ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 4 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 4 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Bosh muharrir: **Botir Elov**

Bosh muharrir o‘rinbosari: **Shahlo Hamroyeva**

Mas’ul kotib: **Oqila Abdullayeva**

Tahrir kengashi

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

Alisher Navo'i Tashkent State University of the Uzbek Language and Literature

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 6-18

TIL KORPUSLARIDA MATNLARNI ANNOTATSIYALASH: AFZALLIK VA KAMCHILIKLARI

Mastura Primova[1]

Annotatsiya. Ushbu maqolada korpus annotatsiyasi, annotatsiya turlari, afzalliklari, kamchiliklari ko‘rib chiqiladi. Korpus ma’lum maqsadda yig‘ilgan matnlar majmuyini tashkil etuvchi til birliklari yig‘indisi, tabiiy tildagi elektron shaklda saqlanadigan yozma va og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan online yoki offline tizimda ishlaydigan matnlar jamlanmasi. Tilshunoslikka oid tadqiqotda fakt bilan ish ko‘riladigan hollarda material yig‘ilishi, sistemaga solinishi lozim. Bunday katta hajmli ishni bajarishda korpus vaqt va mehnatni tejaydigan ish quroli vazifasini bajaradi. U texnik jarayonni tezlashtiruvchi vosita bo‘libgina qolmay, muayyan tilning zamonaviy shakliga xos axborot tizimi ham bo‘lib, kutilmagan savolga javob bera oladigan, til hodisasi bilan shug‘ullanadigan soha oldiga avval ko‘rilmagan dolzarb muammolarni qo‘ya oladigan tizim.

Kalit so‘zlar: *Post-tegging, lemmatizatsiya, sintaktik tahlil, annotatsiya, coreferens annotatsiya, pragmatik annotatsiya, stilistik annotatsiya.*

Kirish

Zamonaviy axborot texnologiyalari tilning funksional imkoniyatlaridan foydalanish borasida benihoya imkoniyatlar yaratadi. Kompyuter tarjimasi, avtomatik tahrir va tahlil, yozma matnni ovozlashtiruvchi nutq sintezatorlari, og‘zaki nutqni yozma matnga aylantiruvchi nutqni tanish dasturlari, elektron lug‘atlar, lingvistik mobil ilovalar, tezaurus (til xazinasi)lar va til ontologiyasi. Ayniqsa, zamonaviy elektron lug‘atlar tuzish va undan foydalanish madaniyatini shakllantirish til imkoniyatini egallashda samarador ekanligi o‘z isbotini topgan. Xususan, tilning imkoniyatini namoyon qilish va egallash borasida dunyo miqyosida tez sur’atlarda yaratilayotgan til korpuslarining roli beqiyos. Til korpuslari - til bo‘yicha tadqiqot va amaliy topshiriqlar yechimi uchun zarur ish quroli. U oddiy elektron kutubxonadan farqlanadi. Elektron kutubxonaning maqsadi -

[1]*Primova Mastura Hakim qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi o‘qituvchisi.
E-pochta: *primovamastura@navoiy-uni.uz*
ORCID: 0000-0002-0241-4659

xalqning ijtimoiy-siyosiy, ma'naviy, iqtisodiy hayotini aks ettiruvchi badiiy va publitsistik asarlarni nisbatan to'liq qamrab olish. Elektron kutubxona matnlari til nuqtayi nazaridan ishlov berilmaganligi sababli tadqiqotlar uchun noqulaylik tug'diradi. Chunki elektron kutubxona ilmiy tadqiqot materiali bazasini tayyorlash maqsadida tuzilmaydi, balki milliy ma'naviy merosni to'plashni maqsad qilgan bo'ladi. Til korpusi esa elektron kutubxonadan farqli o'laroq, tilni o'rganish va tadqiq qilish uchun foydali va qiziqarli matnlarni to'plashni nazarda tutadi. Ko'p hollarda korpus annotatsiyalari korpusning belgilari bilan bog'liq hisoblanadi. Shuningdek, tilshunoslik tadqiqotlarda korpuslardan foydalanish orqali lingvistik ma'lumotlarni olishda ishlatiladi. Har doim ham korpusdan ma'lumotlarni ajratib olib bo'lmaydi. Bunday hollarda lingvistik tahlilni korpusga kodlash orqali amalga oshiriladi. "Elektron korpusga og'zaki va/yoki yozma til ma'lumotlarining izohlovchi lingvistik ma'lumotlarni qo'shish" jarayoni **korpus annotatsiyasi** deb ataladi [Leech, 1997. 2]. Korpus annotatsiyasi korpusga qo'shimcha qiymat beradi, ya'ni, korpus osonlikcha hal qila oladigan tadqiqot savollari doirasini sezilarli darajada kengaytiradi. Keng ma'noda ta'riflangan korpus annotatsiyasi ham matnli/kontekstual ma'lumotni, ham izohli lingvistik tahlilni kodlashni nazarda tutishi mumkin bo'lsa-da, adabiyotda tez-tez uchraydigan ikki tushunchaning o'zaro bog'liqligi ko'rsatiladi. Bu yerda atama tor ma'noda qo'llanilib, faqat lingvistik kodlashda ishlatiladi. Masalan, nutq qismini (POS) teglash va korpus matnida sintaktik tahlil qilishda.

Asosiy qism

Annotatsiya - Korpus lingvistikasi doirasida qaralganda berilgan matnga bevosita aloqasi bo'lmagan, ammo uning qaysidir qismi haqida lingvistik yoki ekstralingvistik axborot beruvchi umumiy ma'lumot. Annotatsiya o'z ichiga metama'lumot va teglarni qamrab olishi mumkin. Ayrim manbalarda [Leech, Vilson, 1994] korpus annotatsiyasi deganda, til korpusida matnning elektron shakliga kodlash qo'shilgan izohlovchi, lingvistik ma'lumotlarni qo'shish amaliyoti ham tushuniladi. McEnery annotatsiyalashni uch usulda amalga oshirish mumkinligini yozadi: to'la avtomatlashtirilgan, yarim avtomatlashtirilgan va qo'l mehnati yordamida. Ammo hech bir usul mukammal ish bermasligi, mutlaqo xatosiz natija olib bo'lmasligini aytgan.

Razmetka – McEnery razmetka va meta ma'lumotni annotatsiyalash jarayonining bir qismi sifatida baholagan [McEnery, Hardie,

2012. 47]. Boshqa manbalar bilan tanishganimizda esa razmetkaga annotatsiyalash jarayonining o‘zi sifatida baho berilganligiga guvoh bo‘ldik [Myfilology.ru]. Ingliz manbalarida razmetka markup termini bilan yuritiladi [McEnery, Hardie, 2012. 47]. Jahon amaliyotida razmetkalashning standart prinsiplari ishlab chiqilgan. U SGML yoki Standard Generalized Markup Language [https://www.techtarget.com/] deb ataladi. E’tibor qaratish lozim bo‘lgan jihat shundaki, SGMLda standart annotatsiya emas, balki annotatsiyalash jarayonini tashkil etish metodologiyasi aks etadi. SGML asosida ishlab chiqilgan va keng ommalashgan tillarga HTML yoki XMLni misol qilish mumkin.

Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo‘linadi [https://ucrel.lancs.ac.uk/]: semantik teg, sintaktik teg va grammatik teg. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur. Annotatsiya (razmetka) va tegning farqini quyidagi jadvalda aniqroq ko‘rish mumkin [https://knowledge.autodesk.com/]:

1-jadval. Annotatsiya va teg terminining ta’riflanishi

Annotatsiya	Teg
Muayyan komponent yoki segment haqida jadval yoki grafik ko‘rinishidagi ma’lumot	Muayyan komponent yoki segment uchun unikal identifikator
Komponent yoki segment uchun bir nechta annotatsiya bo‘lishi mumkin	Komponent yoki segment uchun yagona nom beriladi va u teg deb ataladi
Bir qancha ta’riflar to‘plamini matn shaklida o‘zida jamlaydi va unda teglar ham aks etadi.	Gap emas, uning bo‘laklari haqida ma’lumot beruvchi muayyan formatdagi data hisoblanadi

Korpus belgisi korpusning tarkibiy qismlari va har bir matnning matn tuzilishi haqida nisbatan obyektiv tekshiriladigan ma’lumotlarni beradi. Shuningdek, korpus annotatsiyasi izohlovchi lingvistik ma’lumotlar bilan bog‘liq. “Annotatsiyani “*tarjimon*” deb atash orqali izohlash hech bo‘lmaganda ma’lum darajada inson ongining matnni tushunish mahsuli ekanligini bildiramiz” [Leech, 1997. 2]. Misol uchun, so‘zning nutq qismi noaniq bo‘lishi mumkin, buni korpus belgisidan ko‘ra korpus annotatsiyasi sifatida aniqlash osonroq hisoblanadi. Boshqa tomondan, ma’ruzachi yoki yozuvchining jinsi odatda obyektiv tekshiriladi va bu izohlash emas, balki belgilash bo‘ladi.

Korpus annotatsiyasi

Korpus belgisi kabi izoh korpusga qiymat qo‘shadi. Leech [1997. 2] korpus annotatsiyasiga shunday ta’rif beradi: “korpusni kelajakdagi tadqiqot va ishlanmalar uchun lingvistik ma’lumot manbayi sifatida boyitib, korpus keltiradigan foydaga hal qiluvchi hissadir”. McEnery [2003. 454-455] korpus annotatsiyasi kamida *to‘rtta afzalliklarga* ega ekanligini ko‘rsatadi.

• **Annotatsiya korpusdan ma’lumotlarni bir necha xil usullar orqali olish** ancha oson hisoblanadi. Leachning fikriga ko‘ra, nutqning bir qismini teglashsiz, xom korpusdan sifat sifatida chapni ajratib olish qiyin, chunki uning turli ma’nolari va qo‘llanishlarini faqat orfografik shakli yoki kontekstidan aniqlash mumkin emas. Masalan, o‘ngning teskari ma’nosiga ega chap orfografik shakli sifatdosh, qo‘shimcha yoki ot bo‘lishi mumkin. U o‘tgan zamon qo‘shimchasi yoki o‘tgan zamon shakli bo‘lishi ham mumkin. Nutqning tegishli qismiga izohlar bilan chapning bu turli xil qo‘llanilishini bir-biridan osongina ajratish mumkin. Shuningdek, korpus annotatsiyasi, inson tahlilchilari va mashinalariga o‘zlari qodir bo‘lmagan tahlillardan foydalanish va olish imkonini beradi [McEnery, 2003. 454].

Misol uchun, Xitoy tilini bilmasangiz ham, agar sizda to‘g‘ri izohlangan Xitoy korpusi bo‘lsa, ushbu korpusdan foydalangan holda Xitoy tili haqida ko‘p narsalarni bilib olishingiz mumkin. Korpusdan ma’lumotlarni olish tezligi - izohli korpusning yana bir afzalligi hisoblanadi. Agar biror kishi kerakli lingvistik tahlilni amalga oshirishga qodir bo‘lsa ham, agar biror kishi korpusning o‘ziga izoh berishdan boshlash kerak bo‘lsa, xom korpusni izohli korpusni o‘rganish kabi tez va ishonchli tarzda tekshira olishi dargumon.

• **Korpus annotatsiyasi qayta foydalanish mumkin bo‘lgan resursdir**, chunki annotatsiya korpus ichidagi lingvistik tahlillarni qayd etadi, keyinchalik ularni qayta ishlatish mumkin hisoblanadi. Korpus annotatsiyasi odatda qimmat va vaqt talab qilishini hisobga olsak, qayta foydalaniladi [Leech, 1997. 5].

• **Korpus annotatsiyasi ko‘p funksiyalilik uchun uni qayta foydalanish** mumkin hisoblanadi. Korpus dastlab ma’lum bir maqsad uchun izohlangan bo‘lishi mumkin. Biroq, korpus tahlili turli ilovalar uchun va hatto dastlab mo‘ljallanmagan maqsadlarda ham qayta ishlatilishi mumkin.

• **Korpus annotatsiyalari lingvistik tahlilni aniq qayd qilib boradi.** Unda tahlil va tanqidga ochiq bo‘lgan aniq obyektiv rekordni ta’minlaydi [McEnery, 2003].

Yana bir afzalliklaridan biri *korpus annotatsiyasi korpusning o‘zi kabi standart ma’lumot manbasini taqdim etadi*. U o‘zi ifodalovchi til xilma-xilligi uchun standart ma’lumotnoga asoslangan. Korpus annotatsiyasi asosan lingvistik tahlilni asosini ta’minlab, obyektiv ravishda qayd etadi. Shuning uchun ketma-ket tadqiqotlar umumiy asosda taqqoslanishi va qarama-qarshi qo‘yilishi mumkin bo‘ladi.

So‘nggi o‘n yil ichida korpus annotatsiyalariga to‘rtta asosiy kamchiliklari tanqid qilingan:

• **Korpus annotatsiyalari korpusda tartibsizlikni keltirib chiqaradi.** Ular Hunstonning fikriga ko‘ra “matnga juda ko‘p izoh qo‘shilgan bo‘lsa-da, tadqiqotchi annotatsiya belgilaridan tozalangan oddiy matnni ko‘ra olishi muhim” (2002. 94) deb ta’kidlaydi

• **Korpus annotatsiyalari korpus foydalanuvchisiga lingvistik tahlilni yuklaydi.** Garchi korpus annotatsiyalari o‘z mohiyatiga ko‘ra izohli bo‘lsa-da, korpus foydalanuvchilari ushbu tahlilni qabul qilishga majbur emas. Agar xohlasalar annotatsiyani e’tiborsiz qoldirib, o‘zlarining talqinlarini yuklashlari mumkin. Korpus annotatsiyasida matnni talqin qilishdan boshlanadi [McEnery, 2003. 456]. Shuningdek, korpusni izohsiz qoldirish korpus tahlil qilinganda talqin qilish jarayoni sodir bo‘lmaydi degani emas. Aksincha, izohning yo‘qligi tadqiqotchilar xom korpusdan foydalanganda bunday ko‘p talqinlar hali ham sodir bo‘lishini yashiradi. Tahlil hali ham sodir bo‘ladi, u shunchaki aniq ko‘rinishdan yashiringan hisoblanadi. Shuningdek, korpus annotatsiyasi bu borada zaiflik emas, balki afzallik sifatida tan olinadi, chunki u tekshirish uchun ochiq bo‘lgan aniq tahlilning obyektiv rekordini ta’minlaydi - izoh bermaslik shunchaki tahlil qilmaslik emas. Biroq, izohning yo‘qligi tahlilni qayta qurish qiyin yoki hatto imkonsiz bo‘lishini ta’minlaydi.

