda an’anaviy tematik modellaridan ko‘ra ko‘proq mahsuldorlikka egadir. Til modellari rivojlanishda davom etar ekan, BERTopicning moslashuvchan tuzilmasi unga ushbu yutuqlarni o‘z ichiga olish imkonini beradi va bu uning tabiiy tillarni qayta ishlashning jadal rivojlanayotgan sohasida doimiy dolzarbligini ta'minlab turadi.

Foydalanilgan adabiyotlar:

1. Elov B., Alayev R., Aloyev N. (2024). Tematik modellashtirishning zamonaviy usullari. *Digital transformation and artificial intelligence*, *2*(1), 8–16. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i12
2. Elov.B., Alayev N. Matnlarini tematik modellashtirish va tasniflash usullari. barqarorlik va yetakchi tadqiqotlar onlayn ilmiy-amaliy jurnali. Vol. 3 No. 12 (2023). 263-276
3. Elov B., Aloyev N., Yuldashev A. (2023). SVD va NMF metodlari orqali tematik modellashtirish. O`zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 2(6). 55-66
4. Alghamdi, R., & Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1). https://doi.org/10.14569/ijacsa.2015.060121
5. Tao, R., Wei, Y., & Yang, T. (2021). Metaphor Analysis Method Based on Latent Semantic Analysis. *Journal of Donghua University (English Edition)*, *38*(1). https://doi.org/10.19884/j.1672-5220.202010087
6. Darmalaksana, W., Slamet, C., Zulfikar, W. B., Fadillah, I. F., Maylawati, D. S. adillah, & Ali, H. (2020). Latent semantic analysis and cosine similarity for hadith search engine. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(1). https://doi.org/10.12928/TELKOMNIKA.V18I1.14874
7. Ke, Z. T., & Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*.
8. https://doi.org/10.1080/01621459.2022.2123813
9. Churchill, R., & Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3507900
10. Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24). https://doi.org/10.4108/eai.13-7-2018.159623

UDK: 004.912:004.738.4

TABIIY TILNI QAYTA ISHLASH YORDAMIDA NUTQNI IMO-ISHORA TILIGA TARJIMA QILISH TEXNOLOGIYALARI

Jurayev Dilshod Boymuradovich,
Tayanch doktoranti
dilsamtuit@gmail.com
TATU

Anotatsiya. Maskur maqolada O‘zbekistondagi eshitish va gapirish qobiliyati zaif insonlar uchun nutq va matndan Surdo imo-ishora tiliga tarjima tizimini yaratish g‘oyasi taqdim etilgan. Tizim nutq va matnni qabul qilib, tabiiy tilni qayta ishlash moduli yordamida tahlil qiladi hamda ma’lumotlar bazasidan mos keluvchi video va animatsiyalarni tanlaydi. Keyin esa Surdo imo-ishora tilining grammatik qoidalariga mos ravishda ifodalaydi. Asosiy texnologik yechim sifatida tabiiy tilni qayta ishlash moduli qo‘llaniladi.

Kalit so‘zlar: *Imo-ishora tili, tabiiy tilni qayta ishlash, tokenizatsiya, lemmatizatsiya, sintaktik tahlil.*

Har bir mamlakatda, shu jumladan O‘zbekistonda ham, eshitish va nutq imkoniyati cheklangan shaxslar uchun asosiy muloqot vositasi bo‘lgan o‘ziga xos imo-ishora tili ya’ni Surdo imo-ishora tili mavjud. Afsuski, jamiyatning aksariyat qismi ushbu tilni bilmasligi yoki e'tiborsizligi sababli, oddiy insonlar va eshitish imkoniyati cheklangan insonlar o‘rtasida chuqur kommunikatsion bo‘shliq yuzaga kelgan. Bu bo‘shliq nafaqat kundalik muloqotni qiyinlashtiradi, bu insonlarning ta’lim olish, sog‘liqni saqlash va davlat xizmatlaridan foydalanish kabi muhim sohalarda teng imkoniyatlarga ega bo‘lishiga to‘sqinlik qiladi va ularni ijtimoiy hayotdan chetlashtirib quyadi.

Hozirgi kunda O‘zbekistonda, Surdo imo-ishora tilidan foydalanish darajasi, ayniqsa rasmiy va ta'lim sohalarida, yetarli emas. Ta’lim tizimida va o‘qitish jarayonlarda imo-ishora tiliga yetarlicha e’tibor qaratilmasligi hamda ota-onalarning Surdo imo-ishora tilining ahamiyatini to‘liq anglab yetmasligi mavjud vaziyatni yanada murakkablashtirmoqda. Natijada, eshitish imkoniyati cheklangan shaxslarning jamiyatga to‘liq integratsiyalashuvi va ularning salohiyatidan foydalanish imkoniyatlari sezilarli darajada cheklanmoqda.

Shu bilan birga, zamonaviy axborot texnologiyalari, xususan, tabiiy tilni qayta ishlash (NLP) va kompyuter ko‘rishi (Computer Vision) sohalarining rivojlanishi eshitish va nutq imkoniyati cheklangan shaxslar uchun kommunikatsion to‘siqlarni kamaytirish imkonini bermoqda. Imo-ishora tillarini avtomatik aniqlash va tarjima qilish tizimlari bu jarayonda muhim rol o‘ynaydi. Eshitish va gapirish

imkoniyati cheklangan shaxslar bilan ikki tomonlama muloqot o‘rnatish nafaqat aloqa imkoniyatlarini kengaytiradi, balki ularning jamiyat va texnologik taraqqiyotda faol ishtirok etishini ham ta’minlaydi **[6]**.

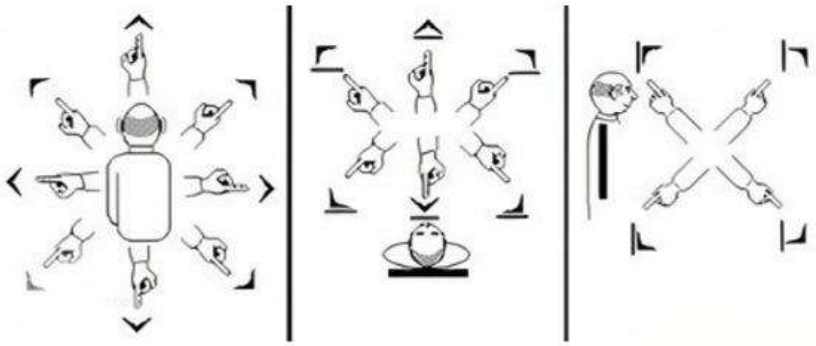
O‘zbek surdo imo-ishora tili. Har bir mamlakatning imo-ishora tili o‘z milliy tili va madaniyatiga asoslangan holda shakllanadi. O‘zbek Surdo imo-ishora tili ham o‘zbek tilining leksik va grammatik xususiyatlaridan ta’sirlangan bo‘lsa-da, o‘ziga xos strukturaga ega bo‘lgan mustaqil til tizimidir. **[1]**.

Imo-ishora tili eshitish imkoniyati cheklangan shaxslar uchun o‘z fikrini erkin ifodalash, balki ijtimoiy hayotda faol ishtirok etish va to‘laqonli hayot kechirish imkoniyatlarini yaratishda muhim rol o‘ynaydi. Eshitish imkoniyati cheklanganlar jamiyatida Surdo imo-ishora tilidan keng foydalanilayotgan bo‘lsa-da, rasmiy ta’lim tizimida uning qo‘llanilishi hanuzgacha yetarli darajada rivojlanmagan. Bu masalalar oliy ta’lim muassalarida kadrlarni tayyorlash dasturlarida Surdo imo-ishora tilidan samarali foydalanishga yetarlicha e’tibor berilmaganida aks etadi. Misol: O‘qitish jarayonida eshitish va gapirish imkoniyati cheklangan o‘quvchilarning ehtiyojlari to‘liq inobatga olinmayapti. Eshitish imkoniyati cheklangan bolalarning ota-onalari ko‘pincha imo-ishora tilining aloqa bo‘shliqlarini bartaraf etishdagi ahamiyatini bilishmaydi**[2]**.

Ushbu maqolada quyidagi masalalarga yechim topish maqsad qilingan:

- Eshitish va gapirish imkoniyati cheklangan shaxslarning muloqot imkoniyatlarini kengaytirish: Surdo imo-ishora tili orqali eshitish va gapirish imkoniyati cheklangan shaxslarning o‘zaro va jamiyat bilan samarali muloqot qilishlariga ko‘maklashish.
- Oliy va o‘rta ta’lim muassasalarida eshitish imkoniyati cheklangan o‘quvchilar va talabalar uchun Surdo imo-ishora tilidan foydalanib o‘qitish shaklini qo‘llash.
- Hukumat rahbarlari, o‘qituvchilar, mutaxassislar, jamoat rahbarlari va jamoatchilik kabi turli guruhlarni Surdo imo-ishora tili so‘zlashuvchilar bilan muloqotini yengillashtirish.

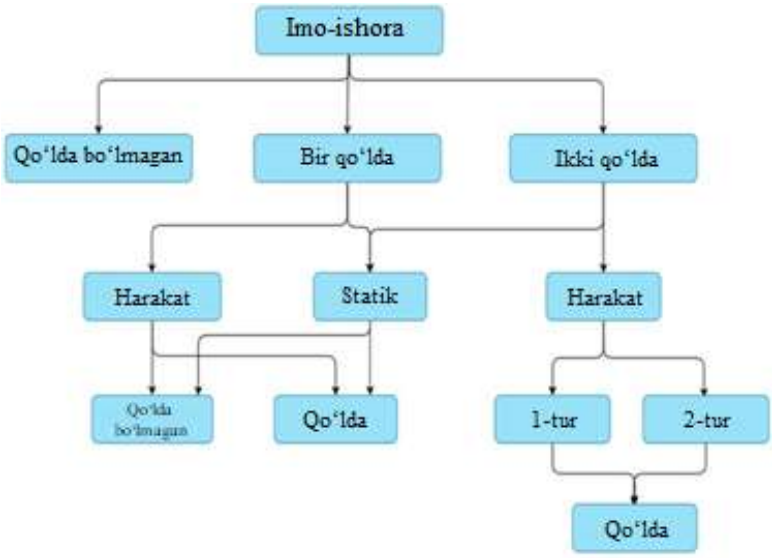
Surdo imo-ishora tili, 1-rasmda ko‘rsatilganidek, imo-ishoralar va ko‘rsatkichlarning turli shakllarini o‘z ichiga oladi.



1-rasm. Imo-ishora tilidagi harakat va ifoda yo‘nalishlari.

Surdo imo-ishora tili faqat qo‘l imo-ishora harakatlari bilan cheklanmay, balki yuzning mimika ifodalari, tana holati va harakatlari kabi komponentlarni ham o‘z ichiga oladi. Ushbu komponentlar imo-ishoralarning ma’nosini kuchaytirish, gap ohangini ifodalash hamda grammatik bog‘liqlikni ifodalash kabi funksiyalarni bajarib beradi. Imo-ishora tilining chuqurroq tahlili asosida uni 2-rasmda tasvirlanganidek uchta asosiy toifaga ajratish mumkin.

Imo-ishora tili nafaqat qo‘l harakatlari, balki yuz ifodalari, tana holati, og‘iz va lablarning hamohang ifodalarini ham o‘z ichiga oladi.Ushbu elementlar imo-ishora tilining ajralmas qismlaridan biri bo‘lib, modulyatsiya vositasi sifatida xizmat qiladi.



2-rasm. Qo‘l imo-ishoralarining turlari.

Bizning og‘zaki nutqda ohang va intonatsiya orqali fikr ifodasi kuchaytirilib, aniqlashtirilsa, surdo imo-ishora tilida esa tana harakati va yuz ifodalari orqali ma’no kuchaytiriladi, gapning ohangi va kayfiyati aniqlik bilan yetkaziladi. Shu tariqa, tana imo-ishoralari muloqot jarayonida semantik yukni oshirish, emotsional fonni shakllantirish va grammatik tuzilmalarni aniqlashtirish vazifasini bajaradi**[5]**.

Shuningdek bir qo‘l bilan amalga oshiriladigan imo-ishoralar ikki toifaga ajratiladi:

- ***Statik harakatlar*** – belgining ma’lum bir holatda qotib turishi.
- ***Dinamik harakatlar*** – belgining harakatlanish jarayonida ifodalanishi.

Ushbu harakatlar qo‘lga tegishli yoki qo‘l bilan bog‘liq bo‘lmagan shakllarda amalga oshirilishi mumkin. Ikki qo‘l ishtirokidagi imo-ishoralar ham statik va dinamik harakatlarga bo‘linadi. Biroq, ular yana quyidagi turlarga tasniflanadi:

- ***0-tur*** – har ikkala qo‘l teng darajada ishtirok etadigan imo-ishoralar.
- ***1-tur*** – ikkala qo‘l ishtirok etadigan, biroq ulardan biri yetakchi, ikkinchisi yordamchi vazifani bajaradigan imo-ishoralar.

Mazkur tasnif imo-ishora tilidagi murakkablik va ifoda boyligini chuqurroq anglash imkonini beradi hamda uni o‘rganish va o‘qitish jarayonida muhim metodologik asos vazifasini bajaradi.

O‘zbek Surdo imo-ishora tili grammatikasi o‘zbek tilining grammatik tizimidan sezilarli darajada farqlanishi mumkin **[1]**. Xususan:

- ***So‘z tartibi***: Surdo imo-ishora tilida gap tuzilishi ko‘pincha *Ega-Olmosh-Fe'l* strukturasi asosida tashkil topadi.
- ***Fe’llarning tuslanishi***: Fe’l shakllari imo-ishoralarda kontekstual ravishda ifodalanadi va ularning zamon, shaxs yoki nisbat shakllari alohida grammatik belgilar orqali emas, balki imo-harakat va yuz ifodalari vositasida bildiriladi.
- ***Ko‘plik va egalik shakllari***: O‘zbek tilidagi kabi ko‘plik yoki egalik qo‘shimchalari imo-ishoralarda aniq grammatik affikslar bilan emas, balki maxsus harakatlar, ko‘rsatkichlar yoki qo‘shimcha mimika orqali ifodalanadi.