• **Annotatsiya korpusni “*ortiqcha baholab*” qo‘yishi mumkin, bu esa uni kamroq kirish, yangilash va kengaytirish imkonini beradi** [Hunston, 2002. 92-93]. Shuningdek, annotatsiya korpusni kamroq kirishni talab qilmaydi. Misol uchun, ko‘plab tahlil qilingan [masalan, Lancaster Parsed Corpus va Suzanne korpusi] va prosodik ravishda izohlangan korpuslar [masalan, London-Lund Corpus va Lancaster/IBM Spoken English Corpus] hamma uchun ochiqdir. Ba’zi korpus yaratuvchilari odatda o‘z korpuslarini annotatsiya qilish uchun juda ko‘p harakat qilishlariga qaramay o‘z korpuslarini iloji boricha kengroq foydalanishga topshirishdan mamnun bo‘lishadi. Ko‘pincha tashkilotlar korpus qurulishini moliyaviylashtirishadi chunki, qimmatli annotatsiyalar resursi ommaga taqdim etiladi. Annotatsiyalangan korpusni (yoki hatto xom korpusni) ommaga

taqdim etmaslikning keng tarqalgan sababi, korpus ma'lumotlariga tegishli mualliflik huquqi bilan bog'liq muammolar uni taqiqlaydi. Bu cheklov mualliflik huquqiga qo'yiladi, annotatsiyalarga emas.

Namuna korpusi ma'lum bir vaqtda ma'lum bir til xilma-xilligini ifodalash uchun mo'ljallangan hisoblanadi. Masalan, LOB va Brown korpuslari 1960-yillarning boshidan yozma Britaniya va Amerika ingliz tilini ifodalaydi deb taxmin qilinadi. Bu ikkita korpus - FLOB va Frown uchun "*yangilanishlar*" mavjud. Unda 1990-yillarning boshidan yozma Britaniya va Amerika ingliz tillarini ifodalaydi va sekinroq til o'zgarishlarini kuzatish uchun ishlatiladi. Doimiy kengayish zarurati faqat dinamik monitor korpus modeli bilan bog'liq hisoblanadi. Bu namuna korpusiga argument sifatida qo'llanilishi shart emas. Aksariyat korpuslar namunaviy korpus ekanligini hisobga olsak, kengaytirilish argumenti unchalik muhim emas, chunki namunaviy korpus hajmi odatda korpus ishlab chiqilganda aniqlanadi. Odatda, korpus yaratilgandan so'ng, kengaytirishga hojat bo'lmaydi.

• **Oxirgi tanqidda korpus annotatsiyasining aniqligi va izchilligi bilan bog'liq.** Korpusga annotatsiya berishning uchta asosiy usuli mavjud - *avtomatik*, *kompyuter* va *qo'lda*. Hunston fikriga ko'ra, "avtomatik izohlash dasturi inson tadqiqotchisi oladigan natijalarga 100% mos keladigan natijalarni berishi dargumon; boshqacha qilib aytganda, xatolar ehtimoli bor". Bunday xatolar matnlar faqat odamlar tomonidan matnlarni tahlil qilinganda ham sodir bo'ladi - hatto eng yaxshi tilshunos ham ba'zida xato qiladi. Shuning uchun annotatsiyalarga inson omillarini kiritish boshqa natijalarga olib kelishi mumkin; Sinkler (1992) ta'kidlashicha, "qo'lda yoki kompyuter yordamida korpus annotatsiyasiga inson omillarini kiritish annotatsiya izchilligini pasayishiga olib keladi". Bu ikki fikrni bir joyga jamlagan holda, nima uchun har qanday tilshunos tahlil qiladi, degan savol tug'ilishi mumkin. Chunki tahlillardagi nomuvofiqlik va noaniqlik haqiqatan ham kuzatilishi mumkin bo'lgan hodisalar bo'lsa-da, ularning ekspert inson tahliliga ta'siri bo'rttirilgan. Bundan tashqari, kompyuter noaniqliklar yoki nomuvofiqliklarning oldini olishning ishonchli vositasi emas: bu ikki nuqta mashina tahliliga ham tegishli bo'lishi mumkin. Avtomatlashtirilgan annotatsiyalar xatolari bo'ladi va bir-biriga mos kelmaydi. Agar annotatsiya dasturi uchun resurslar o'zgartirilsa - leksika o'zgartirilsa, qoidalar qayta yozilsa, vaqt o'tishi bilan dastur natijasi inson tahlilchilari tomonidan ko'rsatilganidan ancha yuqori darajada oshib ketishi mumkin bo'lgan shkalada nomuvofiqlikni ko'rsatadi. Korpus annotatsiyasida nimadan foydalanishimiz kerak: inson tahlilchilarimi yoki kompyu-

termi? Korpus annotatsiyasining ahamiyati keng e'tirof etilganligini hisobga olsak, inson tahlilchisi va mashina bir-birini to'ldirishi kerak, bu esa korpus hal qilish uchun mo'ljallangan tadqiqot savoli uchun noaniqlik va nomuvofiqlikni kamaytirishga qaratilgan aniqlik va izchillikka muvozanatli yondashuvni ta'minlashi kerak.

Yuqoridagi to'rtta tanqidlarni korpus annotatsiyasi rad etish mumkin. Unda annotatsiya faqat lingvistik tahlilni amalga oshirish va uni amalga oshirishni anglatadi.

Korpus annotatsiyasiga qanday erishiladi?

Korpus annotatsiyasi to'liq avtomatik ravishda ishlatiladi: ya'ni yarim avtomatlashtirilgan inson va mashina o'zaro ta'siri orqali yoki inson tahlilchilari tomondan butunlay qo'lda amalga oshiriladi. Uchalasini o'z navbatida qamrab olish uchun avtomatik izohlashda kompyuter dasturchi tomonidan oldindan belgilangan qoidalar va algoritmlarga rioya qilgan holda annotator sifatida yakka o'zi ishlaydi, ammo qoidalar oldindan belgilangan ML algoritmi yordamida mashinani o'rganish (ML) orqali ham mashina tomonidan tanlashi mumkin. Biroq, avtomatik izohlash vositasini ishlab chiqish vaqt va pul talab qilib, ma'lumotlar bazasida katta hajmdagi ma'lumotlarga tez va doimiy ravishda izohlanadi (resurs o'zgarmagan holda). Ba'zan, bu ish allaqachon boshqa joyda amalga oshirilganligini va kerakli izohni bajara oladigan dastur tekin mavjud ekanligini ko'rish mumkin.

Annotatsiyalarning ayrim turlari, masalan, Ingliz, fransuz va ispan tillari uchun lemmatizatsiya va POS belgilarini belgilash, xitoy tili uchun segmentatsiyasi va POS belgilarini ishonchli tarzda mashina yordamida amalga oshirilishi (odatda xatolik darajasi 3%), annotatsiyasiga to'liq avtomatlashtirilgan yondashuvdan iborat. Avtomatlashtirilgan jarayondan olingan ma'lumotlar tahlil qilinganligi kabi bo'lmasa yoki chiqish ishonchli bo'lsada, lekin ma'lum maqsad uchun yetarlicha aniq bo'lmasa (masalan, izohni yaxshilash uchun foydalaniladigan o'quv korpusi), odatda insoniy tuzatish talab qilinadi (ya'ni post-tahrirlash). **Post-tahrirlash** odatda qo'lda izoh berishdan ko'ra tezroqdir. Ba'zi izohlash vositalari inson-mashina interfeysini ta'minlaydi, bu inson tahlilchisiga mashina aniq bo'lmagan noaniq holatlarni hal qilishga yordam beradi. Yarim avtomatik izohlash jarayoni to'liq avtomatlashtirilgan izohga qaraganda ishonchliroq natijalar beradi, lekin u sekinroq ishlaydi va qimmat hisoblanadi. Sof qo'lda izohlash foydalanuvchi uchun hech qanday izoh vositasi mavjud bo'lmaganda yoki mavjud tizimlarning aniqligi qo'lda tuzatishga sarflangan vaqtni sof qo'lda izohlashdan kamroq qilish uchun yetarli

bo‘lmagan hollarda yuzaga keladi. Qo‘lda izohlash qimmat va ko‘p vaqt talab qiladiganligi sababli, odatda faqat kichik korpuslar uchun amal qiladi. Yuqorida aytib o‘tilgan bir nechta istisnolardan tashqari, hozirda katta korpuslarda mavjud bo‘lgan izohlarning aksariyat turlari yarim avtomatik yoki qo‘lda kiritilgan.

Korpus annotatsiyalarining turlari

Korpus annotatsiyasi turli darajalarda paydo bo‘lishi va turli shakllarga ega bo‘lishi mumkin. Masalan, fonologik darajada korpusga bo‘g‘in chegaralari (phonetic/phonemic annotation) yoki prozodik xususiyatlar (prosodic annotation) bilan izohlash mumkin; morfologik darajada korpusga prefikslar, qo‘shimchalar va o‘zaklar bilan izohlash mumkin morphological annotation); leksik darajada korpuslar nutq qismlari (POS tagging), lemmalar (lemmatizatsiya) (lemmatization), semantik maydonlar (semantic annotation) bilan izohlanishi mumkin; sintaktik darajada, korpuslarni tahlil qilish (parsing, treebanking or bracketing) yordamida izohlash mumkin; nutq darajasida korpusga anaforik munosabatlar (coreference annotation), nutq aktlari kabi pragmatik ma’lumotlar (pragmatic annotation) yoki nutq va fikrlarni ifodalash (stylistic annotation) kabi stilistik xususiyatlarni ko‘rsatish uchun izohlash mumkin.

Ulardan eng keng tarqalgan annotatsiya turi **POS tegi** bo‘lib, u ko‘plab tillarda muvaffaqiyatli qo‘llanilgan; **sintaktik tahlil** qilish ham jadal rivojlanmoqda, ayni paytda annotatsiyaning ayrim turlari (masalan, nutq va pragmatik annotatsiya) hozircha nisbatan rivojlanmagan.

POS teglari

POS tegi (grammatik teg yoki morfo-sintaktik izoh deb ham ataladi) korpusdagi har bir so‘zga POS tegi sifatida ham tanilgan nutqning bir qismi mnemonikasini belgilashni anglatadi. POS teglari korpus annotatsiyalarining birinchi keng tarqalgan turlaridan biri bo‘lib, bugungi kunda eng keng tarqalgan turi hisoblanadi. Shuningdek, tahlil va semantik izoh kabi keyingi tahlil shakllarining asosini tashkil etuvchi korpus annotatsiyasining eng asosiy turi hisoblanadi. Biroq, faqat nutq qismlari uchun izohlangan korpuslar keng ko‘lamli ilovalar uchun foydali bo‘lib, omograflarni ajratib ko‘rsatishdan tortib, korpusdagi so‘z sinflarining paydo bo‘lishini hisoblash kabi murakkabroq foydalanishgacha. Ko‘pgina lingvistik tahlillar, masalan, so‘z birikmasi ham POS teglariga tayanadi [Hunston, 2002. 81].

POS yorlig‘i rivojlanishining ilg‘or holatini hisobga olgan holda, u ko‘plab tillar uchun ko‘pgina tadqiqot savollari uchun yetarli aniqlik bilan avtomatik ravishda amalga oshirilishi mumkin.

Lug‘aviy elementlarga avtomatik ravishda POS teglarini tayinlaydigan izohlash vositasi tegger deb ataladi. Ingliz tili uchun eng mashhur va ishonchli teggerlardan biri Lankaster universitetida ishlab chiqilgan CLAWS (Clausibility Based Automatic Word Tagging System) hisoblanadi [Garside, Leech and Sampson, 1987]. Tizim gibrid statistik yondashuvdan foydalanadi, qoidaga asoslangan komponent bilan to‘ldiriladi, g‘ayrioddiy tarzda “idiomlar ro‘yxati” deb ataladi [Garside va Smith, 1997]. Bu tegger umumiy yozma ingliz tilida 97% aniqlikka erishgani xabar qilingan [Garside va Smith, 1997]. Tizim Britaniya Milliy Korpusini belgilash uchun ishlatilgan (BNC, 7.2-bo‘limga qarang). POS teggerlari fransuz [Gendner, 2002], ispan [Farwell, Helmreich, Kasper, 1995], nemis [Hinrichs, Kübler, Myuller va Uhle, 2002], shved [Kutting, 1994], Xitoy [Chang, Chen] kabi tillar uchun ham muvaffaqiyatli ishlab chiqilgan.

Xulosa

Tilni o‘rgatishda lug‘at boyligining ulkanligini ko‘rsata olish, so‘zning qo‘llanish imkoniyatini u yoki bu grammatik qurilma orqali tushuntirish uchun misollar massivini ko‘rsatishda korpus juda qo‘l keladi. Til birligini qidirish kerak bo‘lsa, bunday dasturiy ta’minot, ya’ni korpus tadqiqotchi yoki foydalanuvchiga juda katta yordam beradi. Ilgari tadqiqotchi o‘z ishi uchun misolni topish, ularni kartotekaga ko‘chirish (kompyuter texnologiyalari rivojlanishidan oldingi davrda)ga oylab, ba’zan yillab vaqt sarflagan bo‘lsa, bugun dunyo til korpuslari yordamida sanoqli daqiqada yuzlab misol topishga, tahliliy natijalarni olishga va ular ustida ishlash imkoniga ega bo‘ldi. Maxsus qidiruv tizimi korpusdan ma’lumot olishga mo‘ljallangan bir qancha dasturdan iborat, u statistik axborot va qidiruv natijasini foydalanuvchiga qulay shaklda taqdim eta oladi. Tilda qanday jarayon kechayotganligini aniq tasavvur qilish uchun korpus qamrovini yanada kengaytirish, nafaqat yozma nutq, balki og‘zaki nutq materialidan ham foydalanish maqsadga muvofiq. Bunday korpus yordamida taraqqiyot natijasida tilda sodir bo‘lgan va kutilayotgan o‘zgarish haqida aniq xulosa chiqarish mumkin. Tilshunoslikka oid tadqiqotda fakt bilan ish ko‘riladigan hollarda material yig‘ilishi, sistemaga solinishi lozim. Bunday katta hajmli ishni bajarishda korpus vaqt va mehnatni tejaydigan ish quroli vazifasini bajaradi. U texnik jarayonni tezlashtiruvchi vosita bo‘libgina qolmay, muayyan tilning zamonaviy shakliga xos axborot tizimi ham bo‘lib, kutilmagan savolga javob bera oladigan, til hodisasi bilan shug‘ullanadigan soha oldiga avval ko‘rilmagan dolzarb muammolarni qo‘ya oladigan tizim.

Korpuslar turli maqsadlarda turli sohalar uchun tuzilishi mumkin. Jumladan, og‘zaki yoki yozma matnlar, bir tilli yoki ko‘p tilli, uslubiga ko‘ra badiiy, publitsistik, folklor, bazasi o‘zgaruvchan yoki o‘zgarmas, matn hajmiga ko‘ra to‘liq matnli yoki fragmentli korpuslar bo‘ladi.

Til korpusi qurilishida lingvistik ta’minot masalasi muhim va murakkab hisoblanadi. Korpuslarda matnlardagi nutq bo‘laklariga mos identifakatorini belgilash jarayoni muammolidir, sababi tilni modellashtirish teglash qoidasi va tilda mavjud qonuniyat bilan bog‘liq. Teglash, xususan, grammatik teglash yoki POS tegging o‘zbek korpus lingvistikasi uchun ham dolzarb masaladir. Chunki maxsus “kodlangan” belgilar tizimi o‘zbek tili bilan bog‘liq NLP masalalarini yechishda birlamchi kalit bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar

Elov B., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlari uchun tf-idf statistik ko‘rsatkichni hisoblash. Science and innovation international scientific journal volume 1 issue 8 uif-2022: 8.2 ISSN: 2181-3337. 1774-1785 b.

O‘zbek tili ta’limiy korpusi - http://uzschoolcorpara.uz/

Elov B., Amirkulov M. Uzbek-English Parallel Corpus Algorithm and Alignment Problem. Central asian journal of literature, philosophy and culture eissn: 2660-6828 | Volume: 04 Issue: 06 June 2023. 71-78 p.

Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and Artificial Intelligence, Vol. 1 No. 3 (2023) 32–42. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v1i317. 117–129 b.

Elov B., Alayev R. O‘zbek tili korpusi va uning imkoniyatlari. O‘zbekiston Informatika va energetika mummolari jurnali. O‘zbekiston Jurnali. - Toshkent, 2023, - № 2. 89-100 b.

Elov B., Hamroyeva Sh., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlarini qayta ishlashda countvectorizer, tf-idf hamda co-occurrence matrix usullarining ahamiyati. ELEKTRON LUG‘ATLAR YARATISHNING NAZARIY VA AMALIY ASOSLARI mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari, 2023-yil 12-dekabr. Andijon. 78-88 b.

Захаров В.П. Корпусная лингвистика: Учебно-метод. Пособие. –СПб., 2005, -48с.

Boliang Zhang, Ajay Nagesh, Kevin Knight. 2020. Parallel Corpus Filtering via Pretrained Language Models. 10.18653/v1/2020.acl-main.756

Elaheh Rafatbakhsh, Alireza Ahmadi. 2019. A thematic corpus-based study of idioms in the Corpus of Contemporary American English. Asian-Pacific Journal of Second and Foreign Language Education,10.1186/s40862-0190076-4

Полицын, С.А., Полицына Е.В. 2018. Применение корпуса текстов для автоматической классификации в комплексе инструментов автоматизированного анализа текстов. Вестник ВГУ. Серия: Системный анализ и информационные технологии,10.17308/sait.2018.2/1224.

ADVANTAGES AND DISADVANTAGES OF CORPUS ANNOTATION

Mastura Primova[1]

Abstract. This article discusses corpus annotations, types of annotations, advantages and disadvantages. A corpus is a collection of linguistic units constituting a collection of texts collected for a specific purpose, written and spoken in natural language, stored in electronic form, a collection of texts hosted in a computerized software-based search engine and operated in an online or offline system. When working with facts in linguistic research, material should be collected and systematized. When performing such large-scale work, the case acts as a working tool, saving time and labor. It is not only a tool for speeding up a technical process, but also an information system specific to the modern form of a particular language, a system that can answer unexpected questions and create unprecedented problems in the field of linguistic phenomena.

Key words: *Post-tagging, lemmatization, parsing, annotation, coreference annotation, pragmatic annotation, stylistic annotation.*

References

Elov B., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlari uchun tf-idf statistik ko‘rsatkichni hisoblash. Science and innovation international scientific journal volume 1 issue 8 uif-2022: 8.2 ISSN: 2181-3337. 1774-1785 b.

O‘zbek tili ta’limiy korpusi - http://uzschoolcorpara.uz/

Elov B., Amirkulov M. Uzbek-English Parallel Corpus Algorithm and Alignment Problem. Central asian journal of literature, philosophy and culture eissn: 2660-6828 | Volume: 04 Issue: 06 June 2023. 71-78 p.

Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and Artificial Intelligence, Vol. 1 No. 3 (2023) 32–42. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v1i317. 117–129 b.

Elov B., Alayev R. O‘zbek tili korpusi va uning imkoniyatlari. O‘zbekiston Informatika va energetika mummolari jurnali. O‘zbekiston Jurnali. - Toshkent, 2023, - № 2. 89-100 b.

[1]*Primova Mastura Hakim qizi* – Alisher Navo’i Tashkent State University of Uzbek Language and Literature Teacher Department of Computational Linguistics and Digital Technologies.
E-mail: *primovamastura@navoiy-uni.uz*
ORCID: https://orcid.org/0000-0002-0241-4659

Elov B., Hamroyeva Sh., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlarini qayta ishlashda countvectorizer, tf-idf hamda co-occurrence matrix usullarining ahamiyati. ELEKTRON LUG‘ATLAR YARATISHNING NAZARIY VA AMALIY ASOSLARI mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari, 2023-yil 12-dekabr. Andijon. 78-88 b.

Zaxarov V.P. Korpusnaya lingvistika: Uchebno-metod. Posobiye. – SPb., 2005. -48s.

Boliang Zhang, Ajay Nagesh, Kevin Knight. 2020. Parallel Corpus Filtering via Pretrained Language Models. 10.18653/v1/2020.acl-main.756

Elaheh Rafatbakhsh, Alireza Ahmadi. 2019. “A thematic corpus-based study of idioms in the Corpus of Contemporary American English”. Atsian-Pacific Journal of Second and Foreign Language Education,10.1186/s40862-0190076-4

Politsin, S.A., Politsina Ye.V. 2018. “Primeneniye korpusa tekstov dlya avtomaticheskoy klassifikatsii v komplekse instrumentov avtomatizirovannogo analiza tekstov”. Vestnik VGU. Seriya: Sistemniy analiz i informatsionnie texnologii,10.17308/sait.2018.2/1224

ISSN 2181-922X 19-27

CLARIN TIZIMIDAGI OG‘ZAKI KORPUSLAR XUSUSIDA

Nilufar Muradova[1]

Annotatsiya. Raqamli texnologiyalarning rivojlangani barcha sohalarda o‘z aksini topmoqda. Xususan, tilshunoslikda til korpuslarini yaratish, tabiiy tilga ishlov berish (NLP), mashina tarjimasi masalalari dolzarb. Albatta, bu tadqiqotlar tilimizning rivojlanishi va yashovchanligini oshirishga xizmat qiladi. Bunda, asosan, til korpuslari ahamiyatlidir. Dunyo tilshunosligida mukammal, kengaytirilgan qidiruv imkoniyatiga ega yirik korpus tizimlari ishlab chiqilgan. Bulardan biri Clarin tizimidir. Clarin – til ma’lumotlarini kashf qilish, o‘rganish, izoh qo‘shish, tahlil qilish, birlashtirish va lingvistik tadqiqot o‘tkazish imkonini beradi. Bu tizimga bir qancha til korpuslari ham kiritilgan. Ushbu maqolada Clarin tizimining maqsad, vazifa, imkoniyatlari; tizimdagi korpus, subkorpus, shuningdek, og‘zaki korpuslar tavsiflangan. Og‘zaki korpuslarning turlari, imkoniyatlari va qidiruv tizimi bayon etilgan. Xususan, chex tili og‘zaki korpusi haqida umumiy ma’lumot, korpusning ishlash tizimi, subkorpusdan farqi, qidiruv imkoniyatlari o‘rganilgan.

Kalit so‘zlar: *Clarin tizimi, korpus, og‘zaki korpus, lemmatizatsiya, qidiruv tizimi.*

Kirish

Korpus lingvistikasi jahon kompyuter lingvistikasining juda tez rivojlanib kelayotgan sohasi bo‘lib, bu borada ancha yutuqlarga erishilgan. Oliy ta’lim muassasalarida korpus lingvistikasi fan sifatida ham o‘qitiladi. Bu sohaning predmeti korpus yaratish nazariyasi va amaliyoti bo‘lsa, fan sifatida korpusning o‘ziga xosligi, dasturlash asoslari kabi jihatlari o‘qitiladi. Korpus lingvistikasi kompyuter lingvistikasining tarkibiy qismi bo‘lib, til korpusini yaratish, kompyuter texnologiyasi yordamida ulardan foydalanishning umumiy nazariyasi va amaliyoti bilan shug‘ullanadi [Захаров, 2005. 48]. Dunyo tilshunosligida korpus tuzish, foydalanish va uni ishlab chiqish tamoyillari rivojlangani sari nafaqat yangi tizimlar

[1]*Muradova Nilufar Baxdir qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti kompyuter lingvistikasi mutaxassisligi magistranti.
E-pochta: *muradovanilufar06@gmail.com*
ORCID: 0009000184741863

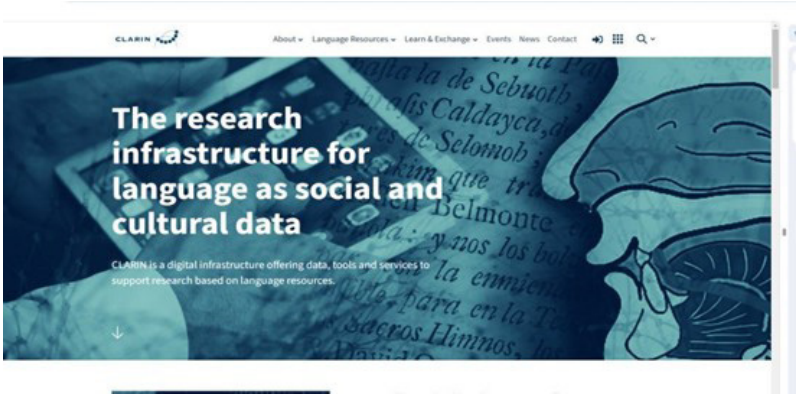
balki, mukammal korpuslar ham ishlab chiqilmoqda. Mana shunday tizimdan biri Clarin korpuslar jamlanmasidan iborat tuzilmadir. Clarin - til ma'lumotlarini kashf qilish, o'rganish, izoh qo'shish, tahlil qilish, birlashtirish va lingvistik tadqiqot o'tkazish imkonini beradi. Bu dastur 2019-yil fevral oyida tashkil qilingan (https://www.clarin.eu/). Clarin - bu til manbalariga asoslangan tadqiqotlarni qo'llab-quvvatlash uchun ma'lumotlar, vositalar va xizmatlarni taklif qiluvchi tizim.

Clarin tizimining maqsadi butun Yevropa va undan tashqaridagi barcha raqamli til manbalari va vositalariga gumanitar va ijtimoiy fanlar tadqiqotchilarini qo'llab-quvvatlash uchun yagona onlayn muhit yaratish. Clarinning vazifasi gumanitar va ijtimoiy fanlar bo'yicha tadqiqotlar uchun til ma'lumotlari va vositalarini almashish, ulardan foydalanish va barqarorligini qo'llab-quvvatlash uchun tizimlarni yaratish va saqlashdir.

Asosiy qism

"Og'zaki korpuslar" o'zbek tilida bo'lib, ingliz tilida "oral corpora" deb tarjima qilingan. "Og'zaki korpus" lingvistik tahlil uchun ishlatiladigan og'zaki til namunalari to'plamini anglatadi, asosan suhbatlar, intervyular, nutqlar yoki og'zaki muloqotning boshqa shakllarini yozib olish. Bu korpuslar tilning fonetika, sintaksis, semantika, nutq va sotsiolingvistika kabi turli jihatlarini o'rganish uchun qimmatli manbadir.

Clarindan foydalanish va unga til manbalari va vositalarini joylashtirish uchun tizimdan ro'yxatdan o'tish hamda a'zo bo'lish kerak. Hozirgi kunda 23 ta shahar va davlatlar ro'yxatdan o'tgan va a'zo bo'lgan. Bular: Avstriya, AQSH, Buyuk Britaniya, Belgiya, Bolgariya, Xorvatiya, Kipr, Chexiya, Daniya, Estoniya, Finlandiya, Germaniya, Gretsiya, Vengriya, Islandiya, Italiya, Latviya, Litva, Niderlandiya, Janubiy Afrika, Shvetsariya, Kolumbiya, Portugaliya. Clarin tizimining bir qancha markazlari mavjud. Markazning 3 ta turi mavjud: B markazlari, C markazlari va K markazlari. Ularning har biri turli xil tajriba va xizmatlarni taklif etadi. Maxsus Clarin markazlari (Clarin centre) 50 ga yaqinni tashkil etadi. Barcha tashkil etilgan Clarin markazlarining to'liq ro'yxati https://www.clarin.eu/content/overview-clarin-centres saytida mavjud. Clarin tizimi 6 bo'limdan iborat (1-rasm).



1-rasm. Clarin tizimi interfeysi

Birinchi bo‘limda (About) tizim haqida umumiy ma’lumot, Clarin tizimining markazlari, boshqaruv tizimi, a’zo bo‘lgan davlatlar va ishlash texnologiyasi haqidagi dastlabki ma’lumotlar aks etgan.

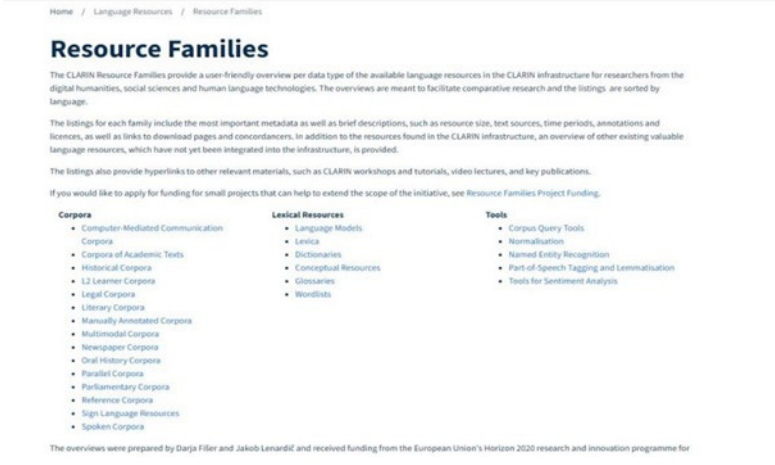
Ikkinchi bo‘limda (Language Resources) foydalanish uchun qulay til manbalari. Bu yerda siz Clarin tizimining til manbalariga osongina kirishingiz mumkin. Nutq va til ma’lumotlarining keng turlarini, shuningdek, ma’lumotlarni qayta ishlash uchun dasturiy vositalar va xizmatlardan foydalanishingiz mumkin. Yozma va og‘zaki korpuslar, shu jumladan, multimodal resurslari va ma’lumotlar bazasi mavjud. Undan tashqari Clarin til ma’lumotlarini izohlash, tahlil qilish yoki birlashtirish uchun turli xil vositalar va xizmatlarni taklif etadi. Taklif qilinayotgan narsalarni ko‘rib chiqish yoki ehtiyojlaringizga mos keladigan maxsus vositalarni tanlash mumkin. Shuningdek, tizimdagi mavjud til resurslarining ma’lumotlar turi bo‘yicha foydalanuvchilarga qulay ma’lumot beradi.