Shu sababli, og‘zaki nutqdan imo-ishora tiliga so‘zma-so‘z tarjima qilish ko‘pincha noto‘g‘ri yoki tushunarsiz natijalarga olib keladi. Tarjima jarayonida O‘zSITning o‘ziga xos grammatik xususiyatlarini e’tiborga olish zarur. Aks holda, muloqotda semantik buzilishlar va tushunmovchiliklar yuzaga kelishi mumkin.

Taklif etilayotgan usul: Ushbu maqolada taqdim etilgan taklif qilingan tizim eshitish qobiliyati zaif odamlarga boshqa eshitish qobiliyatiga ega odamlar bilan osongina muloqot qilishda yordam beradigan real vaqt rejimida nutqni Surdo imo-ishora tiliga o‘tkazish tizimidir. Tizim asosan oltita komponentdan iborat:

1. ***Kirishni qayta ishlash moduli*** – foydalanuvchi tomonidan kiritilgan ma’lumotni tizim uchun mos shaklga keltirish vazifasini bajaradi. Ushbu modul ikki asosiy funksiyani o‘z ichiga oladi:

- *Nutqni matnga aylantirish (Speech-to-Text, STT)* - Agar foydalanuvchi ma’lumotni ovozli nutq shaklida kiritsa, birinchi bosqichda nutqni matnga

aylantirish jarayoni amalga oshiriladi. Bunda o‘zbek tiliga moslashtirilgan maxsus Speech-to-Text (STT) modelidan foydalanish talab etiladi. STT tizimining aniqligi butun tarjima tizimi samaradorligiga bevosita ta’sir ko‘rsatadi. Biroq, muloqot jarayonidagi fon shovqinlari, turli hudud dialektlari va individual talaffuz xususiyatlari STT modulining ishlash sifatiga sezilarli darajada ta’sir qilishi mumkin [2].

- *To‘g‘ridan-to‘g‘ri matn kiritish* - Foydalanuvchi ma’lumotni bevosita matn shaklida kiritishi ham mumkin. Bunday holatda, Speech-to-Text bosqichi o‘tkazib yuboriladi va tizim matn bilan bevosita ishlay boshlaydi.

2. Tabiiy tilni qayta ishlash (NLP) moduli - tizimning yadrosi bo‘lib, foydalanuvchi tomonidan kiritilgan matnni lingvistik jihatdan tahlil qiladi va uni Surdo imo-ishora tiliga tarjima qilish uchun tayyorlaydi (4-rasmga qarang).

Ushbu modul quyidagi asosiy bosqichlardan iborat [2]:

- *Dastlabki ishlov berish* — Matn keraksiz belgilardan (masalan, tinish belgilari va ortiqcha bo‘shliqlardan) tozalanadi va normallashtiriladi. Masalan, barcha harflar kichik harf shakliga keltiriladi, sonlar esa so‘z shaklida ifodalanadi. Ushbu bosqich matnni bir xil va tizimli ko‘rinishga keltirishga xizmat qiladi.
- *Tokenizatsiya* — Matn alohida birliklarga — tokenlarga (odatda so‘zlar yoki gap bo‘laklariga) ajratiladi. O‘zbek tilida so‘z o‘zagiga ko‘plab qo‘shimchalar qo‘shilishi mumkinligi sababli, dastlab so‘z darajasidagi tokenizatsiya amalga oshiriladi.
- *Nomuhim so‘zlarni olib tashlash (Stopword Removal)* — Matnda tez-tez uchraydigan, lekin semantik yuklamasi past bo‘lgan so‘zlar (masalan, "va", "bilan", "uchun", "bu", "shu") aniqlanib, olib tashlanadi. Bu jarayon tarjima sifatini oshirish va asosiy ma’noli birliklarga e’tiborni jamlash uchun muhimdir.
- *Stemming va Lemmatizatsiya* — Har bir so‘zning morfologik tahlili amalga oshiriladi va u lug‘aviy asosiy shakliga (lemma) keltiriladi. Masalan, "kitoblarimni" so‘zi "kitob" lemmaga qisqartiriladi. Ushbu bosqich imo-ishoralar ma’lumotlar bazasi bilan mos ishlashni ta’minlash uchun zarur.
- *Nutq qismlarini belgilash (POS-Tagging)* — Har bir tokenning nutqdagi turkumi (masalan, ot, fe‘l, sifat, ravish va boshqalar) aniqlanadi. Ushbu ma’lumot keyingi bosqichda sintaktik tahlil va grammatik qoidalarni to‘g‘ri qo‘llash uchun asos bo‘lib xizmat qiladi.
- *Sintaktik tahlil (Parsing)* — Gapning grammatik strukturasini aniqlash jarayoni bo‘lib, bunda so‘zlar o‘rtasidagi munosabatlar va frazaviy birliklar

tahlil qilinadi. Bu bosqich imo-ishora tiliga mos grammatik tuzilma hosil qilish uchun muhim ahamiyat kasb etadi.

3. Tarjima va natijani generatsiya qilish moduli - Taklif etilayotgan tizim Surdo imo-ishora tiliga nutq va matn asosida avtomatik tarjima qilish uchun mo‘ljallangan bo‘lib, quyidagi asosiy bosqichlardan iborat [3]:

- *Qoidalarga asoslangan tarjima qilish* orqali kiritilgan matn imo-ishora tili grammatik tuzilmasiga mos ravishda so‘zlar ketma-ketligi qayta tartiblanadi.
- *So‘zlarni aniqlash* mexanizmi yordamida barqaror frazalar aniqlanib, agar mavjud bo‘lsa, yagona imo-ishora belgisi bilan ifodalanadi.
- *Ma’lumotlar bazasidan qidirish* orqali so‘zlar yoki iboralarga mos imo-ishora videolari yoki animatsiyalari topiladi.
- *Zaxira mexanizmi* sifatida mos keluvchi imo-ishora topilmaganda, barmoq alifbosi yordamida harflab ko‘rsatish funksiyasi qo‘llaniladi.
- *Natijani generatsiya qilish* orqali tanlangan imo-ishora videolari va animatsiyalar imo-ishora tili grammatik qoidalariga mos tartibda birlashtirilib, foydalanuvchiga taqdim etiladi. Animatsiyalar 3D avatar texnologiyalari yordamida yuz ifodasi va tana harakati kabi imo-ishoralar orqali ham ifodalanadi.

Xulosa. Ushbu maqolada o‘zbek tilidagi nutq va matnni Surdo imo-ishora tiliga tarjima qiluvchi tizim konsepsiyasi va metodologiyasi taklif etildi. Tizim NLP bosqichlari tokenizatsiya, lemmatizatsiya, sintaktik hamda imo-ishora tili grammatik qoidalarini qo‘llash orqali foydalanuvchi kiritgan ma’lumotni mos ravishda imo-ishora ketma-ketligiga aylantiradi. Natijada, foydalanuvchi tomonidan kiritilgan matnning mazmuni ma’lumotlar bazasidan tanlangan video yoki animatsiyalar yordamida ifodalanadi.

Taklif etilayotgan tizim eshitish va nutq qobiliyati cheklangan shaxslar uchun ta’lim, kundalik hayot va davlat xizmatlarida muloqot imkoniyatlarini sezilarli darajada kengaytirishga xizmat qiladi. Shu bilan birga, sifatli imo-ishora ma’lumotlar bazasini yaratish, o‘zbek tili uchun NLP vositalarini takomillashtirish, imo-ishora tili grammatikasini standartlashtirish va imo-ishora tili elementlarni tabiiy tarzda hosil qilish tizim rivojlanishining asosiy muammolari sifatida ko‘riladi.

Kelgusida tizimning dastlabki prototipini ishlab chiqish, uni real foydalanuvchilarda sinovdan o‘tkazish va natijalarga asoslangan takomillashtirish ishlari amalga oshirilishi rejalashtirilgan. Tadqiqot natijalari O‘zbekistonda inklyuziv jamiyatni rivojlantirish va kommunikatsion to‘siqlarni texnologiyalar orqali kamaytirishga hissa qo‘shishi kutilmoqda.

Foydalanilgan adabiyotlar:

1. Jurayev D.B, Ochilov M.M, Abdullayeva M.I. Nutqni imo-ishora tiliga tarjima qilish tizimlarida matnlarga ishlov berish. TATU xabarlari, 4/68(2023).
2. Mukhammadjon Musaev, Dilshod Juraev, Mannon Ochilov, Malika Abdullaeva. “Text processing in speech translation systems into sign language”. International Scientific Conference on Modern Problems of Applied Science and Engineering MPASE2024, Samarkand, Uzbekistan • 2–3 May 2024, AIP Conf. Proc. 3244, 030062 (2024) https://doi.org/10.1063/5.0241738
3. Zhu, D., Czehmann, V., & Avramidis, E. (2023). Neural Machine Translation Methods for Translating Text to Sign Language Glosses. *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 12523–12541. https://aclanthology.org/2023.acl-long.700/
4. Moryossef, A., Jiang, Z., Müller, M., Ebling, S., & Goldberg, Y. (2023). Linguistically Motivated Sign Language Segmentation. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12703–12724. https://doi.org/10.18653/v1/2023.findings-emnlp.846
5. Camgoz, N. C., Koller, O., Hadfield, S., & Bowden, R. (2020). Sign Language Transformers: Joint End-to-end Sign Language Recognition and Translation. https://arxiv.org/abs/2003.13830
6. Shahin, N., & Ismail, L. (2024). From Rule-Based Models to Deep Learning Transformers Architectures for Natural Language Processing and Sign Language Translation Systems: Survey, Taxonomy and Performance Evaluation. https://arxiv.org/abs/2408.14825

RAQAMLI TILSHUNOSLIK: BLINGVIZM VA SUN’IY INTELLEKT O‘RTASIDAGI ALOQA

Ruziyeva Madina Ilhom qizi,
Tayanch doktoranti
madinaruziyeva67@gmail.com
ToshDO‘TAU

Annotatsiya. Zamonaviy tilshunoslik raqamli texnologiyalar bilan chambarchas bog‘liq holda rivojlanmoqda. Raqamli tilshunoslik, ya’ni tilshunoslikning texnologik va raqamli vositalar yordamida o‘rganilishi, so‘nggi yillarda juda muhim yo‘nalishga aylandi. Bu soha sun’iy intellekt (SI) va blingvizm (inglizcha “blended linguistics”) kabi yangi tushunchalarni o‘z ichiga oladi. Su’iy intellektning tilshunoslikdagi ahamiyati hamda blingvizmning bu jarayondagi roli kognitiv tilshunoslik, avtomatik tarjima, nutqni tanib olish va tilni qayta ishlash kabi sohalarda sezilarli tarzda o‘sdi. Ushbu maqolada raqamli tilshunoslik, blingvizm va sun’iy intellektning tilshunoslikdagi o‘rni, ularning o‘zaro aloqasi va kelajakdagi rivojlanish istiqbollari haqida so‘z boradi.

Annotation. Modern linguistics is developing in close connection with digital technologies. Digital linguistics, that is, the study of language through technological and digital tools, has become a highly important field in recent years. This area includes new concepts such as artificial intelligence (AI) and blended linguistics. The significance of artificial intelligence in linguistics, as well as the role of blended linguistics in this process, has notably increased in areas such as cognitive linguistics, machine translation, speech recognition, and natural language processing. This article discusses the role of digital linguistics, blended linguistics, and artificial intelligence in the field of linguistics, their interrelations, and future development prospects.

Аннотация. Современное языкознание развивается в тесной связи с цифровыми технологиями. Цифровая лингвистика, то есть изучение языка с помощью технологических и цифровых средств, в последние годы стала очень важным направлением. Эта область включает в себя такие новые понятия, как искусственный интеллект (ИИ) и блендвизм (от англ. "blended linguistics"). Значение искусственного интеллекта в лингвистике, а также роль блендвизма в этом процессе значительно возросли в таких сферах, как когнитивная лингвистика, автоматический перевод, распознавание речи и обработка языка. В данной статье рассматриваются место цифровой лингвистики, блендвизма и искусственного интеллекта в языкознании, их взаимосвязь и перспективы развития в будущем.

Kalit so‘zlar: *raqamli tilshunoslik, sun’iy intellekt, blingvizm, avtomatik tarjima, nutqni tanib olish, texnologik tilshunoslik, lingvistik innovatsiyalar, zamonaviy lingvistika*

Raqamli tilshunoslik – bu tilshunoslikni raqamli vositalar yordamida o‘rganish jarayoni bo‘lib, u kompyuterlar, algoritmlar, big data va sun’iy intellektni qo‘llashni o‘z ichiga oladi. Blingvizm, o‘z navbatida, turli tillarni birlashtirish orqali yangi til tizimlarini yaratishni anglatadi. U tilni avtomatik qayta ishlash, mashinasozlik tarjimasi va nutqni tanib olish tizimlarida keng qo‘llaniladi. Sun’iy intellekt (SI) va blingvizm o‘rtasidagi aloqalar lingvistik tadqiqotlar uchun yangi imkoniyatlar ochib bermoqda. Quyidagi jadvalda raqamli tilshunoslikning asosiy aspektlari aks ettirilgan:

Aspekt	Tavsif	Misollar
Sun’iy intellekt	Tabiiy tilni qayta ishlash uchun texnologiyalar	Avtomatik tarjima nutqni tanib olish
Blingvizm	Bir nechta tillarni birlashtirish tizimi	Mashinasozlik tarjimasi turli tillarni birlashtirish
Kognitiv tilshunoslik	Insonning tilni qanday tushunishi va ishlatishi	Tilni qayta ishlash va o‘rganish algoritmlari

Bugungi kunda raqamli texnologiyalar va sun’iy intellekt (SI) tilshunoslik sohasida inqilob yasamoqda. Ayniqsa, blingvizm – ikki yoki undan ortiq tillarni bilish fenomeni – va sun’iy intellekt o‘rtasidagi o‘zaro bog‘liqlik raqamli tilshunoslikning dolzarb yo‘nalishlaridan biri hisoblanadi.

Blingvizm ikki yoki undan ortiq tillarni mukammal bilish va ulardan faol foydalanish qobiliyatidir. Bu fenomen kognitiv neyrobiologiya, psixolingvistika va sotsiolingvistika sohalarida muhim ahamiyat kasb etadi. Shuningdek, blingvizm inson miyasining moslashuvchanligi va lingvistik ko‘nikmalarni qayta ishlash qobiliyatini aks ettiruvchi muhim hodisalardan biridir. So‘nggi tadqiqotlar shuni ko‘rsatmoqdaki, ko‘p tilli insonlar:kognitiv moslashuvchanlikka ega bo‘lib, muammolarni hal qilish qobiliyati yuqori bo‘ladi. Sun’iy intellekt til o‘rganish jarayonini osonlashtirish va optimallashtirishga qaratilgan texnologik yondashuvlarni taqdim etadi. Masalan, neyron tarmoqlar asosida qurilgan tarjima tizimlari blingvistik jarayonlarga ta’sir ko‘rsatmoqda. Sun’iy intellekt asosida ishlab chiqilgan lingvistik modellar, xususan, tabiiy tilni qayta ishlash (NLP) tizimlari, blingvizmga oid muammolarni hal qilish va uni chuqur o‘rganishda katta imkoniyat yaratmoqda. Quyidagi texnologiyalar bunga misol bo‘la oladi: Nevrologik tarmoqlarga asoslangan tarjima tizimlari (masalan, Google Translate, DeepL) blingvistik jarayonlarni modellashtirishga yordam beradi. Til modeli asosidagi sun’iy intellekt (masalan, ChatGPT, BERT, T5) bir nechta tilda fikrlash va muloqot

qilish ko‘nikmalarini rivojlantirish uchun qo‘llaniladi. Tilni avtomatik aniqlash va o‘rganish tizimlari (masalan, Duolingo, Babbel) ikki tillilik jarayonlarini kuchaytiradi.

Blingvizm va sun’iy intellekt o‘rtasidagi o‘zaro bog‘liqlik. Sun’iy intellekt inson miyasi kabi ko‘p tillilik bilan bog‘liq jarayonlarni tushunishga harakat qilmoqda. Ushbu bog‘liqlikni quyidagi jihatlarda ko‘rish mumkin:

- Ko‘p tillilikka moslashish – SI modellarining mashinaviy o‘rganish orqali turli tillarni o‘zlashtirish qobiliyati blingvizm prinsiplariga o‘xshaydi.

- Interferensiya va kod almashtirish modellarining yaratilishi – SI lingvistik interfeyslar orqali bir tildan ikkinchisiga o‘tish jarayonlarini tushunishga yordam beradi.

- Til o‘rganishda yordamchi vosita sifatida – SI asosidagi vositalar insonlarning blingvistik malakalarini rivojlantirish uchun ishlatiladi.

Kelajakda SI asosida rivojlanayotgan lingvistik tadqiqotlar quyidagilarga yo‘l ochishi mumkin: Xususan, dunyoning rivojlangan mamlakatlarida barcha tillarni qamrab oluvchi universal tarjima tizimlari yaratilishi boshlangan. Bilingv shaxslar uchun moslashtirilgan til o‘rgatuvchi sun’iy intellekt vositalari ishlab chiqiladi. Til o‘rganish jarayonlarini tezlashtirishga xizmat qiluvchi interaktiv modellar yaratiladi. Ko‘p tilli sun’iy intellekt yordamida yanada tabiiy va kontekstga mos muloqot tizimlari ishlab chiqiladi.

Blingvizm va sun’iy intellektning rivojlanishi raqamli tilshunoslikda yangi imkoniyatlar yaratmoqda. Sun’iy intellekt lingvistik jarayonlarni modellashtirish orqali blingvizm fenomenini chuqur o‘rganishga va insonning kognitiv ko‘nikmalarini yaxshilashga yordam beradi. Kelajakda ushbu ikki sohaning uyg‘unlashuvi natijasida til o‘rganish jarayoni yanada rivojlanib, yangi texnologik yutuqlarga yo‘l ochiladi.

Quyidagi jadvalda blingvizm va sun’iy intellekt o‘rtasidagi bog‘liqlik yoritiladi:

Xususiyatlar	Blingvizm	Sun’iy intellekt
O‘rganish usuli	Tajriba asosida	Ma’lumotlar asosida
Moslashuvchanlik	Juda yuqori	Modelga bog‘liq
Tilni tushunish	Kontekstga bog‘liq	Ma’lumotlar bazasiga bog‘liq

Sun’iy intellektning blingvistik kompetensiyaga ta’siri. Til o‘rgatuvchi dasturlar, avtomatik tarjima tizimlari va tovushli yordamchilar blingvistik kompetensiyani rivojlantirishda muhim o‘rin tutadi. Masalan, GPT va DeepL kabi tizimlar real vaqtda tarjima qilish imkoniyatini beradi. Blingvizmning inson kognitiv

tizimiga ta’siri Nevropsixologik tadqiqotlar shuni ko‘rsatadiki, blingv shaxslarning miyasi monolingv (bitta tilni biladigan) shaxslarga nisbatan moslashuvchanroq bo‘ladi. Blingvistik jarayonlar, xususan, dorsolateral prefrontal korteks va anterior singulat korteksning faoliyatini kuchaytiradi . Ushbu jihat sun’iy intellekt modellarini ko‘p tillilikka moslashuvchan qilish uchun ishlatilishi mumkin.[1]

Sun’iy intellekt va lingvistik modellar. Transformer modellar tabiiy tilni qayta ishlash sohasida inqilob qildi. Ushbu model bir nechta tillarda ishlashga moslashgan bo‘lib, tarjima, so‘z tushunish va nutq sintaksisini qayta ishlashda yuqori samaradorlik ko‘rsatmoqda. Multilingvistik o‘qitish uchun Meta AI tomonidan yaratilgan M2M-100 modeli bir nechta tillarni bir vaqtning o‘zida tarjima qilish qobiliyatiga ega bo‘lib, u 100 dan ortiq tillarni o‘z ichiga oladi. [2]

Sun’iy intellekt yordamida blingvizmni modellashtirish. Google Brain tadqiqotlari sun’iy intellekt yordamida inson miyasi kabi kod almashtirish va interferensiya jarayonlarini tushunish va simulyatsiya qilishga harakat qilmoqda. Duolingo va boshqa til o‘rgatish platformalari adaptive learning metodikasidan foydalangan holda, sun’iy intellekt yordamida blingvizmga oid pedagogik usullarni taklif qilmoqda.[3]

Bilingvizm SI uchun katta imkoniyatlar ochadi: ko‘p tillilikka asoslangan tarjima tizimlarini yaxshilaydi. Til o‘rganish platformalari uchun moslashuvchan darsliklar yaratadi. Ovozli yordamchi tizimlarni turli tillarda faol ishlashga moslaydi. Kognitiv SI modellarida inson miyasi faoliyatini chuqurroq tushunishga xizmat qiladi.

Blingvizm va sun’iy intellekt o‘rtasidagi aloqani quyidagi jadval orqali aniqroq tushunamiz:

№	Asosiy jihatlar	Tavsif
1	Sun’iy intellektning lingvistik modellari	Transformer modellar ikki tillilik jarayonlarini avtomatlashtirishga yordam beradi.
2	Blingvizm va kognitiv jarayonlar	Blingv shaxslarning miyasi ko’proq moslashuvchan bo‘lib, bu lingvistik qobiliyatlarni rivojlantiradi.
3	Ikki tillilik va tarjima tizimlari	Google Translate va M2M-100 kabi modellar blingvizm aosida ishlaydi.
4	Kod almashtirish va interferensiya	Blingvizmda kuzatiladigan kod almashtirish jarayonlari SI yordamida modellashtirilmoqda.
5	Kelajakdagi istiqbollar	Kelajakda SI yordamida real vaqtda tarjima qiluvchi ko‘zoynaklar va smartfon ilovalari ishlab chiqilishi mumkin. Masalan, Google Glass tarjima funksiyasi kabi.

Raqamli tilshunoslik, blingvizm va sun’iy intellekt bo‘yicha amaliy takliflar:

1. Bilingval bolalar uchun sun’iy intellekt yordamida o‘quv platformalarini ishlab chiqish. SI asosida tillarni o‘rgatadigan mobil ilovalar va veb-platformalar yaratish. Bilingval bolalar uchun maxsus moslashtirilgan interaktiv darslik va mashqlar tizimini ishlab chiqish.
2. Til o‘rganuvchilar uchun sun’iy intellekt asosidagi moslashuvchan tarjima tizimlarini rivojlantirish. SI yordamida matnlarni kontekstga qarab mos tarjima qilish tizimlarini takomillashtirish. O‘zbek tilini ingliz, rus va turk tillari bilan bog‘laydigan avtomatik tarjima tizimlarini rivojlantirish.
3. O‘zbek tilidagi sun’iy intellekt modellarini rivojlantirish. O‘zbek tilidagi nutqni matnga aylantirish va aksincha, matndan nutqqa o‘giruvchi texnologiyalarni takomillashtirish. O‘zbekcha chatbot va virtual yordamchilarni ishlab chiqish.
4. Sun’iy intellektga asoslangan lug‘atlar va til o‘rganish vositalarini yaratish. O‘zbek va boshqa tillar uchun AI asosida ishlaydigan interaktiv lug‘atlar ishlab chiqish. Blingvistik kompetensiyani oshirish uchun sun’iy intellekt asosida adaptiv til o‘rgatish tizimlarini yaratish.
5. Tilshunoslik tadqiqotlari uchun sun’iy intellekt vositalaridan foydalanish. O‘zbek tilida matn tahlil qilish va semantik analiz tizimlarini yaratish. Sun’iy intellekt yordamida o‘zbek tilining grammatik va sintaktik xususiyatlarini chuqurroq tadqiq qilish.
6. O‘zbek tilshunoslari va IT mutaxassislari o‘rtasida hamkorlikni kuchaytirish. Oliy ta’lim muassasalari va IT kompaniyalari o‘rtasida qo‘shma loyihalar tashkillashtirish. Sun’iy intellekt va lingvistika bo‘yicha tadqiqot markazlarini tashkil etish.
7. Bilingval shaxslarning til qobiliyatini baholash va rivojlantirish uchun sun’iy intellekt vositalaridan foydalanish. Talabalar va til o‘rganuvchilar uchun AI asosida ishlaydigan diagnostika testlarini yaratish. O‘zbek tilida avtomatik esse baholash tizimini ishlab chiqish.

Sun’iy intellekt tizimlarining tilga oid modellari bilingvizmga asoslanadi. Bu modellar inson miyasi kabi ikki yoki undan ortiq tilda ma’lumotni qabul qilish, qayta ishlash va uni kontekstga mos holda ifodalashga harakat qiladi.[4] Ayniqsa, neyron tarmoqlar asosida qurilgan mashina tarjimasi va ko‘p tilli til modellari (masalan, ChatGPT, Google Translate, M2M-100, NLLB-200) bilingvizm mexanizmlarini chuqur o‘rganish va undan amaliy foyda olish imkonini beradi.[5] Shuningdek, bilingvizmni o‘rganish orqali SI tizimlari inson kognitsiyasini yanada chuqurroq anglay boshlaydi. Bu esa turli tillarda ishlovchi foydalanuvchilar uchun intellektual yordamchilarni, ko‘p tilli ta’lim resurslarini, tilni avtomatik aniqlash, tarjima, nutq tanish, til korreksiyasi va boshqa xizmatlarni rivojlantirishga turtki bo‘ladi. O‘zbek

tilshunosligi uchun bu soha alohida dolzarbdir. Chunki sun’iy intellekt asosida o‘zbek tilini rivojlantirish, uni ko‘p tilli tizimlarga integratsiya qilish, korpuslar yaratish va avtomatik tarjima vositalarini ishlab chiqishda bilingvizmga oid yondashuvlar katta imkoniyatlarni taqdim etadi.[6] Shu bois, bu yo‘nalishdagi ilmiy izlanishlar nafaqat nazariy, balki amaliy jihatdan ham yurtimizda dolzarb bo‘lib borayotgan sun’iy intellekt sohasiga katta hissa qo‘shadi.

Foydalanilgan adabiyotlar:

1. Bialystok, E. Bilingualism: The Good, the Bad, and the Indifferent. Bilingualism:Language and Cognition,2009- 12(1), 3-11. https://doi.org/ 10.1017/ S1366728908003477
2. Fan, A., Bhosale, S.,Schwenk, H.,Ma, X.,El-Kishky, A.,Goyal, N.,& Edunov, S. 2020.
3. Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., … & Dean, J. (2017). Google’s Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. Transactions of the Association for Computational Linguistics, 5, 339-351. https://doi.org/10.1162/tacl_a_00065 Press.
4. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All You Need. Advances in Neural Information Processing Systems (NeurIPS). https://arxiv.org/abs/1706.03762
5. OpenAI. (2023). GPT-4 Technical Report.
6. Toshkent davlat o‘zbek tili va adabiyoti universiteti ilmiy jurnali. (2021). Sun’iy intellekt va lingvistik model tizimlari.

SUN’IY INTELLEKT KUTUBXONALARI YOQDAMIDA MATNLARNI QAYTA ISHLASH

Berdiyev Jahongir Botir o‘g‘li,
Kompyuter lingvistikasi mutaxassisligi magistranti
berdiyevjahongir94@gmail.com
ToshDO‘TAU

Annotatsiya. Zamonaviy texnologiyalar taraqqiyoti sun’iy intellekt (SI) sohasini jadal rivojlantirmoqda. Ayniqsa, tabiiy tilni qayta ishlash (Natural Language Processing — NLP) yo‘nalishida erishilayotgan yutuqlar turli sohalarda matn bilan ishlash jarayonini avtomatlashtirish va optimallashtirish imkonini bermoqda. Ilgari faqat inson aqli orqali bajarilgan matn tahlili, his-tuyg‘ularni aniqlash, tarjima, xulosa chiqarish kabi vazifalar bugungi kunda sun’iy intellekt yordamida yuqori aniqlikda amalga oshirilmoqda. Matn bilan ishlashda asosiy vazifa uni kompyuter tushunadigan shaklga keltirishdir. Buning uchun maxsus algoritmlar va modellardan foydalaniladi. Shu jarayonda TensorFlow, PyTorch, Scikit-learn kabi SI kutubxonalari muhim o‘rin egallaydi. Ushbu kutubxonalar yordamida neyron tarmoqlarni o‘rgatish, mavjud modellarni sozlash va yangi yondashuvlarni sinovdan o‘tkazish mumkin.