Uchinchi bo‘limda (Learn Exchange) amaliy tadqiqotlar, foydalanuvchilarni jalb qilish bo‘yicha mavjud tadbirlar, ishlab chiqilgan vositalar va tizimdan foydalanayotgan taniqli tadqiqotchilar bilan suhbatlar joylashgan.

To‘rtinchi bo‘limda (Events) tadbirlar, konferensiyalar, seminarlar va Clarin tizimi bilan bog‘liq boshqa tadbirlar ro‘yxati mavjud. Shuningdek, tadbirlar o‘tkazilgan vaqti, joyi, tadbir nomi, shahri ham to‘liq ko‘rsatilgan. Foydalanuvchi istagan tadbir namoyishi va materiallari bilan tanishish va foydalanish imkoniyatiga ega.

Beshinchi bo‘limda (News) tizimdagi va tizim bilan bog‘liq yangiliklar va o‘zgarishlar, tizimga yangi qo‘shilgan lingvistik tahlillar natijasi, eng so‘nggi yangiliklar bayon etilgan.

Oltinchi bo‘limda (Contact) tizim bilan bog‘lanish manzillari joylashtirilgan. E-mail, telefon raqam, yangiliklar sayti, aloqa uchun saytlar ko‘rsatilgan. Ushbu tizimning “Resource families” qismida ko‘pgina til korpuslari jamlangan. Xususan, vazifasi va maqsadiga ko‘ra 15 ta korpus mavjud bo‘lib, ular o‘z ichiga yana bir qancha subkorpuslarni qamrab oladi (2-rasm).


Home / Language Resources / Resource Families

Resource Families

The CLARIN Resource Families provide a user-friendly overview per data type of the available language resources in the CLARIN infrastructure for researchers from the digital humanities, social sciences and human language technologies. The overviews are meant to facilitate comparative research and the listings are sorted by language.

The listings for each family include the most important metadata as well as brief descriptions, such as resource size, text sources, time periods, annotations and licences, as well as links to download pages and concordancers. In addition to the resources found in the CLARIN infrastructure, an overview of other existing valuable language resources, which have not yet been integrated into the infrastructure, is provided.

The listings also provide hyperlinks to other relevant materials, such as CLARIN workshops and tutorials, video lectures, and key publications.

If you would like to apply for funding for small projects that can help to extend the scope of the initiative, see Resource Families Project Funding.

Corpora
- Computer-Mediated Communication Corpora
- Corpora of Academic Texts
- Historical Corpora
- L2 Learner Corpora
- Legal Corpora
- Literary Corpora
- Manually Annotated Corpora
- Multimodal Corpora
- Newspaper Corpora
- Oral History Corpora
- Parallel Corpora
- Parliamentary Corpora
- Reference Corpora
- Sign Language Resources
- Spoken Corpora

Lexical Resources
- Language Models
- Lexica
- Dictionaries
- Conceptual Resources
- Glossaries
- Wordlists

Tools
- Corpus Query Tools
- Normalisation
- Named Entity Recognition
- Part-of-Speech Tagging and Lemmatisation
- Tools for Sentiment Analysis

The overviews were prepared by Darja Fišer and Jakob Lenardič and received funding from the European Union's Horizon 2020 research and innovation programme for

2-rasm. Clarindagi subkorpuslar

1. Computer-Mediated Communication Corpora. Kompyuter vositasidagi aloqa korpuslari. Kompyuter vositasidagi onlayn muloqotlarni o‘z ichiga oladi. Masalan, yangiliklar, saytdagi sharhlar, ijtimoiy tarmoqdagi ilovalar. Onlayn muloqotda ko‘pgina til o‘zgarishlariga uchraydi. Imloviy, ishoraviy xatoliklar kuzatiladi. Clarin tizimi 23 ta CMCni taqdim etadi. Ularning aksariyati, sloveniyaliklar uchun, shuningdek, golland, chex, fin, nemis, italyan, litva tillari uchun ham mavjud.

2. Corpora of Academic Texts. Akademik matnlar korpusi. Akademik matnlar ilmiy jurnallarda chop etilgan ilmiy maqola, tezis, konferensiya materiallari va monografiyalarni o‘z ichiga oladi. Ushbu tizimda 24 ta akademik matnlar korpusi mavjud, ulardan 2 tasi ko‘p tilli va 22 tasi bir tilli hisoblanadi.

3. Historical Corpora. Tarixiy korpus.

4. L2 Leaner Corpora. Ushbu turdagi korpusning o‘ziga xos xususiyatidan biri, yangi til o‘rganuvchilar o‘z xatolarini bilib borishlari mumkin.

5. Legal Corpora. Huquqiy korpus. Qonun hujjatlari, huquqiy hujjatlar, sud qarorlari va shunga oid materiallarni qamrab oladi.

6. Literary Corpora. Adabiy korpus. Adabiy korpus she’r va

nasriy asarlar, masalan, qissa, hikoya, roman, dramalarni o‘z ichiga oladi. Ushbu tizimda 45 ta adabiy korpusga kirish mumkin.

7. Manually Annotated Corpora. Izohli korpus. Lingvistik ma’lumotlarni qamrab olgan.

8. Multimodal Corpora. Multimediali korpus. Vidio, tasvir va yozuv ko‘rinishida ma’lumotlarni taqdim etadi.

9. Newspaper Corpora. Gazeta korpuslari. Gazeta to‘plamlari, ommaviy-axborot vositalari mavjud.

10. Oral History Corpora. Og‘zaki tarixiy korpus. Bir shaxs yoki ma’lum guruhning tarixi bilan bog‘liq bo‘lgan ma’lumotlarini to‘plash bilan shug‘ullanadi.

11. Parallel Corpora. Parallel korpus. Bunday korpuslar til o‘rganuvchilar uchun asosiy manba bo‘lib xizmat qiladi. Asosan, tarjima sohasida keng qo‘llaniladi.

12. Parliamentary Corpora. Parlament korpuslari.

13. Reference Corpora. Ma’lumotlar korpusi.

14. Sign Language Resources. Imo-ishoralar tili resurslari.

15. Spoken Corpora. Og‘zaki korpus. Bunda og‘zaki nutq, dialog, efirga uzatilgan ko‘rsatuvlarni o‘z ichiga oladi. Tilshunoslikda, xususan, dialektologiyada muhim manba bo‘lib xizmat qiladi.

Har bir korpusda korpus nomi, korpus tili va korpus haqida dastlabki umumiy ma’lumotlar berilgan bo‘lib, ularga kirish uchun havola va yuklab olish imkoniyati ham mavjud. Birgina og‘zaki korpus (faqat og‘zaki matnlardan tarkib topgan korpus [Hamroyeva, 2020. 34]) tarkibida 148 ta og‘zaki korpus mavjud bo‘lib, ulardan 134 tasida og‘zaki va matn ko‘rinishida, 14 tasida faqat transkripsiyasi mavjud. Korpuslarning aksariyati bir tilli bo‘lib, quyidagi 15 tilni qamrab oladi: Arab, chex, golland, eston, fin, fransuz, nemis, venger, italyan, nepal, norveg, polyak, sloven, ispan va shved. Aksariyat hollarda korpusni to‘g‘ridan-to‘g‘ri ma’lumotlar bazasidan yuklab olish yoki foydalanish uchun qulay onlayn qidiruv muhiti orqali foydalanish mumkin. Og‘zaki tilning korpuslari o‘z-o‘zidan yoki rejalashtirilgan nutqning transkripsiyalarini o‘z ichiga oladi, masalan, efirga uzatilgan yangiliklar yoki olingan rivoyatlar va dialoglar. Ular ko‘pincha ilova qilingan yozuvlar bilan izohlanadi. Ular fonologiya, suhbat tahlili va dialektologiya kabi turli xil lingvistik tadqiqotlar uchun qimmatli manbadir.

Og‘zaki korpusda bir tilning bir nechta korpuslari mavjud. Ular bir-biridan yaratilgan maqsadi va vazifasiga ko‘ra farqlanadi. Masalan ingliz tilining 5 ta korpusi kiritilgan bo‘lsa-da, ular mavzuviy jihatdan farqlanadi. 1. Aviadispetcherlar va uchuvchilar

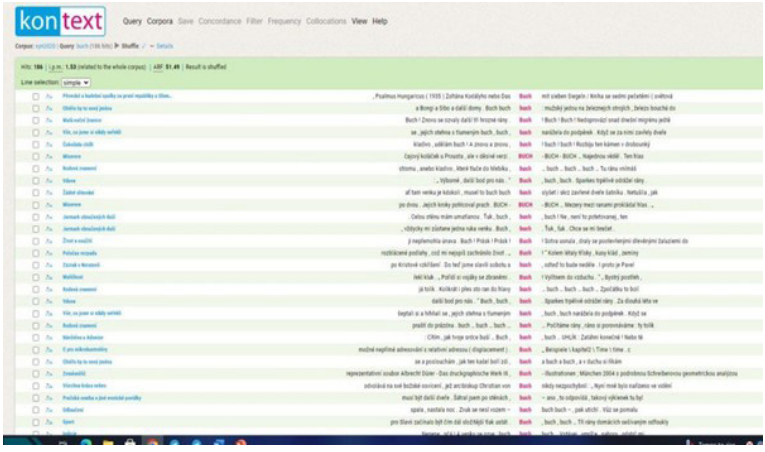
o‘rtasidagi aloqa yozuvlari mavjud. 2. Radio yangiliklaridan yozuvlar va matnlar mavjud. 3. Ushbu korpusda intervyu mavjud. 4. Yozilgan ma’ruzalar va seminarlar mavjud. 5. Ushbu korpusda ingliz tilidagi asosiy bog‘langan nutq jarayonlarini kiritish uchun mo‘ljallangan 460 ta qisqa jumlalar to‘plami mavjud. Ma’lumotlar bazasida audio fayllar, ovoz to‘lqin shakllari mavjud. Har bir til korpusida korpus turi haqida (Corpus), korpus tili (Language), korpus haqida umumiy ma’lumot (Description), yuklab olish bo‘limi (Availability) mavjud. Birgina chex tilining tizimda 3 xil korpusi joylangan.

1. Ko‘p bosqichli transkripsiyaga ega dialektal korpus. Hajmi 100000 so‘zdan iborat. Orfografik va fonetik (dialekt xususiyatlari) transkripsiyalangan, lemmatizatsiyalangan. Ushbu korpusda an’anaviy dialektologik material, asosan, monolog tipidagi nutqlar mavjud.

2. Norasmiy chex tilining korpusi (transkripsiya va audio). Hajmi 2,8 mln so‘zdan iborat. Ushbu korpus norasmiy suhbatlarni o‘z ichiga oladi.

3. Og‘zaki chex tili ma’lumotlar korpusi. Hajmi 770 000 token, 7 324 daqiqani tashkil etadi. Lemmatizatsiyalangan. Nutqlar, asosan dialog shaklda mavjud.

Dialekt korpusi butun Chexiya Respublikasida qo‘lga kiritilgan an’anaviy mintaqaviy dialektlarni taqdim etadi. Dialekt materiali Chexiya Respublikasining barcha dialektal mintaqalaridan kelgan ovoz yozuvlarini transkripsiya qilish orqali olingan. Korpus ikki darajadan iborat. Qadimgi dialektal daraja 1950-yillarning oxiridan 1980-yillargacha bo‘lgan davrda yozilgan yozuvlarni o‘z ichiga oladi. Yangi daraja 1990-yillardan hozirgi kungacha bo‘lgan davrni o‘z ichiga olgan ikkala qatlam uchun ham bizda bugungi kunda umuman uchramaydigan arxaik dialektal elementlarni o‘z ichiga olgan til ma’lumotlari mavjud. Dialekt korpusining ikkinchi versiyasi 220 000 dan ortiq so‘zlarni o‘z ichiga oladi. Birinchi dialektal korpusda 9745 ta, ikkinchi dialektal korpusda esa 43 628 ta nutq mavjud. Ushbu dialektal korpusda “buch” (kitob) so‘zi orqali qidiruv amalga oshirilganda quyidagi Kwic formatdagi natija olindi. Ushbu so‘z 186 ta matnda uchraydi (3-rasm).



3-rasm. Qidiruv natijasi ko‘rinishi

Ko‘rsatilgan barcha ma’lumotlar avtomatik tahlil orqali olinadi. Ma’lumotlarning aniqligi va ishonchliligi korpusdagi teglashning mukammalligiga bog‘liq. Shu sababli lemmatizatsiya, ya’ni asosiy so‘zshaklini ajratish va morfologik teglash muhim rol o‘ynaydi. Tizimdagi va korpusdagi ma’lumotlarda doimiy ravishda yangilanib, to‘ldirilib boriladi.

Xulosa

Olimlar korpusning maqsadi hamda oldiga qo‘yilgan vazifasiga qarab korpusning bir qancha tasnifini keltirishadi. Clarin tizimidagi korpuslar saqlash shakliga ko‘ra (ovozli, yozma, aralash) aralash; matn tiliga ko‘ra (bir tilli, ko‘p tilli) bir tilli, janriy mansubligiga ko‘ra (adabiy, dialektal, og‘zaki, publitsistik, aralash) aralash; korpusga kirish imkoniyatiga ko‘ra (erkin, tijorat korpuslari, yopiq) erkin; maqsadiga ko‘ra (tadqiqiy, illyustrativ) illyustrativ; dinamikasiga ko‘ra (dinamik, turg‘un) dinamik (chunki bu korpuslar yangilanib, to‘ldirilib boriladi); qo‘shimcha axborotga egaligiga ko‘ra (teglangan, teglanmagan) teglangan. Clarin tizimi mukammal korpuslar yig‘ilgan tizimdir. Unda nafaqat korpuslarni undan tashqari barcha raqamli til manbalari va vositalariga gumanitar va ijtimoiy fanlar tadqiqotchilarini qo‘llab-quvvatlash uchun yagona online muhit mavjudligi, keng qidiruv imkoniyatiga egaligi bilan ahamiyatlidir. Ma’lum bir tilda yaratilgan korpuslarning mukammalligi foydalanuvchiga ham, yangi til o‘rganuvchiga ham qulaylik yaratadi. Korpus qanchalik mukammal razmetkalangan bo‘lsa, uning qidiruv imkoniyati ham, olingan natija ham shunchalik aniq mukammal bo‘ladi.