Ushbu maqolada matnlarni qayta ishlash jarayonida sun’iy intellekt kutubxonalarining tutgan o‘rni, ularning imkoniyatlari va amaliy qo‘llanilishi haqida so‘z yuritiladi. Shuningdek, yetakchi tadqiqotchilar ishlanmalari asosida iqtiboslar keltiriladi va turli kutubxonalar taqqoslanadi.

Abstract. The rapid advancement of modern technologies has significantly accelerated the development of artificial intelligence (AI). In particular, the field of Natural Language Processing (NLP) has witnessed substantial progress, enabling the automation and optimization of various text analysis processes. Tasks such as sentiment detection, text classification, summarization, and machine translation, which were once performed solely by human experts, are now being carried out by intelligent systems with high accuracy. One of the core challenges in text processing is converting natural language into a machine-readable format. To achieve this, various algorithms and deep learning models are employed. In this context, open-source AI libraries such as TensorFlow, PyTorch, Scikit-learn play a crucial role. These libraries provide efficient tools for training neural networks, fine-tuning pre-trained models, and experimenting with innovative approaches.

This article explores the role and capabilities of AI libraries in text processing tasks. It further discusses practical applications and comparative analysis of leading frameworks, referencing key contributions from prominent researchers in the field.

Аннотация. Стремительное развитие современных технологий существенно ускорило прогресс в области искусственного интеллекта (ИИ). Особенно заметны достижения в сфере обработки естественного языка (Natural Language Processing — NLP), что позволило автоматизировать и оптимизировать различные задачи анализа текстов. Такие операции, как определение тональности, классификация текста, реферирование и машинный перевод, ранее выполнявшиеся исключительно человеком, сегодня реализуются интеллектуальными системами с высокой точностью. Одной из основных задач при работе с текстами является преобразование естественного языка в формат, понятный для машин. Для этого используются специализированные алгоритмы и модели глубокого обучения. В этом контексте важную роль играют открытые библиотеки ИИ, такие как TensorFlow, PyTorch, Scikit-learn. Эти инструменты позволяют эффективно обучать нейронные сети, адаптировать уже готовые модели и тестировать новые подходы.

В данной статье рассматриваются роль и возможности библиотек искусственного интеллекта при обработке текстов. Также проводится сравнительный анализ популярных инструментов с опорой на ключевые научные исследования и авторитетные источники.

Kalit so‘zlar. *TensorFlow, PyTorch, Scikit-learn, tokenizatsiya, vektorlashtirish, klasslar.*

Kirish. Matnni qayta ishlash sun’iy intellektning tabiiy tilni qayta ishlash (NLP) sohasiga kiradi. Uning asosiy maqsadi mashinalarga inson tilini tushunish, talqin qilish va yaratish imkoniyatini berishdir. Chuqur o‘rganishning rivojlanishi bilan matnga asoslangan vazifalarni avtomatlashtirish va takomillashtirish uchun sun’iy intellekt kutubxonalari muhim rol o‘ynamoqda. Bu tizim matnni tasniflash, hissiyotlarni tahlil qilish, mashina tarjimasi va boshqa ko‘plab NLP vazifalarini bajarish imkonini beradi.

Sun’iy intellekt kutubxonalari yordamida ma’lum bir vazifa bajarish va matnga ishlov berish bir qancha bosqichlarni o‘z ichiga oladi. Quyida bu bosqichlarni ko‘rib chiqamiz.

Tokenizatsiya va to‘ldirish(padding).

Tabiiy tilni qayta ishlash (NLP) texnologiyalarining rivojlanishi bilan matnni raqamli formatga aylantirish usullari muhim ahamiyat kasb etmoqda. “Tokenizatsiya” va “to‘ldirish” kabi jarayonlar NLP algoritmlarining samaradorligi va ishlash tezligini ta’minlovchi asosiy tarkibiy qismlardir. Matnni qayta ishlashning birinchi bosqichi xom matnni mashinani o‘qitish modeli tushunadigan shaklga aylantirishni o‘z ichiga oladi. Bu jarayon *tokenizatsiya* bilan boshlanadi, u matnni

tokenlar deb ataladigan kichikroq birliklarga, odatda, so‘zlar, bo‘g‘inlar yoki harflarga ajratadi.

Tokenizatsiya – bu matnni kichik bo‘laklarga ajratish jarayonidir. Bu kichik bo‘laklar **tokenlar** deb ataladi. Tokenlar, odatda, so‘zlar, bo‘g‘inlar yoki boshqa ma’lumotli segmentlar bo‘lishi mumkin. Tokenizatsiya tabiiy tilni qayta ishlashning birinchi va eng muhim bosqichlaridan biridir. Tokenizatsiya jarayoni matn ustida bajariladigan amallarga bog‘liq holda bir qancha turlarga bo‘linishi mumkin (Qarang: 1-jadval). Masalan, *bolalar bugun maktabga ketishdi* gapini turli xil tokenlarga ajratish mumkinligini quyidagi jadvalda ko‘rish mumkin:

1-jadval. Token turlari

Token turlari	**Token misoli**
Gap shaklida	Bolalar bugun maktabga ketishdi
So‘zshakl	Bolalar, bugun, maktabga, ketishdi
Asos va qo‘shimcha	Bola, lar, bugun, maktab, ga, ket, ish, di
Bo‘g‘in shaklida	Bo, la,lar, bu, gun, mak, tab, ga, ket, ish, di
Harf shaklida	B, o, l, a, l, a, r, b, u, g, u, n, m, a, k, t, a, b, g, a, k, e, t, i, sh, d, i

Yuqoridagi jadvaldan ko‘rinib turibdiki, matnlarni turli shaklda tokenizatsiya qilish mumkin ekan. Bu, albatta, matn ustida bajariladigan vazifaga bog‘liq bo‘ladi. Masalan, harf shaklidagi tokenizatsiya fonetik tahlil, yozuv qoidalari va shu kabi xususiyatlarni o‘rganishda ishlatiladi. Yoki so‘zshakl darajasidagi tokenizatsiya matn tasnifi mashina tajrimasi kabi vazifalar uchun ishlatiladi.

TensorFlow kutubxonasida tokenizatsiya modullari va ularning bir qancha klasslari mavjud. Quyida ularning bir qismini ko‘rib chiqamiz.

tf.keras.preprocessing.text moduli. Bu modul **TensorFlow** kutubxonasida matnlarni qayta ishlash uchun taqdim etilgan yordamchi funksiyalardan iborat. Tokenizer klassi matnni butun sonlar ketma-ketligiga aylantirish orqali matnni oldindan qayta ishlash uchun ishlatiladi, bu erda matndagi har bir so‘z yoki token ma’lumotlar to‘plamidagi chastotasi asosida noyob butun songa ko‘rsatiladi[1]. Bu modulda matnlarni **tokenizatsiya qilish, ketma-ketlikka aylantirish** va **NLP** modellar uchun tayyorlash jarayonini osonlashtirish uchun ko‘plab funksiyalar mavjud. Undagi “Tokenizer” klassi tokenizatsiya uchun keng qo‘llaniladi. “Tokenizer” matn korpusini butun sonlar ketma-ketligiga aylantiradi, bunda har bir butun son alohida so‘z yoki tokenni ifodalaydi. Masalan, “TensorFlow mashinali o‘qitishda mashhur kutubxona” degan oddiy jumla berilgan bo‘lsa, tokenizatsiya uni [“TensorFlow”, “mashinali”, “o‘qitishda”, “mashhur”, “kutubxona”] so‘zlariga bo‘ladi, so‘ngra ular korpusdagi chastotasiga qarab mos keladigan butun son

indekslariga aylantiriladi, ya’ni [[2, 3, 1, 4, 5]] shaklidagi natijani beradi. Bu modelga xom matn emas, balki raqamli tasvirlar bilan ishlash imkonini beradi. UnicodeCharTokenizer klassi matnni belgilarga bo‘lib tokenlashtiradi. Ya’ni matnni individual belgilar (harflar, raqamlar, maxsus belgilar) shaklida tokenlarga ajratadi. Bunda gap, so‘zlar yoki bo‘g‘inlar emas, balki har bir belgi alohida token sifatida ishlatiladi. WhitespaceTokenizer klassi bo‘sh joylarni ajratuvchi element sifatida ko‘rib, har bir bo‘sh joydan keyin yangi token yaratadi. Bu tokenizatsiya turi, asosan, so‘zlar orasidagi bo‘sh joylar bo‘yicha ajratish uchun ishlatiladi.

torchtext moduli. Bu modul PyTorch kutubxonasida matnlarni qayta ishlash uchun mo‘ljallangan maxsus vositalarni o‘z ichiga oladi. torchtext yordamida matnlarni tokenizatsiya qilish, lug‘at (vocabulary) yaratish, embedding (semantik vektorlar) hosil qilish va matnlar bilan ishlovchi datasetlar yaratish imkoniyati mavjud. torchtextdan foydalanib, siz matnli ma’lumotlarni qayta ishlash jarayonini avtomatlashtirishingiz mumkin. Bu jarayon tokenizatsiya, so‘zlarni raqamlarga aylantirish, lug‘at yaratish va hokazolarni o‘z ichiga oladi. Bu xususiyat bu amalni bajarish uchun murakkab kodlarni yozish shart emasligini bildiradi[2]. Bu moduldagi get_tokenizer() funksiyasi yordamida ingliz tilida keng qo‘llaniladigan tokenizatorlar, jumladan basic_english, spacy, moses va boshqa variantlar bilan matnni bo‘lish mumkin. Vocab klassi esa barcha tokenlarni indekslash va ularning embedding ifodalarini saqlash uchun ishlatiladi. Masalan, "mashina o‘rganish sohasida PyTorch juda samarali" degan jumla get_tokenizer() yordamida [“mashina”, “o‘rganish”, “sohasida”, “PyTorch”, “juda”, “samarali”] shaklida tokenlarga ajratiladi, so‘ngra ular Vocab yordamida indekslanadi va nn.Embedding orqali neyron tarmoqlar uchun tayyorlanadi. torchtext shuningdek, maxsus TextClassificationDataset, DataLoader va BucketIterator funksiyalarini taqdim etadi, bu esa yirik matnlar ustida o‘qitish jarayonini samarali tashkil etish imkonini beradi. PyTorch tokenizatsiyani fleksib tarzda sozlash imkonini berishi bilan boshqa kutubxonalardan ajralib turadi, ayniqsa LSTM, GRU, Transformer modellar bilan ishlaganda bu juda muhimdir.

scikit-learn.feature_extraction.text moduli. SciKit Learn - bu mashinani o‘rganish loyihalari uchun keng kutubxona bo‘lib, bir nechta tasniflagich va tasniflash algoritmlari, o‘qitish va ko‘rsatkichlarni yig‘ish usullari va kiritilgan ma’lumotlarni oldindan qayta ishlash uchun mo‘ljallangan[3]. Bu modul Scikit-learn kutubxonasida matnlarni raqamlashtirish va oldindan ishlashga ixtisoslashgan komponentlarni o‘z ichiga oladi. Asosan, CountVectorizer va TfidfVectorizer klasslari matnni vektor ko‘rinishiga keltirishda keng qo‘llaniladi. CountVectorizer matndagi har bir so‘zni indekslab, ularning chastotasiga qarab vektor yaratadi. Masalan, "Scikit-learn matnli tahlil uchun foydali vosita" jumlasi quyidagicha vektorga aylanishi mumkin: [“scikit”, “learn”, “matnli”, “tahlil”, “uchun”, “foydali”, “vosita”] → [1, 1, 1, 1, 1, 1, 1]. TfidfVectorizer esa so‘z chastotasi va

umumiy korpusdagi kamyoblik darajasini hisobga olib, har bir so‘zga mos og‘irlik (weight) beradi. Bu metod ayniqsa soxta yangiliklarni aniqlash, ijtimoiy tarmoq postlarini tahlil qilish, yoki hujjatlarni kategoriyalarga ajratishda juda foydalidir. Scikit-learn kutubxonasi shuningdek, tayyorlangan vektorlarni MultinomialNB, SVM, LogisticRegression va boshqa klassifikatorlarga uzatish imkonini beradi. Bu modul matnlarni qayta ishlashni chuqur o‘rganish tarmoqlarisiz bajarishni xohlagan foydalanuvchilar uchun yengil va tezkor yechim taqdim etadi. Shuningdek, Pipeline va GridSearchCV kabi vositalar orqali butun ishlov jarayonini avtomatlashtirish mumkin.

tf.keras.preprocessing.sequence moduli TensorFlow Keras kutubxonasining bir qismi bo‘lib, ketma-ketlik ma’lumotlari bilan ishlash vositalarini taqdim etadi. U mashinani o‘qitish modellari uchun matnli yoki ketma-ket ma’lumotlarni tayyorlash uchun yordamchi dasturlarni, ayniqsa, tabiiy tilni qayta ishlash (NLP) vazifasini o‘z ichiga oladi. Bu moduldagi TimeseriesGenerator klassi ketma-ket ma’lumotlarni qayta ishlash uchun, xususan, prognozlash, anomaliyalarni aniqlash yoki tasniflash kabi davriy vazifalar uchun yordamchi dasturdir. Bu LSTM, GRU yoki CNN kabi modellarga vaqt seriyasidagi ma’lumotlarni uzatishni osonlashtirib, siljish oynalari ketma-ketligini va ularga mos keladigan maqsadlarni yaratishga yordam beradi. Ya’ni bu klass kutilayotgan natijani bashorat qilishga yordam beradi. NgramTokenizer klassi esa ketma-ket elementlardan tashkil topgan n-gramm tokenlarini yaratuvchi tokenizatorni chaqiradi. N-gramm tokeni ketma-ketlikdagi n ta elementdan tashkil topgan tokendir.