Foydalanilgan adabiyotlar

Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. – Globe edit, 2020.

Hamroyeva Sh. Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati: Terminologik lug‘at. – Globe edit, 2020.

Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005.

Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2013.

https://www.clarin.eu/

https://www.clarin.eu/content/overview-clarin-centres

SPECIFICALLY ORAL CORPUSES IN THE CLARIN SYSTEM

Nilufar Muradova[1]

Abstract. The development of digital technologies is reflected in all areas. In particular, the issues of creating language corpora, natural language processing (NLP), machine translation are relevant in linguistics. Of course, these studies serve to increase the development and viability of our language. Mainly, language corpora are important in this. In world linguistics, large systems with excellent, extended search capabilities have been developed. One of these is the Clarin system. Clarin allows you to discover, explore, annotate, analyze, integrate and perform linguistic research on language data. Several language corpora are also included in this system. In this article, the purpose, mission, capabilities of the Clarin system; corpus, subcorpus, and verbal corpuses in the system are described. The types, possibilities and search system of oral corpora are described. In particular, general information about the spoken corpus of the Czech language, the working system of the corpus, the difference from the subcorpus, and search possibilities were studied.

Key words: *Clarin system, corpus, verbal corpus, lemmatization, search system.*

References

Hamroyeva Sh. O‘zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Monografiya. – Globe edit, 2020.

Hamroyeva Sh. Korpus lingvistikasi atamalarining qisqacha izohli lug‘ati: Terminologik lug‘at. – Globe edit, 2020.

Zaxarov V.P. Korpusnaya lingvistika. Uchebno-metodicheskoye posobiye. – Sankt-Peterburg, 2005.

Zaxarov V.P., Bogdanova S.Y. Korpusnaya lingvistika. – Irkutsk: IGLU, 2013.

https://www.clarin.eu/

https://www.clarin.eu/content/overview-clarin-centres

[1]*Muradova Nilufar Baxodir qizi* – Alisher Navoiy nomidagi o‘zbek tili va adabiyoti universiteti kompyuter lingvistikasi mutaxassisligi magistranti.
E-pochta: *muradovanilufar06@gmail.com*
ORCID: 0009000184741863

ISSN 2181-922X 28-37

IBORALARNI MOSLASHTIRISH (PHRASE ALIGNMENT)DA OTLI VA FE'LLI SO'Z BIRIKMALAR MOSLIGI

Noila Matyakubova[1]

Annotatsiya. Moslashtirish jarayoni tabiiy tilga ishlov berish (NLP)ning ko'plab vositalari samaradorligini oshirish uchun juda muhim omil hisoblanadi. Asosan turli tuzilishga ega bo'lgan tillarni moslashtirishda so'z va iboralarni moslashtirish muhim ahamiyat kasb etadi. Bu moslashtirish bosqichlari gaplarni va abzatslarni moslashtirish bosqichlarini tog'ri va samarali ishlashini ta'minlab beradi. Ushbu maqolada o'zbek va ingliz tillaridagi so'z birikmalarini moslashtirishda yuzaga keluvchi holatlar ko'rib chiqiladi.

Kalit so'zlar: *So'z birikmalari, otli birikma, fe'lli birikma, token, moslashtirish jarayoni, erkin birikma, turg'un birikma.*

Kirish

So'z birikmalari gapda malum ma'nolarni ifodalab keluvchi birikmalar bo'lib, ma'lum grammatik qurilmaga ega bo'lishi bilan oddiy so'zlardan ajralib turadi. NLPning ko'plab sohalarida, asosan, mashina tarjimasida, parallel matnlarni moslashtirish jarayonida manba til va maqsadli tilda qo'llaniladigan iboralar va ularning grammatik tuzilishi, semantikasi haqida to'liq tushunchaga ega bo'lish talab etiladi. Asosan turlicha grammatik tuzilishga ega bo'lgan tillarda ularni tuzilishida va ma'nosida katta farq bo'ladi. Ingliz va o'zbek tillari aynan shunday tillar qatoriga kiradi va ba'zi hollarda moslikni topishda bir nechta murakkabliklar keltirib chiqaradi [MacCartney, Galley, 2008]. O'zbek-ingliz so'z birikmalarining to'liq jamlanishi va mashinani o'qitish (Machine learning) jarayoni orqali moslashtirish vositalarini, mashina tarjimonlarini va parallel matnlar bilan ishlovchi turli vostalarning samaradorligini oshirish mumkin.

[1]*Matyakubova Noila Shakirjanovna* – Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti tayanch doktoranti.
E-pochta: *matyakubovanoila@navoiy-uni.uz, nailya89mm@mail.ru*
ORCID: 0009-0009-3154-723X

Asosiy qism

O'zbek tilida so'z birikmalari birikma qismlarining grammatik tabiatiga ko'ra va tuzilishiga ko'ra tasnif qilinadi, ya'ni grammatik tabiatiga ko'ra tasnif qilinganda hokim a'zoning qaysi so'z turkumi ekanligi va tobe a'zoning sintaktik vazifasi asosga olinadi. Ular turg'un va erkin so'z birikmalari ko'rinishida uchraydi, erkin birikmaning o'zi ham o'z navbatida teng bog'lanish va tobe bog'lanish usullarida yasaladi [Abdullayev, 1976]. Tarkibiy jihatiga ko'ra ingliz tilidagi so'z birikmalari bilan deyarli bir xil bo'lib ***bosh so'z*** (Head word) va ***ergash so'z*** (Auxilary word)dan tashkil topadi [Brinton, 2000]. Har ikkala tilda bosh so'z qaysi so'z turkumiga kirsa, aynan o'sha so'z turkumli so'z birikmasi deb yuritiladi. Tuzilishiga ko'ra har ikkala tilda oddiy so'z birikmasi va murakkab so'z birikmasi deb yuritiladi, ammo ingliz tilida o'zbek tilidan farqli ravishda qo'shma so'z birikmasi ham mavjud. Ingliz tilida so'z birikmalarning bir nechta turlari mavjud bo'lib, biz ularni otli birikma (noun phrase), fe'lli birikma (verb phrase), sifatli birikma (adjective phrase), ravishli birikma (adverbial phrase), predlogli birikma (prepositional phrase), verbal phrase, absolute phrase va bundan tashqari frazema (phrasal verb), kollakatsiya (collocation) va ibora (idiom) ko'rinishida uchratamiz [Brinton, 2000]. Ammo o'zbek tilda ular ingliz tilidagidan birmuncha farq qiladi va otli, fe'lli, sifatli, ravishli va modal so'zli birikma ko'rinishida yasaladi. Parallel matnlarni tarjima qilinganda asos tildagi gapda uchraydigan so'z birikmalari maqsadli tilda ham birikma holatida ham tarqoq holatda kelishi mumkin. Ushbu turdagi so'z birikmalari moslashtirish jarayonida bir nechta chalkashliklarga olib keladi. Maqolada o'zbek va ingliz tilidagi otli va fe'lli so'z birikmalarining yasalishi va gapda qo'llanilishini batafsil ko'rib chiqamiz.

O'zbek tilida otli birikma (NP) bosh so'z vazifasida ot keladi, u bitta yoki ikkita so'zlarning bog'lanishidan hosil bo'ladi. NP qismlari bitishuv, muvofiqlashuv hamda boshqaruv yo'li bilan grammatik aloqaga kirishadi va quyidagicha yasaladi:

• Bosh so'z bo'lib ot keladi: bag'rikeng inson, mard yigit.

• Bosh so'z bo'lib sifat keladi: onaning o'zi, Noilaning o'zi.

• Bosh so'z bo'lib otlashgan sifat keladi: talabalardan eng aqllisi, gullarning sarasi.

• Bosh so'z bo'lib otlashgan son keladi: talabalardan biri, ularning uchovi.

• Bosh so'z bo'lib otlashgan sifatdosh keladi: bizning aytayotganimiz, ularning o'qiyotganlari.

• Bosh so'z bo'lib otlashgan ravish keladi: talabalarning ko'pi, so'zlarining ozi.

• Bosh so‘z bo‘lib otlashgan undov so‘z keladi: nochorlarning dod-voyi.

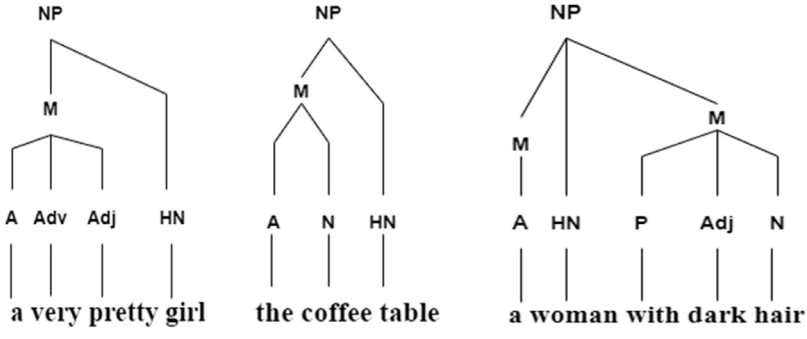
• Bosh so‘z bo‘lib otlashgan taqlid so‘z keladi: kosalarning qasir-qusiri, otlarning dupur-dupuri.

Ingliz tilida NP ot so‘z turkumiga sifat, olmosh yoki artikl kabi aniqlovchilarning birikishi orqali hosil qilinadi va ular erkin birikma hisoblanadi. Misol uchun, **my family, a tall building, a very pretty girl.** O‘z o‘rnida NP uch turga bo‘linadi: ***simple NP (sodda), compound NP (qo‘shma), complex NP (murakkab).***

1-Jadval. Ingliz tilida NPning yasalishi

NP	Yasalishi	Misol
Simple NP	Aniqlovchi + Ot	The black car, a friendly dog
Compound NP	Aniqlovchi + ot + ot	The swimming pool, a cotton shirt
Complex NP	Aniqlovchi + ot + aniqlovchi (ergash gap bilan aniqlanishi ham mumkin)	The lady waiting outside (The lady who is waiting outside)

Moslashtirish jarayonidagi tokenizatsiyadan foydalanilganda hosil bo‘ladigan tokenlardagi tafovut ham aynan birikmalardagi qatnashgan so‘zlarning grammatik jihatdan har ikkala tilda mavjudligi va ularning gapdagi o‘rniga ham bog‘liq bo‘ladi. Misol uchun ba’zi hollarda otga birikib kelgan qism otning oldida emas otdan keyin ham joylashishi mumkin. Bunday hollarda otdan oldin va keyin kelgan so‘zlar aniqlovchi vazifasini bajarib, ingliz tilida otdan oldin keladiganlari premodifier va otdan keyin kelganlari modifier deb ataladi. Bir nechta modifierlarning qatnashishi asosan murakkab NPda kuzatiladi va otga predlogli birikmaning birikishi orqali sodir bo‘ladi. Buni quyidagi misollarda ko‘rib chiqishimiz mumkin:



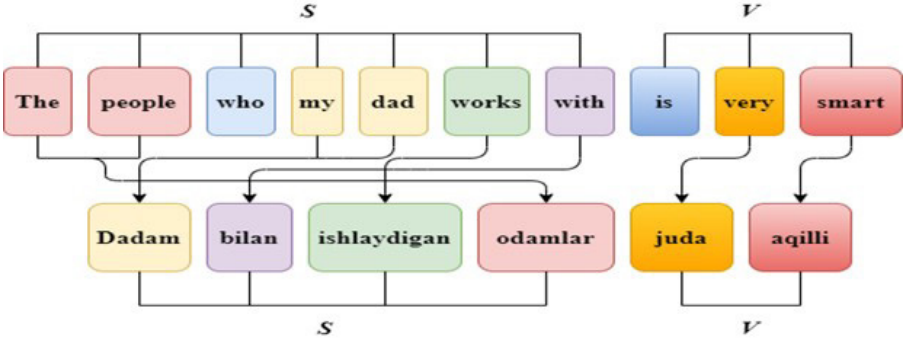
1-rasm. Sodda NP, qo‘shma NP, murakkab NP

Berilgan misollarda A- artikl, Adv- ravish, Adj- sifat, N- ot, HN- asosiy ot, M- aniqlovchi, NP- otli birikmani ifodalaydi. Ot oldidan kelgan artikl, ravish va sifat otga birikib NPni hosil qilyapti. Qo'shma va murakkab NP bir nechta otlar qatnashishi orqali yasalib, ulardan bittasi asosiy ot (head noun) hisoblanadi va gapning kesimi aynan o'sha HNga moslashishi talab etiladi. Misol uchun, "**A child holding a lot of toys is my son**" gapini tahlil qiladigan bo'lsak, gapning egasi **A child holding a lot of toys**, kesim bo'lib kelgan **is** murakkab NP ko'rinishida kelgan yaxlit egaga emas HNga moslashishi talab etiladi. Bunda HN va kesim oralig'idagi barcha so'zlar modifier hisoblanadi. Bundan tashqari murakkab NP **relative clause (RC)** yordamida ham yasalib, NPda kelgan HNdan keyin bog'lovchi va ega kesimli gap keladi, ya'ni qo'shma gap ko'rinishida. Ammo bunday shaklda yasalgan qo'shma gaplar o'zbek tiliga tarjima qilinganda, sodda gap ko'rinishiga keladi, chunki ingliz tilidagi bog'lovchiga birikib kelgan ega kesimli bo'lakning vazifasi HN aniqlash bo'lib modifier vazifasini bajaradi va gapning istalgan qismida kelishi mumkin.

$$S\ [\text{RC}]+V+O\ [\text{RC}]$$

NP gapda ega va to'ldiruvchi vazifasida kelishi mumkin. Bunday holatda RC yordamida yasalgan murakkab NP bog'lovchi yordamida yasalishi yoki sifatdosh yordamida yasalishi ham mumkin. Har ikkala holatda ham o'zbek tiliga bir xil tarjima qilinadi, ammo tokenlar sonida va gap turlarida farq qiladi. Ingliz tilida berilgan gap qo'shma gap shaklida bo'lsa, o'zbek tilida sodda gap ko'rinishida keladi.

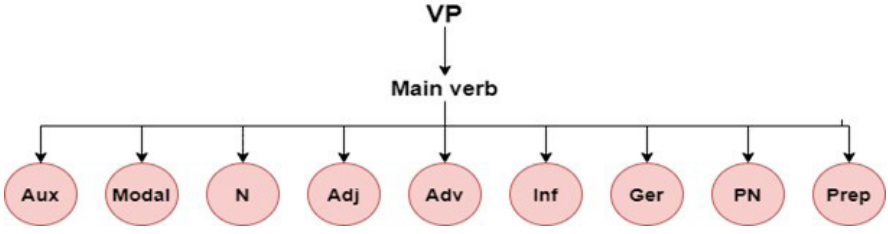
The man who is working in the garden is my dad.	***Bog'da ishlayotgan kishi mening dadam.***
The man working in the garden is my dad.	



2-rasm. Murakkab NP qatnashgan gapning moslashuvi

O‘zbek tilida fe’lli birikma (VP) bosh so‘z sifatida fe’lning kelishi bilan ifodalanadi va fe’lning boshqa boshqa so‘zlar bilan birikib kelishi orqali hosil bo‘ladi. Misol uchun: *tez gapirmoq, kulib yubormoq, diqqat bilan tinglamoq.*

Ingliz tilida VP gaplarni grammatik jihatdan to‘g‘ri tuzish, gapning ma’nosini to‘liq yetkazish uchun zarurdir. VP odatda asosiy fe’l va bir yoki bir nechta yordamchi fe’llarni birikib kelishidan hosil bo‘ladi. Bundan tashqari fe’l, sifat, ravish, ot, fe’lning otlashgan shakli (infinitive/gerund) va inkor yuklamalari bilan ham birikib kelishi mumkin. Ingliz tilida hozirgi va o‘tgan oddiy zamondagi gaplarning darak shakli asosiy fe’lning o‘zi bilan yasaladi, ammo ularning inkor va so‘roq shakllari va qolgan barcha zamonlar VP yordamida yasaladi.



3-rasm. Ingliz tilida VPning yasalishi

VP yasalishi jihatdan ***essential*** va ***non-essential VP*** ga ajraladi. Essential VP gapning grammatik jihatdan to‘g‘ri shakllantirish uchun juda muhim bo‘lib, asosan ko‘makchi va modal fe’llar yordamida yasaladi. Misol uchun, **“will have gone”** fe’l birikmasi Future perfect tense da yasalishi uchun muhim bo‘lsa, **“will probably have gone”** non-essential fe’l birikmasi hisoblanadi, chunki bunda probably gapda ehtimollik holatini ifodalash uchun ishlatilib, gapning ma’nosini ifodalshda ahamiyatli bo‘lsa gapning grammatik jihatdan to‘g‘ri shakllantirish uchun muhim emas. Phrase alignment (birikmalarni moslashtirish) jarayonida o‘zbek va ingliz tilidagi birikmalarini moslashtirishda yuzaga keluvchi ko‘plab tafovutlar aynan VPlarning bir-biridan tubdan farq qilishida hisoblanadi.

2-Jadval. Ingliz tilidagi turg‘un VP larning yasalishi

Tenses	Auxilaries	Modals
Present Simple	-	Modal+V1
Present Continuous	To be (am/is/are) + Ving	Modal be + Ving
Present Perfect	To have been + V3	Modal have been + V3
Present Perfect Continuous	To have been + Ving	Modal have been + Ving
Past simple	-	Modal have been + V3

Past Continuous	To be (was/were) + Ving	Modal have been + Ving
Past Perfect	Had been + V3	Modal have been + V3
Past Perfect Continuous	Had been + Ving	Modal have been + Ving
Future Simple	Will + V1	Modal+V1
Future Continuous	Will be + Ving	Modal be + Ving
Future Perfect	Will have been + V3	Modal have been + V3
Future Perfect Continuous	Will have been + Ving	Modal have been + Ving

Ushbu jadvalda berilgan VP da qatnashgan ko'makchi va modal fe'llar gapda zamonlarning to'g'ri shakllantirish uchun muhim hisoblanadi, ammo o'zbek tiliga tarjima qilganimizda ko'makchi fe'llar tarjimaga ega bo'lmaydi va tokenlardagi asosiy farqlar aynan shundan kelib chiqadi. Misol uchun, "*He **has been working** in the office since 7 p.m*" ushbu gap Present Perfect Continuous zamonida yasalib gapda murakkab kesim bo'lib kelgan VP, **has been working**, uchta tokendan tashkil topgan bo'lsa, "*U soat yettidan beri **ishlayapti***", o'zbek tilidagi tarjimasida, ***ishlayapti***, ya'ni bitta tokenga to'g'ri keladi.

3-Jadval. Qoidaga asosan so'z birikmalarini moslashtirish

So'z birikmasining turi	O'zbek tilida	Ingliz tilida	Tokenlardagi farq	Moslashtirishdagi muammolar va sababi
Otli birikma	Talabalardan biri	One of the students	2	O'zbek tilida mavjud bo'lmagan grammatik shakllarning qo'llanilishi
	Bag'ri keng inson	a generous person	Yo'q	Ayni ma'noni ifodalovchi lekemaning tarjimasi yo'q, "bag'rikeng" so'zi ko'chma ma'noda kelgan; O'zbek tilida mavjud bo'lmagan grammatik shakllarning qo'llanilishi

Fe'lli birikma	Yo'qolgan bo'lsa kerak	Must have been lost	1	O'zbek tilida mavjud bo'lmagan grammatik shakllarning qo'llanilishi
	Diqqat bilan tingladi	Listened carefully	1	Ingliz tilida mavjud bo'lmagan grammatik shakllarning qo'llanilishi

Xulosa

So'z birikmalarini moslashtirish birmuncha murakkab jarayon bo'lib, umumiy moslashtirish jarayoniga qaraganda ancha chalkashliklarga olib keladi. Turg'un so'z birikmalarining jamlanishi va erkin birikmalarning qoidalarini mashinaga to'liq o'rgatish orqali bu muammolarni hal qilish mumkin hisoblanadi. Olib borilayotgan ilmiy ishimiz doirasida o'zbek va ingliz tilidagi turg'un so'z birimalari bazasi va erkin so'z birikmalarining qoidalar to'plami tuzilib, inson aralashuvi orqali kompyuterga o'rgatilmoqda va bu "Aligner" dasturiy vositasining samaradorligini oshiruvchi birlamchi omillardan deb hisoblanmoqda.

Foydalanilgan adabiyotlar

B. MacCartney, M. Galley, Ch. Manning, "A Phrase-Based Alignment Model for Natural Language Inference". Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008.

Abdullayev F.A. O'zbek tili grammatikasi. 2-jild sintaksis.18-39-betlar. 1976.

Laurel J. Brinton, "The Structure of Modern English" John Benjamins Publishing Company Amsterdam /Philadelphia, 2000.

Y. Arase, J. Tsujii, "Compositional Phrase Alignment and Beyond". Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1611–1623, November 16–20, 2020.

Abdullayeva O., Xudayarova S. O'zbek tilshunosligida so'z birikmasiga ta'rif, tavsif va tasnif masalasi. // International scientific-theoretical conference on the topic: «Problems of research and education of the Uzbek language» www.myscience.uz.

Bas Aarts, Liliane Haegeman, "English Word Classes and Phrases", Bas Aarts and Liliane Haegeman, January 2008.

Guelailia Ahmed, "Lists of general phrases that can be very helpful to write a good research paper." https://www.researchgate.net/publication/340666349

Khin Thandar Nwet, "Developing Word to Phrase Alignment for Myanmar-English Machine Translation", 13th International Conference on Computer Applications, 2015.

R. Sennrich, Volk, Martin (2011). "Iterative, MT-based sentence alignment of parallel texts." In: NODALIDA 2011, Nordic Conference of Computational Linguistics, Riga, 11 May 2011 - 13 May 2011.

https://www.masterclass.com/articles/what-is-a-verb-phrase

ALIGNING NOUN AND VERB PHRASES IN PHRASE ALIGNMENT

Noila Matyakubova[1]

Abstract. The alignment process is a critical factor in maintaining much of the performance of natural language processing (NLP). Aligning words and phrases is important in alignment of the languages that have different grammatical sentence structures. These alignment steps ensure that sentence and paragraph alignment steps work correctly and efficiently. In this article alignment process of Uzbek and English phrases have been analyzed.

Key words: *Phrasal verbs, verb phrase, noun phrase, token, alignment, free phrase, fixed phrase.*

References

B. MacCartney, M. Galley, Ch. Manning, "A Phrase-Based Alignment Model for Natural Language Inference". Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008.

Abdullayev F.A. O'zbek tili grammatikasi. 2-jild sintaksis.18-39-betlar. 1976.

Laurel J. Brinton, "The Structure of Modern English" John Benjamins Publishing Company Amsterdam /Philadelphia, 2000.

Y. Arase, J. Tsujii, "Compositional Phrase Alignment and Beyond". Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1611–1623, November 16–20, 2020.

Abdullayeva O., Xudayarova S. O'zbek tilshunosligida so'z birikmasiga ta'rif, tavsif va tasnif masalasi. // International scientific-theoretical conference on the topic: «Problems of research and education of the Uzbek language» www.myscience.uz.

Bas Aarts, Liliane Haegeman, "English Word Classes and Phrases", Bas Aarts and Liliane Haegeman, January 2008.

Guelailia Ahmed, "Lists of general phrases that can be very helpful to write a good research paper." https://www.researchgate.net/publication/340666349

Khin Thandar Nwet, "Developing Word to Phrase Alignment for

[1]*Matyakubova Noila Shakirjanovna* – PhD student of Tashkent State University of Uzbek Language and Literature named after Alisher Navo'i.
E-mail: matyakubovanoila@navoiy-uni.uz, *nailya89mm@mail.ru*
ORCID: 0009-0009-3154-723X

Myanmar-English Machine Translation", 13th International Conference on Computer Applications, 2015.

R. Sennrich, Volk, Martin (2011). "Iterative, MT-based sentence alignment of parallel texts." In: NODALIDA 2011, Nordic Conference of Computational Linguistics, Riga, 11 May 2011 - 13 May 2011.

https://www.masterclass.com/articles/what-is-a-verb-phrase

ISSN 2181-922X 38-48

DIALEKTAL KORPUSLARNING UMUMIY TAVSIFI: TAJRIBA VA TAHLIL

Ruxsora Muftillayeva[1]

Annotatsiya. Tildagi dialektal o‘zgarishlar - bu inson nutqining xilma-xilligini yoritib beradigan qiziqarli tadqiqot sohasi. Ushbu abstrakt dialektal shevalarning umumiy tavsifini taqdim etadi, bu til hodisalarini o‘rganish uchun eksperimental usullar va analitik yondashuvlarga e’tibor beradi. Dialektal shevalarni tekshirish orqali tadqiqotchilar dialektlarning tarixiy evolyutsiyasi va geografik taqsimoti haqida qimmatli fikrlarni ochib berishlari mumkin. Ushbu abstrakt dialektal shevalarni o‘rganishda keng qo‘llaniladigan eksperimental texnikalar, ma’lumotlarni yig‘ish usullari va statistik tahlillar haqida umumiy ma’lumot beradi. Bundan tashqari u tilshunoslikdagi dialektal o‘zgaruvchanlikni tushunishning muhimligini va uning tilni saqlash, madaniy merosga ta’sirini ta’kidlaydi.

Kalit so‘zlar: *dialektal korpus, lingvokulturologik, intervyu, sintaktik, morfologik.*

Kirish

Dialektal korpus deganda, ma’lum dialektlardan yoki tilning mintaqaviy turlaridan lingvistik ma’lumotlarni to‘playdigan matnlar yoki yozuvlar to‘plami tushuniladi. Ushbu korpuslar ko‘pincha keng ko‘lamli ma’lumotlar to‘plash harakatlari natijasida, ma’lum bir dialektdagi ona tilida so‘zlashuvchilar bilan o‘zaro aloqalarni o‘z ichiga olgan holda yaratiladi. Dialektal korpus bilan ishlash tajribasi ma’lum bir mintaqaning til boyligini o‘rganishni o‘z ichiga oladi. Ushbu korpuslar bilan ishlaydigan tilshunoslar va tadqiqotchilar bir dialektni boshqasidan ajratib turadigan o‘ziga xos xususiyatlar, lug‘at, grammatika, talaffuz va boshqa jihatlar haqida tushunchaga ega bo‘ladi. Dialektal korpusni tahlil qilish orqali tadqiqotchilar dialektal o‘zgaruvchanlik, vaqt o‘tishi bilan til o‘zgarishi, til aloqasi va tildan foydalanishga ta’sir qiluvchi ijtimoiy omillar kabi lingvistik hodisalarni o‘rganishlari mumkin. Ushbu tadqiqot tilning rivojlanishi, madaniy o‘ziga xoslik til va jamiyat o‘rtasidagi munosabatlarni tushu-

[1]*Muftillayeva Ruxsora Toshmuhammad qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi mutaxassisligi magistranti
E-pochta: *ruxsoramuftillayeva851@gmail.com*
ORCID: 0009-0007-3046-3953

nishimizga yordam beradi. Koʻpgina Yevropa tillarining dialekt korpuslari hozirda mavjud boʻlib, ular odatda ma'lum bir mamlakatning turli mintaqalaridagi materiallarni oʻz ichiga oladi. Portugaliya korpusi –The syntax-oriented Corpus of Portuguese dialects -Portugal lahjalarining sintaksisiga yoʻnaltirilgan korpus (COSYPOR) turli portugal lahjalari sintaksisiga qaratilgan lingvistik manba. U portugal tilida soʻzlashadigan turli mintaqalardan sintaktik ma'lumotlarning keng qamrovli toʻplamini taqdim etishga qaratilgan boʻlib, tadqiqotchilarga dialektlardagi sintaktik tuzilmalarni oʻrganish va solishtirish imkonini beradi.

Asosiy qism

Oʻzbek shevalari bazasining milliy korpusini yaratishda hali tajribamiz yoʻqligini inobatga olsak, bu jarayonda tilning tabiatidan qat'iy nazar dunyo tilshunos olimlari bilan birgalikda yaratilgan korpuslarni koʻzdan kechirishga toʻgʻri keladi [Xolova, 2022. 31].