torch.nn.utils.rnn va **torchtext** modullari PyTorch kutubxonasida ketma-ketlik ma’lumotlari bilan ishlashda muhim rol o‘ynaydi. Har qanday chuqur o'rganish modeli torch.nn modulining quyi sinfi yordamida ishlab chiqilgan bo'lib, u chiqishni qaytaruvchi oldinga (kirish) kabi usuldan foydalanadi[4]. Bu modullar yordamida NLP va vaqt ketma-ketligi asosidagi vazifalar uchun matnli ma’lumotlarni samarali tayyorlash mumkin. torch.nn.utils.rnn.pad_sequence funksiyasi turli uzunlikdagi ketma-ketliklarni bitta batchga joylashtirish uchun ularni maksimal uzunlikka tenglashtirib (padding), tensor holatiga keltiradi. Bu esa LSTM yoki GRU kabi modellar uchun bir xil o‘lchamdagi kiritish (input) yaratishga imkon beradi. pack_padded_sequence va pad_packed_sequence funksiyalari esa modelga joylashtirilgan (packed) yoki asl shakliga qaytarilgan (unpacked) ma’lumotlar bilan ishlash imkonini beradi. Shuningdek, torchtext.data moduli ichida mavjud bo‘lgan BucketIterator va Field klasslari yordamida ma’lumotlar optimallashtirilgan tarzda modelga uzatiladi. Bu komponentlar ayniqsa, vaqt seriyalari, til modellari yoki ketma-ketlik asosida klassifikatsiya vazifalarida muhim ahamiyatga ega. torchtext.transforms.NGramTransform esa matndan n-gram tokenlar yaratish imkonini beradi, bu esa murakkab til tuzilmalarini modelga tushuntirishda foydalidir.

sklearn.feature_extraction.text va **sklearn.model_selection** modullari Scikit-learn kutubxonasida matn va vaqt ketma-ketligi asosidagi ma’lumotlarni qayta ishlash uchun foydalaniladi. CountVectorizer va TfidfVectorizer klasslari yordamida matnli ketma-ketliklarni vektorlashtirish, ya’ni raqamli ifodalarga aylantirish mumkin. Ayniqsa TfidfVectorizer N-gram parametrlarini belgilash orqali ketma-ketlikdagi tuzilmalarni (masalan, bigram yoki trigram) aniqlash imkonini beradi. Bu matnlardagi kontekstni chuqurroq tushunishga xizmat qiladi. Vaqt asosidagi vazifalarda TimeSeriesSplit klassi yordamida modelni o‘qitish va test qilish uchun vaqt tartibida bo‘lingan datasetlar yaratish mumkin. Bu ayniqsa prognozlash va vaqtga bog‘liq klassifikatsiya vazifalari uchun foydalidir. Scikit-learn modullari LSTM yoki GRU kabi chuqur tarmoqlarni to‘g‘ridan-to‘g‘ri taqdim etmasa-da, statistik asoslangan modellarga tayyorlangan ketma-ketliklarni berish orqali NLP vazifalarini muvaffaqiyatli bajarishga xizmat qiladi. Bu kutubxona o‘zining soddaligi va mustahkam poydevori bilan mashhur.

tf.keras.preprocessing.image moduli rasm ma’lumotlarini oldindan ishlash uchun ishlatiladi. U rasmlarni disk yoki veb saytlardan yuklash, uning o‘lchamlarini o‘zgartirish, normalizatsiya qilish va modelga tayyorlash uchun bir qator yordamchi funksiyalar va sinflarni keltiradi. ImageDataGenerator klassi, asosan, tasvirlarni oldindan ishlash va real vaqt rejimida ma’lumotlarni boyitish vazifasini bajaradi. Bu klass mashinani o‘rgatish jarayonida tasvirlar bilan ishlashni osonlashtiradi va modelning umumlashgan qobiliyatini oshirishga yordam beradi. Bu klassning asosiy vazifalari tasvirlarni neyron tarmoqqa kiritishdan oldin qayta ishlaydi, tasvirlarni kerakli o‘lchamlarga keltiradi, modelni turli xil sharoitlarga moslashtirish uchun real vaqt rejimida tasvirlarni o‘zgartiradi, ma’lumotlarni boyitadi (ya’ni aylantirish, kesish, o‘zgartirish, gorizontal/vertikal shaklda aylantirish va shu kabilar), katta hajmdagi tasvirlar bilan ishlash (diskdagi katta hajmdagi tasvirlarni yuklamasdan, real vaqt rejimida guruhlab yuklab o‘rganish imkonini beradi).

Biroq real dunyo matn ma’lumotlari uzunligi bo‘yicha farq qilishi mumkin va neyron tarmoqlar, odatda, belgilangan o‘lchamdagi kirish ma’lumotlarini kutishadi. Buni hal qilish uchun barcha ketma-ketliklarning bir xil uzunlikka ega bo‘lishini ta’minlash uchun *to‘ldirish*(padding) ishlatiladi. To‘ldirish juda muhim, chunki mashinani o‘qitish modellari, ayniqsa, neyron tarmoqlari, odatda, ma’lum hajmdagi kirishlarni talab qiladi. To‘ldirish turli uzunlikdagi ketma-ketliklarni bir xil uzunlikka keltirishni ta’minlaydi, bu esa partiyalarni qayta ishlashga imkon beradi. TensorFlow eng uzun ketma-ketlikning uzunligiga mos keladigan qisqaroq ketma-ketliklarni nol (yoki boshqa qiymatlar) bilan to‘ldirish uchun “pad_sequences” funksiyasini taqdim etadi. Bu partiyalarni qayta ishlash va samarali o‘qitish uchun juda muhimdir.

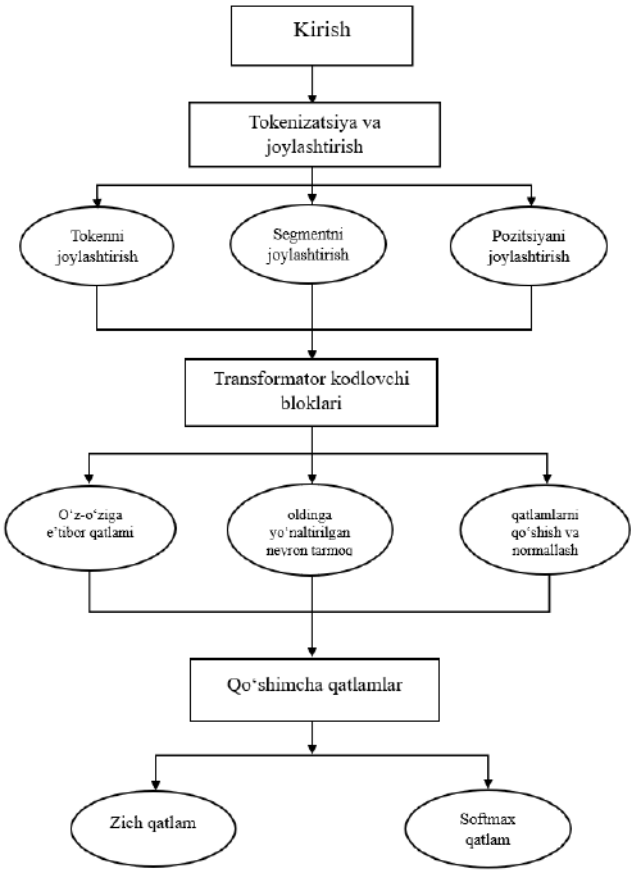
Matnni vektorlashtirish: Matnni raqamli tasvirlarga aylantirish.

Matn tokenizatsiya qilingandan so‘ng, keyingi qadam bu tokenlarni raqamli vektorlarga aylantirishdir. Ushbu transformatsiya matnning semantik ma’nosini saqlab qolgan holda neyron tarmoqqa matnli ma’lumotlar bo‘yicha hisob-kitoblarni amalga oshirish imkonini beradi. Keng tarqalgan yondashuvlardan biri *so‘zlarni joylashtirish* bo‘lib, ular har bir so‘zni uzluksiz vektor fazosida zich vektor sifatida ifodalaydi. Matnni vektorlashtirish juda muhim, chunki kompyuterlar faqat raqamlar bilan ishlaydi. Matnni raqamlarga aylantirish orqali, kompyuter modeliga matnni o‘rganish, uning xususiyatlarini aniqlash va tasniflash mumkin bo‘ladi. Bu jarayonni amalga oshirishda TensorFlow kuchli vosita hisoblanadi. TensorFlowda matnni vektorlashtirish uchun ko‘plab usullar mavjud bo‘lib, ulardan eng ko‘p ishlatiladigan usul TextVectorization qatlamidan foydalanishdir. Bu qatlam matnni sonli vektorlarga aylantiradi va matnni modelga kiritish uchun tayyorlaydi. Shuningdek, bu qatlam matnni tokenizatsiya qilish, vektorlashtirish, to‘ldirish va oldindan ishlov berish kabi vazifalarni amalga oshiradi. Bu esa kodni ixsham va soda qiladi. Bundan tashqari, matnni vektorlashtirishda so‘zlarning ma’nosini yaxshiroq ifodalash uchun *so‘z joylashtirish* funksiyasidan foydalaniladi. Joylashtirishlar har bir so‘zni yuqori o‘lchamdagi (masalan, 100 yoki 300 o‘lchamli) vektor bilan ifodalaydi. Bu usul so‘zlarning semantik ma’nosini saqlashga yordam beradi. Joylashtirish so‘zlar orasidagi semantik munosabatlarni qamrab oladi, masalan, o‘xshash ma’noga ega so‘zlar o‘xshash vektor ko‘rinishlariga ega bo‘ladi.

TensorFlow-da “tf.keras.layers”dagi “Embedding(Joylashtirish)” qatlami trening davomida ushbu joylashtirishlarni o‘rganish uchun ishlatiladi. Bu qatlam butun sonlar ketma-ketligini (tokenlashtirilgan so‘zlar) oladi va ularni ma’lum o‘lchamdagi zich vektorlar bilan taqqoslaydi. Joylashtirish qatlami, odatda, xom ma’lumotlarni, ya’ni so‘zlar, matnlar, tasvirlar yoki shunga o‘xshash elementlarni tahlil qilish va model uchun tushunarli shaklga o‘tkazish uchun xizmat qiladi. Joylashtirish qatlamlari juda katta hajma ega bo‘lgan xom ma’lumotlarning hajmini qisqartirib, modelga samarali tarzda ishlatish imkonini beradi. Misol uchun, model “shoir” so‘zini joylashtirish maydonida “yozuvchi” so‘ziga yaqin vektor sifatida ko‘rsatadi. Chunki bu so‘zlar bir-biriga o‘xshash konterkstlarda ishlatiladi.

Tokenizatsiya va vektorlashtirish jarayonlari tugagandan so‘ng qayta ishlangan matn mashinani o‘qitish modellariga kiritishga tayyor bo‘ladi.

TensorFlowda matnli modellarni yaratish. TensorFlow matnli ma’lumotlar uchun kuchli modellar yaratishga imkon beruvchi ochiq manbali mashinani o‘qitish kutubxonasidir. Oldindan ishlov berish va vektorlashtirishdan so‘ng, qayta ishlangan ma’lumotlar mashinani o‘qitish modeliga kiritilishi mumkin. TensorFlowda matnli modelni yaratish uchun bir qancha ilg‘or arxitekturalardan foydalanish mumkin. Shulardan biri Hugging Face Transformatorlar kutubxonasida ishlatiladigan BERT modeli arxitekturasidir(Qarang: 2.1.1-chizma).



1-chizma. BERT modeli arxitekturasi

Diagramma transformator modellarining ishlash jarayonini tasvirlaydi. Matn avval tokenizatsiya qilinib, token, segment va pozitsiya embeddinglari joylashtiriladi. So‘ngra, o‘z-o‘ziga e’tibor qatlami va oldinga yo‘naltirilgan neyron tarmoq orqali kodlanadi. Qatlamlarni qo‘shish va normallashtirish barqarorlikni oshiradi. Yakunda zich qatlam va Softmax qatlami natijani hisoblab, tasniflashni amalga oshiradi. Bu model NLP vazifalari uchun samarali ishlatiladi.

Model arxitekturasi aniqlangandan so‘ng, TensorFlowning yuqori darajali API “tf.keras” modelni kompilyatsiya qilishni, yo‘qotish funksiyasini va optimallashtiruvchini belgilashni va modelni ma’lumotlar to‘plamida o‘rgatishni osonlashtiradi. TensorFlow, shuningdek, aniqlik yoki F1 balli kabi standart ko‘rsatkichlar yordamida modelning ish faoliyatini oson baholash imkonini beradi.

Xulosa

Xulosa qilib aytganda, matnni qayta ishlash, NLPning asosiy yo‘nalishlaridan biri bo‘lib, mashinalarga inson tilini tushunish va qayta ishlash imkonini beradi.

Sun’iy intellektning turli ochiq manbali kutubxonalari matnni tasniflash, hissiyotlarni tahlil qilish va mashina tarjimasi kabi vazifalarda muhim ahamiyatga ega.

Matnni qayta ishlashda tokenizatsiya, to‘ldirish va vektorlashtirish asosiy jarayonlardir. Tokenizatsiya matnni kichik bo‘laklarga ajratib, modelga tahlil qilish imkonini beradi. To‘ldirish matn uzunliklarini tenglashtiradi, vektorlashtirish esa matnni raqamli formatda ifodalaydi. Sun’iy intellekt kutubxonalari bu jarayonlarni samarali amalga oshirish uchun kerakli vositalarni taqdim etadi.

BERT modeli va joylashtirish usullari semantik munosabatlarni yaxshilaydi, so‘zlar orasidagi o‘xshashlikni aniqlashga yordam beradi, bu esa matnni tahlil qilishda samarali natijalar beradi.

Foydalanilgan adabiyotlar:

1. Bilal Sharma. Tokenizer and TextVectorization? Difference, 2024.
2. https://www.shirdell.ir/index.php/machine-learning/138-what-is-torchtext-and-what-features-does-it-provide-to-developers
3. Sebastian. NLP: Text vectorization methods using Scikit learn. Berlin, 2019.
4. https://www.educba.com/torch-dot-nn-module/

UDK: 811.512.133`42

LINGVISTIK TADQIQOTLARDA YANGI YONDASHUV: ABDULLA QODIRIYNING “O‘TKAN KUNLAR” ROMANI ASOSIDA YARATILGAN PARALLEL KORPUSI VA UNING TAVSIFI

O‘tkirova Fotima Baxtiyor qizi,
Kompyuter lingvistikasi mutaxassisligi magistranti
otkirovafotima@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada korpus lingvistikasi sohasining bir qismi bo‘lgan parallel korpus va uning tilshunoslikdagi ahamiyati haqida so‘z boradi. Lingvistika va adabiyot o‘rtasidagi ilmiy izlanish va muammolarga yangicha yondashuv orqali yechim berish to‘g‘risida ma’lumot beriladi. Abdulla Qodiriyning “O‘tkan kunlar” romani asosida yaratilgan o‘zbek-rus hamda rus-o‘zbek parallel korpusi tavsifi yoritiladi. Asarning har ikki korpusda ifodalangan til imkoniyatlari, muallif hamda tarjimon qalamiga xos xususiyatlar ko‘rsatiladi.

Abstract. This article discusses the parallel corpus, which is part of the field of corpus linguistics, and its significance in linguistics. Information is provided on scientific research and solutions to problems between linguistics and literature through a new approach. The description of the Uzbek-Russian and Russian-Uzbek parallel corpus created on the basis of Abdulla Qodiriy’s novel “O‘tkan kunlar” is highlighted. The linguistic possibilities of the work, expressed in both corpora, and the peculiarities of the author’s pen are shown.

Аннотация. В данной статье рассматривается параллельный корпус, являющийся частью корпусной лингвистики, и его значение в лингвистике. Предоставляется информация о научных исследованиях и решении проблем между лингвистикой и литературой посредством нового подхода. Освещается описание узбекско-русского и русско-узбекского параллельного корпуса, созданного по роману Абдуллы Кадыри “Минувшие дни” Указываются языковые возможности произведения, выраженные в обоих корпусах, особенности, присущие перу автора.

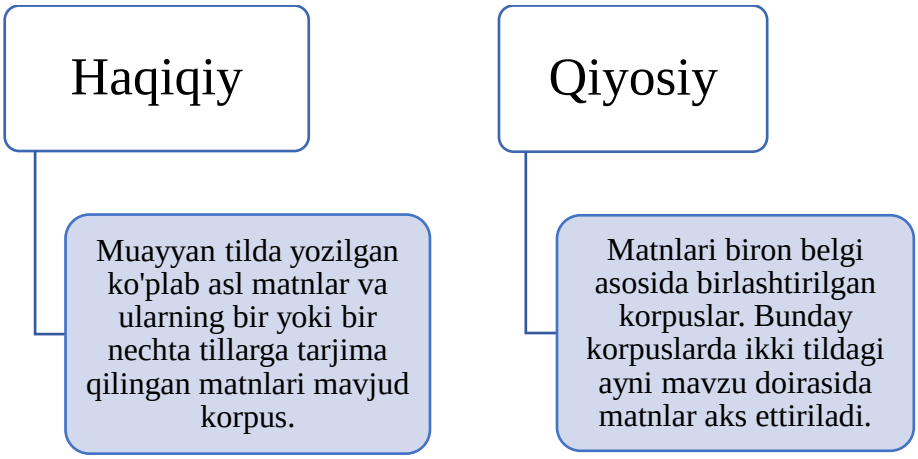
Kalit so‘zlar: Korpus lingvistikasi, korpus, parallel korpus, tezaurus, aligner tizimi, n-grammlar, konkordans.

Keyingi yillarda lingvistik tadqiqotlar sohasida yangicha yondashuvlar hamda metodologiyalarning paydo bo‘lishi tilshunoslikni yangi bosqichga olib chiqdi. Bu yangiliklardan biri parallel korpusning yaratilishi va uning lingvistik izlanishlarda keng foydalanishning yo‘lga qo‘yilishi bo‘ldi. **Parallel korpus** – biror tilning matnlarini boshqa tilga tarjima qilingan matnlar bilan o‘z ichiga olgan maxsus

ma’lumotlar bazasi sanaladi. Bunday korpus turi tilshunoslarga nafaqat tarjima jarayonini tahlil qilish, balki tilning leksik, sintaktik, morfologik va semantik xususiyatlarini chuqur o‘rganish imkonini yaratadi.

Parallel korpuslar yoxud **tarjimalar korpusi** (ingl. *parallel / translation corpus*) – bu asl tildagi matnlar va uning boshqa tildagi tarjimasi mavjud korpus. Qoidaga ko‘ra, bunday korpuslar faqat ikkita tilda tayyorlanadi, shu bois undagi o‘zaro mutanosib matnlar *bitekst* ham deyiladi. **Bimatn (bitekst)** – asliy til matni va uning boshqa til(lar)dagi tarjimasi mavjud matnlar [1].

Parallellik nizomiga ko‘ra korpuslar ikki asosiyga turga bo‘linadi:



Har ikki parallel korpus ham lingvistik tadqiqotlarda muhim ahamiyat kasb etib bormoqda. Jumladan, tarjima tadqiqotlari, leksik tahlil, sintaktik tahlil, stilistik tahlil borasida til imkoniyatlari va tarjimonlarning tarjima usullarini aniqlash, ikki til borasidagi qiyosiy tahlillarni amalga oshirish hamda taqqoslashga yordam beradi. Parallel korpus yaratish jarayonida ba’zi qiyinchiliklar mavjud. Birinchi navbatda, yuqori sifatli tarjimalar topish va ularni tahlil qilishda yuzaga keladigan muammolarni keltirib chiqaradi. Ikkinchidan, ko‘p tilli korpuslarni yaratish tildagi matnlarning o‘zaro muvofiqlashgan(alignment) shaklini talab qiladi. Shunga qaramay, Parallel korpusning bir qator afzalliklarini sanab o‘tamiz. Masalan, zamonaviy va keng qamrovli yondashuvlarni amalga oshiradi. Bu esa bir necha tillar yuzasidan keng qamrovli tahlil olib borish mumkinligini anglatadi. U tilni o‘rganish va tarjimani tahlil qilishda qo‘llaniladi. Tarjimaning to‘g‘riligi va mazmuni parallel matnlar yordamida baholanishi tarjima sifatini belgilashda qo‘l keladi. Bunday korpuslar yordamida tilning semantik tizimi va so‘zlarning kontekstual ma’nolarini o‘rganish mumkin.

Badiiy asar bo‘yicha parallel matnlar korpusi, avvalo, ikki yoki bir necha tillar grammatikasidagi o‘xshash va farqli jihatlarni tahlil qilish imkonini beradi. Parallel matnlar korpusi har ikki tildagi badiiy uslub xususiyatlarini qiyoslashda

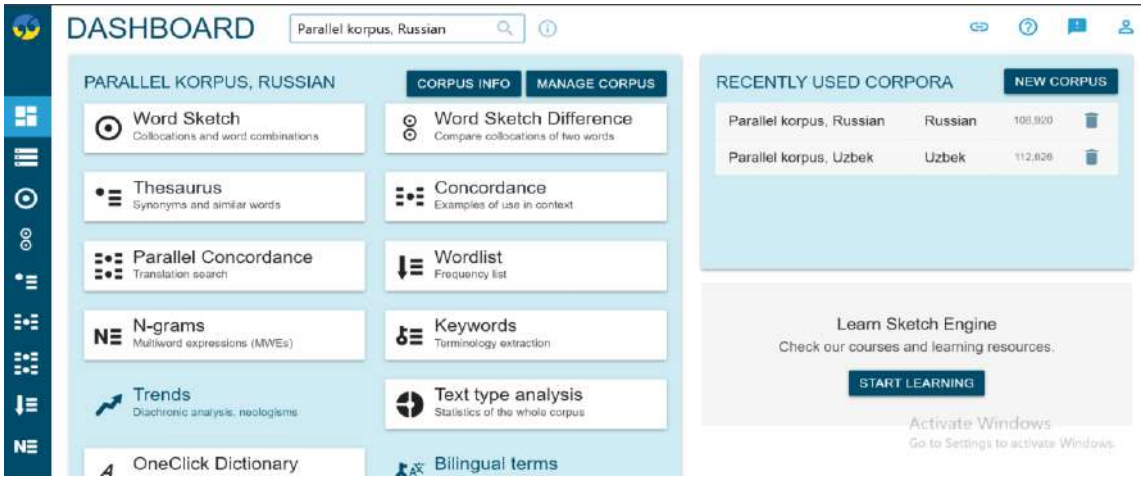
ahamiyatlidir [2]. Parallel korpuslar badiiy asarlarning original va tarjima qilingan variantlarini birlashtirib, tarjimadagi o‘zgarishlarni, til va madaniyatdagi farqlarni, leksik va sintaktik strukturalarni solishtirishga imkon beradi.

Quyida yaratilgan Abdulla Qodiriyning “O‘tkan kunlar” romani va uning rus tiliga tarjima qilingan nusxasi asosidagi parallel korpus ham tilshunoslik, tarjimashunoslik va adabiyotshunoslik uchun muhim ahamiyatga ega. Parallel korpus tuzishda asarning o‘zbek tilida 2007-yilda nashrdan chiqqan nusxasi tanlab olindi. Rus-o‘zbek parallel korpusi uchun Muhammadnodira Safarova tomonidan “Минувшие дни” nomi ostida 2009-yilda nashr qilingan nusxasi tanlab olindi. Korpusni yaratishda “SketchEngine” dasturidan foydalanilgan. Korpus haqidagi statistik ma’lumotlarni quyidagi jadvallarda ko‘rish mumkin.

1-jadval. O‘zbek-rus hamda Rus-o‘zbek parallel korpusi ma’lumotlari

№	O‘zbek-rus PK	Rus-O‘zbek PK
Words	112,626	108,920
Tokens	158,194	140,759
Sentences	8,651	7175
Paragraphs	728	782

Ushbu jadvalda bir asarning ikki tildagi nusxalaridagi farqlarni yaqqol ko‘rish mumkin. Jumladan, parallel korpusda asar rus tiliga tarjima qilinganda o‘zbek tilidagi nusxasida mavjud so‘zlarning 4000 tadan ko‘proqqa qisqarganini ko‘rish mumkin. Yoki tarjima nusxada gaplarning qariyb 1500 taga qisqargani ko‘zga tashlanadi. Parallel korpus tuzishdan avval asarning ikki tildagi ayni nusxalarining elektron formatlari topib olindi. Keyin dasturga muvaffaqiyatli yuklanib, ikki tilli parallel korpus ishga tushdi (1-rasm).



1-rasm. Parallel korpus interfeysi

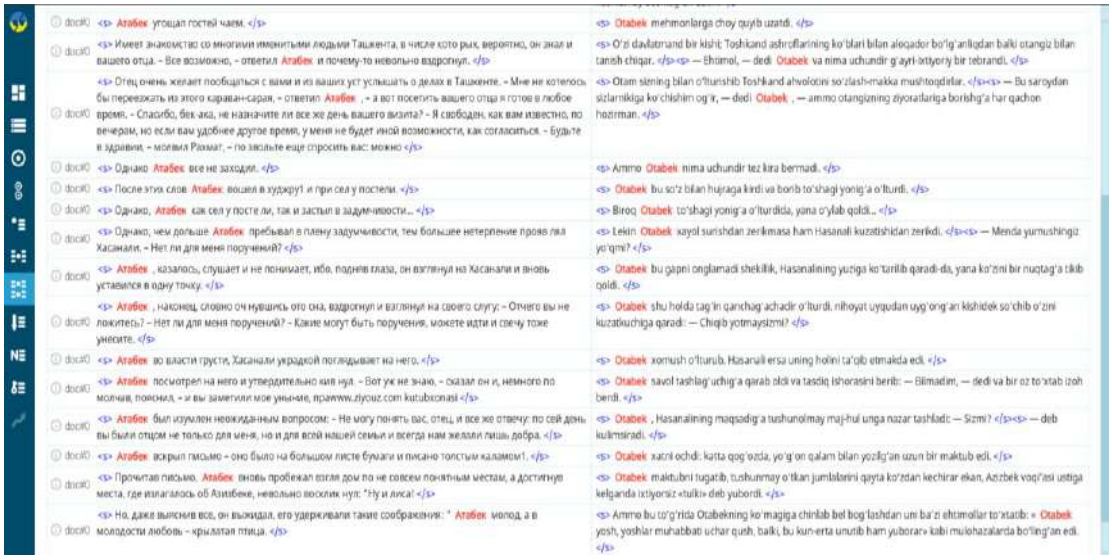
Asar tahliliga kirishishdan avval kalit so‘zlarni topib olish kerak bo‘ladi. Quyidagi jadvalda ikki korpusda topilgan dastlabki 15 ta kalit so‘zlarni ko‘rish mumkin (2-jadval).