COSYPOR yozma va ogʻzaki matnlarni, shuningdek, portugal lahjalarida soʻzlashuvchilardan toʻplangan audio yozuvlarni oʻz ichiga oladi. Korpus turli mavzular va janrlarni, jumladan suhbatlar, intervyular, hikoyalar va boshqalarni qamrab oladi. Tilshunoslik tadqiqoti va tahlilini osonlashtirish uchun matnlar nutq qismi teglari va sintaktik tahlil kabi lingvistik ma'lumotlar bilan izohlanadi. Korpus tilshunoslar, tadqiqotchilar va portugal lahjalari sintaksisini oʻrganishga qiziqqan talabalar tomonidan foydalanish uchun moʻljallangan. U sintaktik oʻzgarishlarni oʻrganish va turli dialektlarning oʻxshash va farqlarini tushunish uchun qimmatli manba boʻlib xizmat qiladi. COSYPOR ma'lumotlarini tahlil qilish orqali tadqiqotchilar portugal lahjalarining sintaktik tuzilmalari haqida tushunchaga ega boʻlishlari va tilshunoslik sohasiga hissa qoʻshishlari mumkin. Resurs turli mazmundagi korpuslarga (asosan, Portugal va Braziliya gazetalari matnlari, shuningdek, portugal fantastika asarlari toʻplami, Braziliya elektron pochta xabarlari va boshqalar) kirish imkonini beradi. Umuman olganda, toʻplamning aksariyat qismini portugal tilining brazil tilidagi versiyasini aks ettiruvchi matnlar egallaydi. Deyarli barcha matnlar morfologik belgilar bilan ta'minlangan. Korpusning umumiy hajmi 70,8 million soʻzdan foydalanishni tashkil etadi. Ulardan 69,8 millioni morfologik, mantiqiy izohlangan. Korpusga kirish bepul. Umuman olganda, Portugal lahjalarining sintaksisiga yoʻnaltirilgan korpusi portugal lahjalari sintaksisini oʻrganish uchun qimmatli vosita boʻlib, tadqiqotchilarga portugal tilida soʻzlashadigan turli mintaqalardagi sintaktik oʻzgarishlarni oʻrganish va tahlil qilish imkonini beradi.

Corpus Oral y Sonoro del Español Rural- Ispaniya korpusi umumiy nom bilan COSER deb nomlanuvchi korpus dialektal korpus bo'lsa ham, lekin u an'anaviy dialektologiyaga qiziqish obyekti bo'lgan ma'lumot beruvchilarning nutqi bilan cheklangan: qishloqda yashovchi aholi asosan yoshi katta, maktab ma'lumoti o'rtacha va ular suhbatlashgan hududda tug'ilgan. Bugungi kunda (2022-yil dekabr) ma'lumotlariga ko'ra 2961 nafar ma'lumot beruvchi ro'yxatga olingan.

1-jadval. COSER korpusining qatnashuvchilari haqidagi ma'lumotlar

Axborot beruvchilar	**Soni**	**O'rtacha yoshi**
Erkaklar:	1.415 (47,8%)	75 yosh
Ayollar:	1.546 (52,1%)	73,6 yosh
Jami:	2.961	74,2 yosh

COSERni tashkil etuvchi yozuvlar 1990-yildan 2022-yil dekabrigacha bir qator so'rov kompaniyalaridan olingan. Ushbu korpus ishi bir nechta tadqiqot loyihalari ko'magida va "Dialectologia Hispanica", "El español hablado. Variantes peninsulares", Madrid avtonom universitetida ispan filologiyasi bakalavriatiga tegishli ixtiyoriy fanlar (Universidad Autónoma de Madrid, UAM). 2011-yildan hozirgi kunga qadar ular ushbu universitetda ispan tilini o'rganish darajasining "Lengua española. Variedades de la lengua" (3-kurs) fanining ixtiyoriy faoliyati sifatida birlashtirilgan.

2-jadval. COSER korpusining audio yozuvlari haqida m'lumot

Yozib olingan hududlar	Viloyat va orollar	Yozib olingan ma'lumotlarning umumiy hajmi	Har bir intervyu uchun o'rtacha yozib olish	Suhbatlar soni	Matn va audio shaklidagi intervyular (may 2022)
1,415	55	1,910 soat	1soat, 4 min.	1,772	218

2022-yilgacha suhbatlar Pireney yarim oroli va ikkita arxipelagdagi 55 provinsiya yoki orolga tegishli 1415 ta qishloq aholi punktlarida o'tkazilgan. Ularning geografik joylashuvi xaritada ko'rsatilgan, bu yerda ularni viloyat va hududni alifbo tartibida umumlashtiruvchi raqamli kod yordamida aniqlash mumkin (masalan, Alava provinsiyasidagi Berganzo shahrida 0101 kodi mavjud). Ovoz materiallari Pireney yarim orolining katta qismini qamrab oladi. Umuman olganda, COSER hozirda 1910 soatlik yozuvlarga ega. Ularning aksariyati analog formatda yozilgan. Materiallarning yarmi

turli xil ilmiy loyihalar tomonidan qo‘lga kiritilgan va o‘zlarining akademik kurs ishlarining bir qismi sifatida o‘zlari to‘plagan yozuvlarni transkripsiya qilgan UAM bakalavriat talabalarining bir necha avlodlari ishtiroki tufayli, turli xarakterdagi va aniqlikdagi transkripsiyalarga ega. Korpus 2015-yilda BConcord muharriri bilan qayta ko‘rib chiqilgan va standartlashtirilgan 141 ta hududga (taxminan 183 soat) mos keladigan 147 ta transkripsiya ushbu veb-saytda nashr etildi va qidiruv tizimi orqali qidirish mumkin bo‘ldi. 2017-yildan beri korpusga oddiy qidiruv va kengaytirilgan qidiruv rejimlarida ham kirish mumkin (bu lemmalar va morfosintaktik teglar bo‘yicha so‘rovlar o‘tkazish imkonini beradi). 2020-yilda ushbu so‘rovda geografik koordinatalar va joylarning pochta indeksi yoqilgan bo‘lib, ma’lumotlar geografik axborot tizimlarida tahlil qilinishi mumkin va matnni audio bilan sinxronlashtirish tugallangan.

3- jadval. COSER korpusining transkripsiya qilingan so‘zlar haqida ma’lumot

Transkripti bor hududlar	Viloyatlar va orollar	Transkripsiya qilingan soatlar	Umumiy transkripsiya qilingan so‘zlar	Umumiy birliklar (tokenlar)
218	55	295 soat, 48 minut	3,596,205 so‘zlar	4,591,828

Deutsches Referenzkorpus (DeReKo) nemis tilidagi yozma matnlarning katta korpusidir. DeReKo dunyodagi eng yirik korpuslardan biri bo‘lib, 24 milliarddan ortiq so‘zdan iborat. U turli janr va sohalardagi yozma matnlarning keng doirasini o‘z ichiga oladi. Korpus 20-asrdan to hozirgi kungacha bo‘lgan turli davrlardagi matnlarni qamrab oladi. U gazetalar, jurnallar, kitoblar, veb-saytlar va boshqa manbalardan olingan matnlardan tashkil topgan. DeReKo matnlar uchun batafsil lingvistik izohlarni taqdim etadi, jumladan nutq qismlarini teglash, lemmatizatsiya va sintaktik tahlil qilish. Bu uni lingvistik tadqiqotlar va til tahlili uchun qimmatli manbaga aylantiradi. DeReKoga kirish Germaniyaning Mannheim shahridagi Nemis tili instituti (Institut für Deutsche Sprache, IDS) orqali mavjud. Tadqiqotchilar korpusga kirish uchun ariza topshirishlari va undan tadqiqot loyihalari uchun foydalanishlari mumkin.

FRED - ingliz dialektlarining FREIBURG korpusi FRED - bu bir tilli og‘zaki dialekt korpusi bo‘lib, u Angliya, Shotlandiya, shuningdek (to‘liq versiyada) Uels, Hebridlar va Men orolidagi ona tilida

so‘zlashuvchilar bilan to‘liq metrajli audio suhbatlardan tashkil topgan. Korpus audio yozuvlardan (wav formatda aksariyat intervyular to‘liq formatda berilgan) va orfografik transkriptlardan (txt fayllari) iborat.

4- jadval. FRED korpusining umumiy tuzilishi haqida ma’lumot

Loyiha rahbari	Tuzilish vaqti:	Hajmi:	Til:	Matnlar soni:	Davr:	Holati:
Prof. Dr. Bernd Kortmann	2000–2005 (loyiha guruhi “Tipologik nuqtai nazardan ingliz shevasi sintaksisi”)	1 011 396 so‘z (to‘liq versiya 2 496 763, intervyu oluvchining so‘zlaridan tashqari)	Ingliz (Britaniya, Shotlandiya, [Uels, Manx] navlari)	121 intervyu/ transkript	1970–1999	2005- yilda tugallangan

FRED korpusi yana quyidagilarni o‘z ichiga oladi: 1 011 396 ta o‘zgaruvchi so‘z, 123 soat yozib olingan nutq, 121 ta intervyu, 144 ta dialekta so‘zlashuvchi 57 ta turli joylarda, 18 ta okrugda, 5 ta asosiy dialekt hududida yozib olingan.

5-jadval. Yozuv sanasi bo‘yicha matn taqsimoti, FRED-S

Qayd etilgan sana	Matnlar soni	Matn materialining %
1970-1979 yillar	47	42,2%
1980-1989 yillar	56	43,9%
1990-1999 yillar	15	10,3%
Noma’lum	3	3,6 %
Jami	121	100,0%

6- jadval. Matnni dialekt maydoni bo‘yicha taqsimlash, FRED-S.

Dialekt maydoni	Matnlar soni	O‘zgaradigan so‘zlar	Matn materialining %
Janubiy-g‘arbiy (SW)	38	264 863	26,2%
Janubiy-sharqiy (SE)	17	260 643	25,8%
Midlans (o‘rta)	16	152 535	15,1 %
Shimoliy (N)	30	266 955	26,4%
Shotlandiya pasttekisligi (ScL)	20	66 400	6,6%
Jami	121	1 011 396	100,0 %

Har bir matndan oldin matn sarlavhasi mavjud bo‘lib, unda matn identifikatori, shuningdek dialekt hududi, okrug va suhbat bo‘lib, o‘tgan joy (ya’ni, ma’ruzachi qayerdan kelgan), mavjud sotsiolingvistik ma’lumotlar va suhbat sanasi haqidagi ma’lumotlar mavjud.

Og‘zaki tarix bo‘yicha suhbatning o‘ziga xos xususiyatlaridan biri shevalarni o‘zining eng asl ko‘rinishida o‘rganishning muhim sharti bo‘lib, so‘zlovchilar o‘z umrlarining ko‘p qismini ma’lum bir geografik hududda o‘tkazgan bo‘lishi va bu hudud bilan mustahkam aloqada bo‘lishi kerak. Ma’ruzachilarning aksariyati birinchi jahon urushidan oldin tug‘ilgan va suhbat vaqtida 60 va undan katta yoshda bo‘lgan (eng keksa ma’ruzachi 1877-yilda tug‘ilgan; barcha ma’ruzachilarning deyarli 90 foizi 1920-yilgacha tug‘ilgan). FRED-S dagi matnli materialning qariyb 70% 60+ yosh guruhi tomonidan yozib olingan intervyulardir.

7-jadvalda. Intervyu beruvchilar va yosh guruhlari bo‘yicha ma’lumot berilgan

Yosh guruhlari	Ma’ruzachilar soni	O‘zgaradigan so‘zlar	Matn materialining %
0-44 yosh	4	12 287	1,2 %
45-59 yosh	5	40, 258	4,0 %
60+ yosh	71	726,134	71,8 %
Yoshi noma’lum	64	232 558	23,0 %

8-jadval. Intervyu beruvchilarning tug‘ilgan yili bo‘yicha matn taqsimoti

Tug‘ilgan yili	Ma’ruzachilar soni	O‘zgaradigan so‘zlar	Matn materialining %
1870-1879-yillar	1	6899	0,7 %
1880-1889-yillar	5	89 615	8,9%
1890-1899-yillar	28	260 909	25,8 %
1900-1909-yillar	30	281 068	27,8 %
1910-1919-yillar	21	176 507	17,5 %
1920-1929-yillar	7	74 365	7,4 %
1930-1939-yillar	3	23 494	2,3 %
1940-1949-yillar	1	1684	0, 2 %
Noma’lum	48	96 696	9,5 %

Ma’ruzachilarning uchdan ikki qismidan ko‘prog‘i NORM deb ataladi, ya’ni mobil bo‘lmagan keksa qishloq erkaklari bo‘lib, ular odatda o‘n to‘rt yoki undan kichik yoshda maktabni tark etishgan. Intervyu beruvchilarning jami 87 erkak va 52 ayol ma’lumot

beruvchidan iboratdir. Korpusda ayol va erkak ismlari va taxalluslari, bosh harflari va familiyalari uchun turli teglar ishlatilgan (obyektlar, kompaniyalar, brendlar va boshqalar nomlari o‘zgarishsiz qoladi). Nashr etilgan va davom etayotgan tadqiqotlar, jumladan, magistrlik va doktorlik dissertatsiyalari haqidagi dolzarb ma’lumotni loyiha veb-saytidan topishingiz mumkin: http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/.

Nordic Dialect corpus (NDC) – korpus Daniya, Islandiya, Norvegiya, Farer va Shvetsiya kabi davlatlarning og‘zaki so‘zlashuv tillaridan tashkil topgan. U shimoliy german tillarining barcha shimoliy mamlakatlaridagi dialektlarning spontan nutq ma’lumotlaridan iborat. Korpusdagi lingvistik ma’lumotlar turli manbalardan olingan. Korpusda dialektlarda so‘zlashuvchilarning suhbatlari va intervyularidan olingan 2,75 milliondan ortiq so‘z mavjud. U transkripsiya qilingan audio va videoga bog‘langan, xarita funksiyasiga ega va uni turli yo‘llar bilan qidirish mumkin. Korpusning maqsadi shimoliy sintaksis tadqiqoti bo‘lsa ham, korpus umumiy, Norvegiya dialekti korpusi, shved dialekti korpusi va boshqalar bo‘lib, fonologiya, morfologiya va leksikografiya kabi keng ko‘lamli tadqiqot sohalarida qo‘llaniladi. Ma’lumotlar bazasi turli sintaktik hodisalarni aks ettiruvchi jumlalar ro‘yxatiga 207 joydan 924 ta shimoliy lahjada so‘zlashuvchilarning nutqlaridan iborat. Ko‘pgina ma’ruzachilar ma’lumotlar bazasida ham, korpusda ham bir xil. Korpusda joy, yosh, ma’lumot beruvchilarning jinsi yoki sintaktik hodisa turiga qarab saralangan. Nordic dialektal korpusi yaratilgandan boshlab hozirgi vaqtga qadar takomillashtirilib borilmoqda, xususan korpusga bir qancha yangi funksiyalar kiritilgan.