2-jadval. Parallel korpusdagi kalit so‘zlarning qiyosiy tahlili

O‘zbek-rus PK			Rus-o‘zbek PK		
Kalit so‘zlar	**Takrori**	**Baholash**	**Kalit so‘zlar**	**Takrori**	**Baholash**
hasanali	189	1148.97	атабек	650	4555.61
qutidor	202	935.314	кумуш	554	3922.5
yusufbek	139	684.623	Kutubxonasi	430	3055.87
yonig	107	626.397	хасанали	276	1961.45
andek	115	626.039	зайнаб	283	1851.61
ach	110	578.605	атабека	258	1816.63
qolg	119	535.356	узбек-аим	200	1421.87
homid	155	528.829	офтоб-аим	176	1251.14
toshkand	113	523.543	кутидор	172	1222.95
chiqg	86	523.337	кушбеги	120	834.737
andan	75	475.101	хамид	137	804.019
ketkan	70	395.237	курбаши	91	631.597
tkan	282	389.307	маргилан	85	598.323
anlig	58	367.638	кутидора	78	555.139
mundan	67	365.094	Алим	87	509.835

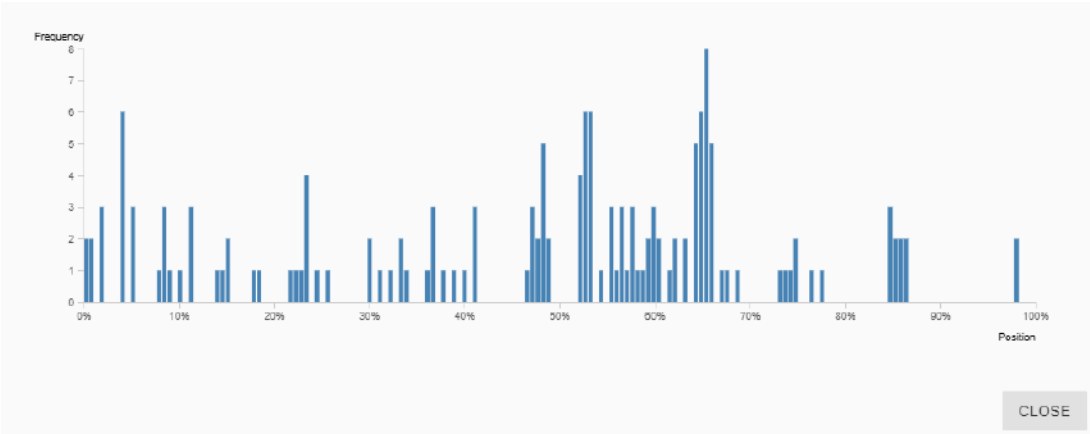
Ushbu jadvalni taqqoslaganda o‘zbek-rus va rus-o‘zbek parallel korpusidagi kalit so‘zlar tarjimada o‘zgarganini ko‘rish mumkin. Jumladan, asl tilda *“Qutidor”, “Xasanali”, “Homid”, “Toshkand”* kabi so‘zlar eng ko‘p qo‘llanilgan bo‘lsa, rus tiliga tarjimada *“Атабек”, “Кумуш”, “Хасанали”, “Зайнаб”* kabi shaxs nomlarida ko‘p takrordan foydalanilgan. Ushbu qiyosiy tahlildan yana bir xulosaga kelish mumkin bo‘ladi. Asarning tarjima nusxasida kishilarning lavozim, mansab hamda laqablari o‘zbek tilidan to‘g‘ridan-to‘g‘ri olib kirilgan. Tarjimon o‘zbek xalqi madaniyati va tarixiga oid ushbu so‘zlar ustida tarjimaga qo‘l urmagan.

Parallel korpusda badiiy asar tahliliga oid muhim jihat prallel konkordansning mavjudligidir. Har ikki asarning gaplar muvofiqligi-yu, ikki tilga xos bo‘lgan leksik tuslanishlar ushbu tahlil jarayonida yuzaga chiqadi. Bunda parallel korpus kalit so‘zlarida yuzaga chiqqan so‘zlarni tanlasak, tahlil yana-da chuqurroq amalga oshiriladi. Masalan, “Otabek” so‘zini qidirdik. Bunda biz konkordans lug‘atda uchta tahlilni amalga oshirdik. Birinchidan, matnlar korpusda to‘g‘ri muvofiqlashtirilganligi namoyon bo‘ldi (2-rasm).

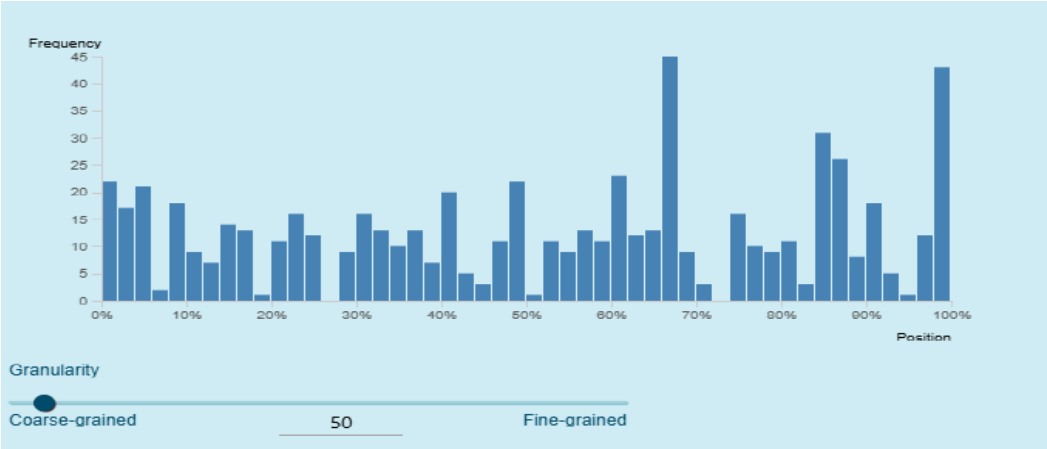


2-rasm. Matnlar muvofiqlashgan(alignment) parallel korpus

Ikkinchi qiyosiy tahlil *“Otabek”* shaxs nomining ikki korpusda ham asarda qo‘llanilganlik darajasini diagrammada ifodalanishi bo‘ldi (3-rasm). Ikki korpus diagrammasida ham ko‘rish mumkinki, asarning 70% qismiga yaqinlashganda har ikki asarda ham *“Otabek”* ismidan eng yuqori darajada foydalanilgan. Va, aksincha, o‘zbek-rus parallel korpusida ushbu ismga asarning 85 foizidan boshlab oxirigacha ko‘p takrorda murojaat qilingan bo‘lsa (4-rasm), rus-o‘zbek parallel korpusida tarjima asarning 85% dan to oxirigacha bo‘lgan qismda umuman tilga olinmagan.

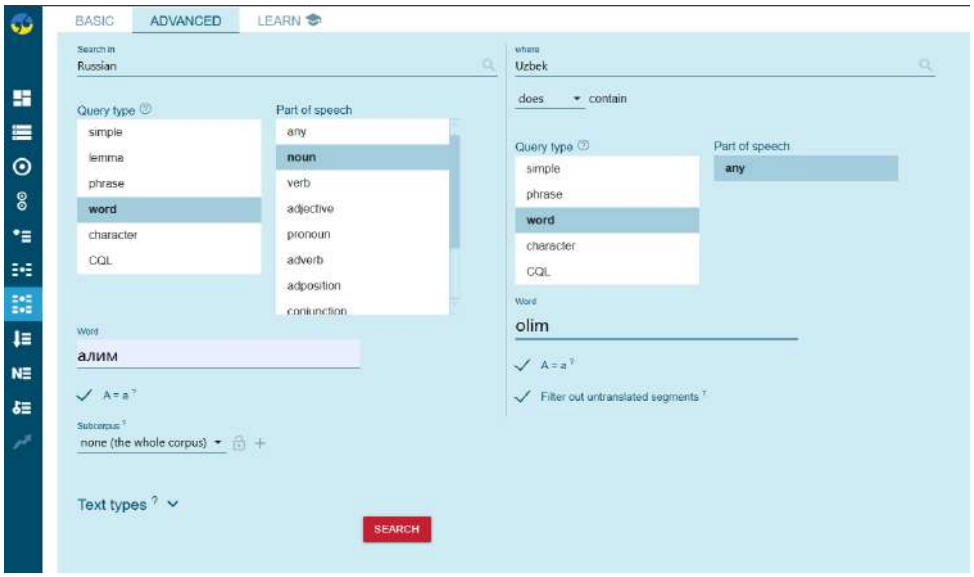


3-rasm. Rus tili parallel korpusida “Амабек” shaxs nomining qo‘llanilishi



4-rasm. O‘zbek tilli parallel korpusda “Otabek” shaxs nomining qo‘llanilishi

Keyingi qiyosiy tahlil parallel korpusda rus tili va o‘zbek tilidagi so‘zlarni qidirish imkoniyatlarini ko‘rsatib berdi. Masalan, har ikki tilli korpusda ham qidiruvni yakka, fraza, so‘z va hokazo shakllar ustida amalga oshirish mumkinligi ko‘rinadi. Biroq keyingi bosqichlarda rus tili parallel korpusining imkoniyatlari kengayadi. Ya’ni qidiruvdagi birlikning so‘z turkumini, o‘rnini aniqlash mumkin bo‘ladi. O‘zbek tilidagi asar korpusida esa so‘z turkumlarini belgilash imkoni mavjud emasligi ayon bo‘ldi (5-rasm).



5-rasm. Parallel konkordans qidiruv imkoniyatlari

Keyingi tahlilni amalga oshirish ham parallel korpus imkoniyatlarini ko‘rsatib beruvchi muhim jihatdir. Bu N-grammlar asosida matnda qo‘llanilgan turli leksik birliklarning bog‘lanish shakllarini topish imkoni mavjud. Badiiy asarning lingvistik mohiyati ham ushbu tahlilda yuzaga chiqadi. Masalan, “O‘tkan kunlar” asarining asl tilida 2-grammlardan foydalanib ko‘rdik. Bunda eng ko‘p qo‘llanilgan so‘z juftliklari, asosan, *“O‘zbekoyim”, “Oftoboyim”, “dedi va”,”shuning uchun”*

kabilar bo‘lsa, rus tilidagi tarjima nusxasida *“об этом”, “с ним”, “что вы”* kabi birliklar ko‘p qo‘llanilgan.

3-jadval. 2-gramm asosida qiyosiy tahlil

	O‘zbek tili PK	**Rus tili PK**
2-gramm natijasi	1037 ta	1367 ta
Takrori	12 690 marta	14 613 marta

Demak, Abdulla Qodiriyning “O‘tkan kunlar” romani yuzasidan yaratilgan o‘zbek-rus parallel korpusida statistik, qiyosiy, lingvistik tahlillarni amalga oshirib, quyidagi xulosalarni qayd etish mumkin bo‘ldi:

1. Abdulla Qodiriyning “O‘tkan kunlar” romani rus tilidagi tarjimasi korpusga joylanganda so‘zlar soni asliyat tildagisidan 4000 taga, gaplar miqdori esa qariyb 1000 taga kamaygan. Bunda tarjima asarda ba’zi gaplar va jumlalar qisqartirilgan.

2. Asar o‘zbek tilidan rus tiliga tarjima qilinganda kishilarning lavozimi, mansab hamda laqabini bildiruvchi leksik birliklar to‘g‘ridan to‘g‘ri o‘zbek tilidan rus tiliga o‘tgan. Bunda asliyat tilga hamda o‘zbek millati tarixiga xos birliklar saqlanib qolgan.

3. Asarning asliyat tilida 2-grammlar tahlili asosida shu aniq bo‘ldiki, kishilarning nomlaridan keyingi o‘rinda keng qo‘llanilgan leksik birliklar qo‘shma olmoshlar, ko‘makchili birikmalardir.

4. Shaxs nomlari tarjima asarda o‘zbek tilidagisidan 3 barobar ko‘p qo‘llanilgan (masalan, “Атабек” so‘zi 650 marta). Bu esa tarjimonga xos xususiyatlardan bir ekanligi namoyon bo‘lgan.

5. O‘zbek-rus parallel korpusida parallel konkordans, tezaurus funksiyalari shuni ko‘rsatdiki, rus tilidagi matnlar tahlilida korpus imkoniyatlari kengayadi. Buni o‘zbek tilida korpus lingvistikasi hamda parallel korpus yangi soha ekani, hali taraqqiy etmagani hamda “Sketch Engine” dasturida ko‘p tilli korpus yaratish imkoniyatlarining samaradorligini oshirish zarurligi bilan izohlaymiz.

Foydalanilgan adabiyotlar:

1. Abjalova M. Parallel korpuslarni yaratish asoslari / Kompyuter lingvistikasining zamonaviy texnologiyalari-CTCL 2023. –Tashkent, 2023. 35-42-b.

2. Akramdjanova M. Parallel matnlar korpusini yaratishning amaliy va nazariy ahamiyati / Ilm-fan va innovatsiya ilmiy-amaliy konferensiyasi (in-academy.uz/index.php/si). –Tashkent, 2023.

3. Mengliyev B., Hamroyeva Sh. Korpus lingvistikasi: Korpus tuzish va undan foydalanish. -Tashkent, 2020.

4. Sobirova N., Akramdjanova M., Sobirov Sh. Parallel korpus tuzish xususiyatlari, tuzilishi, interfeysi / Academic research in Educational Sciences, 2022. 34-b.

5. Abrurakhmonova N., Urdishev K. Corpus based teaching Uzbek as a foreign language. Journal of Foreign Language Teaching and Applied Linguistics (J-FLTAL), 2019.

6. https://app.sketchengine.eu/#dashboard?corpname=user%2Fzukhramarifbaeva%2Fparallel_korpus_russian

MUNDARIJA

SO‘Z BOSHI

Sirojiddinov Sh.S. Tabrik so‘z ……..…… 4

Jo‘raqo‘ziyev N.I. Kompyuter lingvistikasi tadriji zarurati …… 5

Normurodova N.Z. O‘zbek tilini raqamlashtirish yo‘lida …… 6

YALPI YIG‘ILISH MA’RUZALARI

Eşref Adalı Doğal dil işlemedeki gelişmeler…… 7

Normamatov S.E O‘zbek tilining elektron resurslarini yaratish – zamon talabi… 22

Mengliyev B.R. O‘zbek tilining fonematik-fonemik sistemasi xaritasini yaratishning nazariy asoslari…….…….. 23

Toirova G.I. Internet asosida til o‘rgatishda kompyuter texnologiyalari orqali pragmatik kompetensiyani rivojlantirish……..... 30

I SHO‘BA. O‘ZBEK TILINING FORMAL GRAMMATIKASI YARATISH

Elov B.B., Bekmurzayeva M.B. Ot so‘z turkumiga oid birliklarning leksik-grammatik xususiyati…….. 38

Hamroyeva Sh.M., Maxmudjonova G.U. Kam resursli tillarda G2P model tayyorlashda fonetik korpuslarning ahamiyati: o‘zbek tili misolida…….. 44

Abdullayeva O.X., Nizomova F.B. O‘zbek tilida moslashuv yo‘li bilan birikkan so‘z birikmalarini modellashtirish masalasi…… 50

Maxmudjonova G.U. O‘zbek tili fonetikasini TTS tizimlari uchun formallashtirish……... 56

Soliyeva N.Q. O‘zbek tilining bazis so‘zlariga asoslangan kontekst-erkli grammatika ishlab chiqish va uning sintezi …… 63

Sayfiddinova Z.A. Gapdagi boshqaruv, moslashuv munosabatlarining xususiyatlari (kelishiklar orqali ifodalanadigan sintaktik munosabatlar)…….. 69

II SHO‘BA. KORPUS LINGVISTIKASI

Elov B.B., Aloyev N.R. Korpus matnlarini python tili vositasida tematik modellashtirish…….. 81

Миров М.О., Турсынова М.А. Корпустық лингвистиканың дамуындағы профессор асқар жұбановтың рөлі……. 87

Sharipov M.S., Kurbanova L.U. Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga o‘girish mashinali tarjima modelini qurish uchun parallel korpus yaratish 96

Hamroyeva Sh.M., Alimbekova M.X. “Abdurauf Fitrat korpusi” leksikasining statistik tahlili……. 103

Yusupov F., Sharipov M.S., Urazbayeva N.O. Qoraqalpoq tili so‘zlarini lemmalash uchun lingvistik manba yaratish…….. 111
Madatov X.A., Xajibaeva S.M., Xujamov E.J. “O‘zbek boshlang‘ich sinf maktab korpusi”dan bazis so‘zlarni avtomatik ajratib olish…… 116
Adizova N.I. O‘zbek tili korpusi uchun turizmga oid matnlarning lingvistik bazasini yaratish asoslari ……. 123
Abdullayeva O.X. O‘zbek tili korpusi uchun teglash tizimini ishlab chiqish masalasi ……. 130
Matlatipov S.G., Matyoqubova Sh.M., Komilova M.B. INCEpTION platformasi yordamida o‘zbek tili uchun iyerarxik bog‘lik daraxti korpusi yaratish…….. 138
Matlatipov S.G., Komilova M.B. O‘zbek tilida bilidirlgan fikrlarni aspektli sentiment tahlil qilish……. 148
Xusainova Z.Y., Yangiboyeva S.G. Diaxron korpus yaratish bosqichlari (mustaqillik davri nashrlari misolida) …… 162
Elov B.B., Primova M.H., Amirkulov M.A. O‘zbek tili korpusi: tarixiy rivojlanishi, maqsad va vazifalari…….. 168
Matyakubova N.Sh. Parallel korpuslarda sentiment tahlil qilishda “Aligner” vositasining o‘rni…… 185
Mansurova Sh.N. Sifat so‘z turkumini grammatik pos teglash……. 192
Hamroyeva Sh.M., Sulaymonova M.Sh. Diskursni aniqlashda korpusdan foydalanish……. 199
Primova M.H. O‘zbek tili milliy korpusi tuzilishi: asosiy komponentlar va axborot turlari…… 207
Elov B.B., Ahmedova M.K. Til korpusi matnlarida imlo tuzatish…….. 222
Akramova Sh.I. Nusratullo Jumaxo‘ja mualliflik korpusi lingvistik ta’minoti… 229
Urazaliyeva M.Y. Audiomatnlarda morfologik tahlil bosqichi……. 235
Asqarova M.B. Tarjima nazariyasida parallel matnlarning ahamiyati……. 246
Hamroyeva Sh.M., Xolmonova I.A. Mashina tarjimasining sifatini oshirishda parallel korpusning ahamiyati……. 254
Sulaymonova M.Sh. Exploring the stylistic richness of figurative expressions in the uzbek languages…… 263
Maxammadiyeva D.R. Abdulla oripov, muhammad yusuf, shavkat rahmon she’riyati korpusini yaratishning ahamiyati……... 269
Samatboyeva M.T. O‘zbek tili korpusini annotatsiyalash va teglash (umumiy qoida va tizimlar) …….. 275
Xolmonova I.A. O‘zbek-turk parallel korpusi uchun matnlarni tanlash mezonlari 282
Hamroyeva Sh.M., Baxodirov S.R., Muftillayeva R.T., Dialektal so‘zlar va nutq yozuvlarini saqlashning dasturiy yechimi…….. 289

Musayeva M.M. Shevalar va media: o‘zbek tilining raqamli rivojida xorazm shevasining o‘rni…… 298
Muftillayeva R.T. Nurota shevasi (Qarluq lahjasi) korpusining ma’lumotlar bazasi…… 302
Muradova N.B. Zahiriddin muhammad boburning korpusining interfeysi…….. 307
Baxodirov S.R., Muradova N.B. Mualliflik korpusi qidiruv tizimini ishlab chiqish…… 314
Yangibayeva S.G. Diyakron derlemin filolojik araştirmalardaki önemi…….. 322
Xudayberganov N.U., Mansurova Sh.N. Sifat so‘z turkumini grammatik POS teglash…… 327
Bekmurzayeva M.B. Abdulla qahhorning “bemor” hikoyasi parallel korpusi…. 334
Ibragimova M.E. Tohir Malik asarlarida salbiy so‘zlarning statistik tahlili…… 347
Aytbayeva A.X. Abdulla qodiriy asarlarida qo‘llangan tarixiy so‘zlarning korpus tahlili……... 355
Xushmuratova M.B. Korpus vositasida hissiy tahlil (Oybekning “Bolalik” qissasi misolida) …… 364
Hamroyeva Sh.M., Xudoynazarova G. Badiiy asarlar korpusida omonim shakllar tahlili…… 371
Xusainova Z.Y., Raxmatxonova N.I. O‘zbek va Fors tillari uchun parallel korpus: lingvistik va texnologik yondashuvlar…….. 378
Xusainova Z.Y., Bekchanova M. Xorazm shevasidagi matnlarni o‘zbek adabiy tiliga o‘girish dasturini yaratishning amaliy ahamiyati…….. 385

III SHO‘BA. TABIIY TILNI QAYTA ISHLASH (NLP)

Mamatov N.S., **Nuritdinov N.D.** Matnni tahlil qilish bo‘yicha ilg‘or tajribalar 392
Elov B.B., Abdullayeva O.X. Sintaktik parsing turlari…… 403
Sharipov M.S., Adinayev X.S., Yusupova M.M. O‘zbek tili matnlarida undov gaplarni tahlil qilishning qoidaga asoslangan algoritmi…… 413
Sharipov M.S., Adinayev X.S., Yusupova M.M., Qadamboyeva D.D. O‘zbek tili matnlarida savollarni aniqlashning avtomatlashtirilgan tizimi uchun mashinani o‘rganish yondashuvi…… 420
Sharipov M.S., Kurbonova R.U. Markov Model for text generation in Uzbek: practical application and results…… 427
Hamroyeva Sh.M., Qodirova M.M. Yordamchi so‘zlarning qo‘llanish davriga ko‘ra saralovchi lingvistik ta’minot vositasini tuzish masalalari 435
Elov B.B., Samatboyeva M.T. NER: Teglash muommolariga yechimlar…… 442
Abdisalomova Sh.A. O‘zbek tilidagi matnlar koreferensiyasini avtomatik aniqlash tizimining lingvistik ta’minotini shakllantirish xususida……. 449
Elov B.B., Bekmuradova I.Z. Avtomatlashtirilgan til tahlili: slengni aniqlash va teglash muammolari…… 458

Abdullayeva O.X. Bozorqulova O‘. E. O‘zbek tilidagi sodda gaplarning ifoda maqsadiga ko‘ra turlarini modellashtirish masalasi..... 463
Xudayberganov N.U., Karimova Z.D. Ravish so‘z turkumini grammatik pos teglash algoritmlari: nazariy yondashuv, amaliy tatbiq va kelgusidagi yo‘nalishlar……. 468
Xusainova Z.Y., Shirinboyeva M. Matnni tasniflash va ularning sohalar kesimida qo‘llanilish masalasi…… 475
Bekmuradova I.Z. Slenglarni avtomatik teglashda lug‘atga va mashinali o‘qitishga asoslangan yondashuvlar……... 482
Erkinjonova G.B. Zamon qo‘shimchalarini avtomatik aniqlash modeli……... 488
Bozorqulova O‘.E. Turkiy tillardagi treebanklar klassifikatsiyasi…… 493
Karimova Z.D. O‘zbek tilidagi so‘z birliklarini olmosh so‘z turkumi bo‘yicha POS teglashdagi muammolar va yechimlar…….. 500

IV SHO‘BA. MASHINA TARJIMASI

Hamroyeva Sh.M., Erkinjonova G.B. Grammatik teglashda mashinali o‘qitish yondashuvi……. 506
Shamsiyeva G.A. Parallel korpuslarni tarjima xotirasiga integratsiya qilish tamoyillari…….. 511

V SHO‘BA. MATNGA QAYTA ISHLOV BERISH TIZIMLARI

Elov B.B., Sobirova N.G‘. Jahon tajribasida avtomatik matn tahrirlash vositalari 518
Xayotova Z.M. O‘zbek tili matnlari uslublararo perefraz modeli……. 527
Xayotova Z.M. Perifraz generatsiya qilish algoritmlari tahlili…….. 532
Elov B.B., Sobirova N.G‘. Tinish belgilarining matndagi ahamiyati va vazifalari 540
Sobirova N.G‘. Til modellari va chuqur o‘rganish(deep learning) usullari: umumiy tavsif……. 548
Elov B.B., Ibragimova M.E. Matnni perifraz qilish……. 557
Abdullayeva O.X., O‘tkirova F.B. Korpus lingvistikasida parserlarning umumiy tavsifi…… 566
Ahmedova M.K. Avtomatik imlo tuzatish bo‘yicha tadqiqotlar tahlili: tavsif va muammolari…… 573
Matyakubova N.Sh., Avloqulova I. O‘zbek tili matnlarida tinish belgilarni avtomatik qo‘yish…….. 579

VI SHO‘BA. KOMPYUTER LINGVODIDAKTIKASI

Миров М.О., Жұмамұратова А.С. Қажығали мұханбетқалиұлы шығармаларының көркемдік ерекшеліктері……. 589
Mirov M., Zhumagul D. Neurolinguistics in the philological context: research approaches and educational practices…… 595
Бисенгали І., Кенебаева Ұ., Жақанқызы М. Махмұд Қашқари «Диуани лұғат ат-түрік» тілінің морфологиялық және лексикалық сипаты…….. 599

Kessikbayeva A.K., Yessenova K.O. The main differences between a story and a feature…….. 605
Дәулетші Д.Р. Электронды сөздіктер: қазақ лексикографиясының жаңа мүмкіндіктері…… 613
Айтбенбетова А.К., Молдабаева З.М., Имангазина А.А., Султанбаева Н.К. Smart білім беру технологияларының негізгі қағидалары…….. 620
Askarova U. Computer linguistics and its importance today……. 624
Fozilov P.Sh. Fiziologik holat fe’llarining inglizcha-o‘zbekcha tarjima lug‘atlardagi talqini……. 628

VII SHO‘BA KOMPYUTER LEKSIKOGRAFIYASI

Eshmurodova M.D., Xudoykulova N.R. O‘zbek tili dialektlari izohli lug‘atini yaratish model va algoritmi……. 633

VIII SHO‘BA. SEMANTIK TIZIMLAR VA LINGVISTIK ONTOLOGIYALAR

Elov B.B., Xusainova Z.Y. Sentiment tahlilida so‘z turkumlarining roli…….. 645
Elov B.B., Xusainova Z.Y., Abdullayev A.Q. O‘zbek tilidagi matnlarni so‘z turkumlari orqali sentiment tahlil qilishda gap modellarini ishlab chiqish…… 653
Elov B.B., Abdullayev A.Q. Sentiment tahlil uchun katta hajmdagi datasetni yaratish bosqichlari……. 663
Elov B.B., Suyunova M.O. Lexicon-based va machine learning-based modellari asosida sentiment tahlilni amalga oshirish……. 670
Allanazarova S.Y. Sentiment tahlili uchun o‘rgatilgan resurlari kam tillarda ma’lumotlar omborini tuzish…….. 682
Suyunova M.O. Sentiment tahlili va uning jahon tilshunosligida o‘rganilishi…. 691
Axmedova X.I., Omarova D.T. O‘zbek tilidagi frazeologik birliklarning sinonimlik hodisasi……. 700
Xudoykulova N.R., Eshmurodova M.D. Gapning semantik strukturasining ontologik modelini yaratish…… 704
Omarova D.T. Sinonim iboralarni aniqlash masalasi tadqiqi…… 714
Axmedova X.I., Xushmuratova M.B. KNN-algoritmi yordamida so‘z ma’nosini aniqlash (sifat so‘z turkumi misolida) ……. 719
Axmedova X.I., Muxammadyakubova M.E. Wordnet asosida so‘z ma’nosini aniqlash…….. 726
Abdullayeva O.X., Nizomova F.B. O‘zbek tili matnlarida so‘z birikmasini avtomatik aniqlashda omonimiya masalasi…… 733
Xusainova Z.Y., Xusanova D.T., Bekqulova S.F. Sentiment tahlilida so‘z turkumlarining ahamiyati …….. 741
Axmedova X.I., Hasanova G.I. Tabiiy tilga ishlov berish sohasida so‘z ma’nosini aniqlash……. 751

Axmedova X.I., Hayotova G.H. Qaror ro‘yxati algoritmi yordamida so‘z ma’nosini aniqlash……. 759
Axmedova X.I., Ko‘charova I.Sh. Leksik darajadagi perefraz taksonomiyasini amalga oshirish bosqichlari…… 767

IX SHO‘BA. KOMPYUTER LINGVISTIKASIDA TERMINOLOGIYA VA TERMINOGRAFIYA

Odilova K.T. Kognitiv lingvistika nuqtai nazaridan polikodli matnlarning semantik va sintaktik strukturalari……. 775

X SHO‘BA. KOMPYUTER LINGVISTIKASIDA SUN’IY INTELLEKT

Ochilov M.M., Narzullaev O.O. Ko‘p tilli avtomatik nutqni tanish tizimlarini yaratish muammolar va yechimlari……. 782
Alayev R.H., Berdiyev J.B. BERT modeli yordamida matnlarni tasniflash……. 791
Aloyev N.R. BERTOPIC: neyron tematik modellashtirish usuli……... 798
Jurayev D.B. Tabiiy tilni qayta ishlash yordamida nutqni imo-ishora tiliga tarjima qilish texnologiyalari……. 805
Ruziyeva M.I. Raqamli tilshunoslik: Blingvizm va sun’iy intellekt o‘rtasidagi aloqa…… 812
Berdiyev J.B. Sun’iy intellekt kutubxonalari yoqdamida matnlarni qayta ishlash 818
O‘tkirova F.B. Lingvistik tadqiqotlarda yangi yondashuv: Abdulla Qodiriyning “O‘tkan kunlar” romani asosida yaratilgan parallel korpusi va uning tavsifi..... 827