Nordic dialect Corpus v. 4.0: Faqat 1998-2015-yillardagi dialekt yozuvlari va transkripsiyalari. (2019-yil sentyabr)

Yangi qidiruv interfeysida Nordic dialect Corpus uchun foydalanuvchi qo‘llanmasi (2019-yil iyun)

Nordic dialect Corpus v. 3.0: kengaytirilgan Islandiya va Shvetsiya qismi. Islandiyadan 16 nafar, Shvetsiya va Swedia 2000dan 24 nafar yangi informator (2017-yil sentyabr)

Nordic Dialect Corpus v. 2.0 va Nordic Syntax Database uchun yangi qidiruv interfeyslari. (2017)

Nordic Atlas of Language Structures (NALS) jurnali chop etildi (2014)

Nordic dialekt korpusi: kengaytirilgan Islandiya qismi - 6 joydan 20 ta yangi ma’lumot beruvchi (2013)

Nordic Dialect Corpus v. 4.0 uchun new qidiruv interfeyslari (2023-yil noyabr)

9- jadval. Norvegiya sheva korpusi jamlanmasi

Mamlakat	Ma'lumot beruvchilar soni	Hududlar	Identifikatsion belgilar
Daniya	81	15	220, 360
Farer	20	5	64,803
Islandiya	48	8	94,338
Norvegiya	438	111	1,997,920
Shvetsiya (jumladan ovdalian)	150	44	376,868

Ushbu korpus sistemasiga 3 ta kirish tizimi mavjud: 1. EDUGAIN. 2. CLARIN (Skandinaviya shevalar korpusi). 3. FEIDE (Norveg millatiga mansub aholiga foydalanish imkonini beruvchi). Korpusdagi ba'zi yozuvlar ikki tomonlama transkripsiya - orfografiya yoki transkripsiya fonetik yig'ilmalar jamlanmasiga ega. Transkripsiyalarni eshitish yoki ko'rish mumkin bo'lgan audio va video fayl taqdimotlari tayyorlangan.

Helsenki korpusi - Britaniya ingliz dialektlari korpusi bo'lib, asosan, Sharqiy Angliya va Janubiy-G'arbiy hududlardan, Lankashirdan kichik to'plamga ega bo'lgan orfografik transkripsiyalangan audio yozuvlar to'plamidir. Yozuvlar 1970 va 1980-yillarda Finlyandiya aspirantlari tomonidan yig'ilgan. Dialektologik korpusning maqsadi nafaqat dialektologiya, balki sotsiolingvistika, nutq tahlili, morfologiya, sintaksis va fonologiya sohalarida lingvistik tadqiqotlar uchun material taqdim etishdan iborat. Korpus shuningdek, aloqa etnografiyasi, mahalliy odatlar va tarix kabi tilga oid bo'lmagan, ko'p tarmoqli tadqiqotlar uchun material beradi.

10- jadval. Helsenki korpusi haqida ma'lumot

Loyiha rahbarlari	Hajmi	Vaqt davrlari	Til
Ossi Ihalainen Kristi Peitsara Anna-Liisa Vasko	**1.008.641** so'zdan iborat jami **187** ta fayl.	1970-1980-yillar	Ingliz qishloq (Kembridgeshire, Devon, Ile of Ely, Somerset, Suffolk) va shahar (Esseks, Lankashir).

Helsenki korpusining birinchi bosqichi 2006-yilda yakunlangan, ikkinchi bosqichi esa hali ham davom etmoqda. Korpusning asosiy materiallari og'zaki dialekt nutqining audio yozuvlardan iborat. Helsenki korpusi Finlandiya madaniyat jamg'armasi, Finlandiya akademiyasi, Helsenki universiteti tomonidan moliyalashtirilgan.

11- jadval. Korpusga kiritilgan geografik mintaqalar haqida ma'lumot

Viloyat (qishloq)	Qishloqlar soni	Ma'lumot beruvchilar soni
Kembridjeshire	26	38 (+6)
Devon	9	33 (+8)
Ely oroli	20	52 (+6)
Somerset	15	24 (+5)
Suffolk	16	47 (+4)
Hudud (shahar)		
Essex	3	6 (+1)
Lancashir	3	6 (+1)
Jami	92	206 (+31)

Korpusda asosan ma'lumot beruvchilar sifatida erkaklarning ulushi ko'proq. Korpusda ayollar nutqi taxminan beshdan bir qismni tashkil etadi. 1970 va 1980-yillardagi suhbat chog'ida qishloq xabarchilarining aksariyati nafaqaga chiqqan va yoshi 70 dan oshgan yoki undan katta bo'lgan, shuning uchun XX asrning birinchi o'n yilligida ta'lim olgan. 1980-yillarning oxirida to'plangan Essex va Lancashire shahar korpuslari yuqoridagilardan uch avlod ikki jinsli namunalarni berishda farq qiladi, ma'lumot beruvchilar 19 yoshdan 70 yoshgacha. Fayllardan foydalanishga ruxsat olish uchun Anna-Liisa Vasko (anna-liisa.vasko@helsinki.fi) yoki Kirsti Peitsara (kirsti.peitsara@helsinki.fi) bilan bog'lanish kerak.

Ma'lum bir hududning shevasi asosida yaratilgan dialektal korpuslar ham bor. Masalan: Kuban dialektal korpusi shunday korpuslardan hisoblanadi. Kuban dialektal korpusi 18 yil davomida Kubanning g'arbiy qishloqlari an'anaviy madaniyatini o'rganish asosida yaratilgan. Bu loyiha Rossiya gumanitar jamg'armasi, Krasnodar o'lkasi ma'muriyati tomonidan qo'llab-quvvatlangan. 2014-yilda Rossiya gumanitar fondi va Krasnodar o'lkasi ma'muriyati tomonidan "Kuban dialekt madaniyatining elektron korpusini yaratish" loyihasi tasdiqlandi. 2016-yilda Rossiya gumanitar fondidan (2016-yildan buyon u Rossiya fundamental tadqiqotlar fondi tarkibiga kiradi) ekspeditsiya ishlari uchun grant olinadi. Korpus tarkibiga "marosim madaniyati", "ma'naviy madaniyat", "hunarmandchilik madaniyati", "kundalik madaniyat", "xalq hunarmandchiligi va san'ati", "oilaviy turmush tarzi" kabi kichik korpuslar kiradi. Hozirgi vaqtda "Ritual madaniyat" (to'y marosimi ma'ruzasi) va "ma'naviy madaniyat" ("mifologiya", "xalq pravoslavligi" ma'ruzalari) kabi korpuslar mazmunini shakllantirish bo'yicha ishlar olib borilmoqda. 2015-yilda loyi-

hani moliyalashtirish davom etdi. Korpusga turli subkorporalarning nutqlari materiallari bo‘yicha lingvokulturologik tahlil o‘tkazildi. Korpus uchun yig‘ilgan materiallarning keyinchalik lenta va video kassetalar, elektron vositalar, yozilgan dialekt nutqining raqamli versiyalari yaratildi. Turli tematik nutqlarni tahlil qilish jarayonida ularning tarkibiy va semantik birligini tartibga soluvchi mikro-mavzular, tushunchalar to‘plamini aniqlash mumkin bo‘ldi.

Xulosa

Xulosa qilib aytadigan bo‘lsak, dialektal korpuslarni o‘rganish va tahlil qilish turli dialektlarning lingvistik xususiyatlari haqida qimmatli ma’lumotlarni beradi. Dialekt matnlarining katta to‘plamlarini to‘plash va tekshirish orqali tadqiqotchilar bir dialektni boshqasidan ajratib turuvchi lingvistik xususiyatlarni chuqurroq tushunishlari mumkin. Dialektal korpus bilan ishlash tajribasi turli qiyinchiliklar va mulohazalarni o‘z ichiga oladi. Birinchidan, dialektal matnlarni to‘plash va tuzish ko‘p vaqt va mehnat talab qiladigan jarayon. Tadqiqotchilar turli xil manbalardan matnlarni aniqlashlari va to‘plashlari kerak, ular dialektal o‘zgarishlarning keng doirasini ifodalaydi. Korpus qurilgandan so‘ng, tadqiqotchilar turli lingvistik vositalar va usullardan foydalangan holda ma’lumotlarni tahlil qilishadi. Bu tahlil muayyan shevaga xos bo‘lgan fonetik, fonologik, morfologik, sintaktik va leksik xususiyatlarni aniqlashni o‘z ichiga olishi mumkin. Bundan tashqari, tadqiqotchilar lahjani shakllantirgan tildan foydalanish qonuniyatlarini, sotsiolingvistik omillarni va tarixiy ta’sirlarni o‘rganadi. Bunday tadqiqot natijalari tilshunoslik, sotsiolingvistika, antropologiya va tilni saqlash kabi turli sohalarga ta’sir qiladi.

Foydalanilgan adabiyotlar

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.

http://rusling.narod.ru/qqq_corp_nonslav_other.htm

http://www.corpusrural.es

https://freidok.uni-freiburg.de/proj/

Nordic Dialect Corpus (uio.no)

https://www.helsinki.fi/varieng/CoRD/corpora/Dialects/fieldwork

Региональная этнолингвистика (ethnolex.ru)

GENERAL DESCRIPTION OF DIALECTAL CORPSES: EXPERIMENT AND ANALYSIS

Ruxsora Muftillayeva[1]

Abstract. Dialectal variations in language are a fascinating area of study that sheds light on the diversity of human communication. This abstract presents a general description of dialectal corpses, focusing on the experimental methods and analytical approaches used to explore these linguistic phenomena. By examining dialectal corpses, researchers can uncover valuable insights into the historical evolution and geographical distribution of dialects. This abstract provides an overview of the experimental techniques, data collection methods, and statistical analyses commonly employed in studying dialectal corpses. Furthermore, it highlights the importance of understanding dialectal variation in linguistics and its implications for language preservation and cultural heritage.

Key words: *dialectal corpus, language culture logic, interview, syntax, morphology.*

References

Xolova M. o'zbek milliy shevalari korpusini tuzishning lingvistik asoslari (Boysun tumani "j"lovchi shevalari misolida): Fil. fan. bo'yicha falsafa doktori (PhD)...diss. Avtoref. – Termiz, 2022. http://rusling.narod.ru/qqq_corp_nonslav_other.htm

http://www.corpusrural.es

https://freidok.uni-freiburg.de/proj/

Nordic Dialect Corpus (uio.no)

https://www.helsinki.fi/varieng/CoRD/corpora/Dialects/fieldwork

Региональная этнолингвистика (ethnolex.ru)

[1]*Muftillayeva Ruxsora Toshmuhammad qizi* –Master of degree, Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: *ruxsoramuftillayeva851@gmail.com*
ORCID: 0009-0007-3046-3953

ISSN 2181-922X 49-62

JAHON TILSHUNOSLIGIDA TABIIY TILNI MODELLASHTIRISH NAZARIYASI VA AMALIYOTI

Sabura Xudayarova[1]

Annotatsiya. Texnika paydo bo‘lganidan beri bir necha o‘n yillar o‘tdi, bu esa kompyuter texnologiyalari shakllanishi va tez rivojlanishini hayotga olib keldi. Tilshunoslikning kompyuter kabi noan’anaviy sohasining rivojlanishi lingvistik bilim va jarayonlarni modellashtirish zaruriyatini keltirib chiqardi. Ushbu maqolada jahon tilshunosligida tabiiy tilni qayta ishlash hamda lingvistik modellashtirish masalalarida olib borilgan tadqiqotlarning farqli va o‘xshash jihatlarini tahlil qilish orqali to‘xtalib o‘tilgan.

Kalit so‘zlar: *Model, modellashtirish, deduktivlik, anilitik, sintetik, meta-til, N-gramm, ELM dasturi, Open AT, Neyron.*

Tabiiy til va uning turli xil modellari mavjud. Til faoliyatining har xil turlarida ishlash esa, ko‘pincha ushbu modellar o‘rtasidagi munosabatlar noaniq bo‘lib qolishiga olib keldi. Shuning uchun modellashtirishdagi umumiy xususiyatlarni o‘rganish ehtiyoji tug‘ildi. Bu modellarni solishtirish imkonini beradigan lingvofalsafiy asosdir. Ularning o‘zaro asosiy o‘xshashliklari va farqlarini aniqlab berish vazifasi dolzarb masalalardan biridir, chunki bu sohada o‘xshashlik va farqlar mavjud.

Tilshunoslikda model - bu tilshunos tomonidan haqiqiy yoki sun’iy ravishda yaratilgan. Uning xatti-harakatlarini takrorlaydigan, taqlid qiladigan aqliy qurilma (odatda soddalashtirilgan shakl) lingvistik maqsadlar uchun asl nusxaning xatti-harakatidir.

Tilni modellashtirish texnik modellashtirishdan farqli ravishda, o‘ziga xos xususiyatlarga ega. L.N.Murzinning fikricha, obyekti bevosita kuzatuvga berilmagan har qanday fan obyektni modellashtirishga majbur bo‘ladi. Modellashtirish umumilmiy metod bo‘lib, deduktivlik, tafakkur eksperimentidan foydalanish va modelni ideal obyekt sifatida talqin qilish kabilar bilan xarakterlanadi.

[1]*Xudayarova Sabura Shuxrat qizi* -Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti, Kompyuter lingvistikasi yo‘nalishi 2-kurs magistranti.
E-pochta: *xudayorovasabura@navoiy-uni.uz*
ORCID: 0000-0001-7102-0824

Lingvistik modellashtirish berilgan semiotik elementlar o‘rtasidagi munosabatlar tizimi sifatidagi til bilan yoki verbal kommunikatsiya jarayoni sifatidagi til bilan ish ko‟radi.

Demak, modellashtirish struktur tilshunoslik, psixolingvistika, dinamik va kommunikativ lingvistikalarning obyekti sifatida namoyon bo‘ladi. Til modeli haqida fikr yuritar ekan, I.I.Revzin shunday yozadi: “Model quyidagicha quriladi: muayyan fan tomonidan to‘plangan xilma-xil tushunchalar yig‘indisidan birlamchi deb qabul qilinuvchi ayrimlari tanlab olinadi. Birlamchilar o‘rtasidagi munosabatlar aniqlanadi va ular pastulat sifatida qabul qilinadi. Qolgan barcha da’volar qat’iy deduktiv ravishda terminlarda namoyon bo‘ladi. Ular birlamchi tushunchalar orqali aniqlanadi. Shu ma’noda model sistema sifatidagi tilning qismi emas, balqi qandaydir konstrukt, gipotetik ilmiy qurilma sifatida qabul qilinadi” [Tulupova, Pavlenko].

Lingvistik model tushunchasi tarkibiy tilshunoslikda paydo bo‘lgan. K.L.Buler, Z.Z.Harris, Ch.Hokket, ammo XX asrning 60-70-yillarida ilmiy foydalanishga kiradi.

Matematik tilshunoslikning paydo bo‘lishi va g‘oyalarning tilshunoslikka kirib borishi, turli kibernetika usullari shakllanib bordi. Maqolada mualliflar tamonidan modellashtirish haqida fikr yuritilar ekan uning quydagi turlari e’tirof etiladi.

“Tilni bilishning qaysi tomoni mavzu ekanligiga qarab modellashtirish, nutq faoliyati modellari til modellariga bo‘linadi to‘g‘ri va to‘g‘ri farqlash qobiliyatini taqlid qiluvchi grammatik to‘g‘rilik tildagi noto‘g‘ri va funktsional, o‘zaro bog‘liqlik qobiliyatini taqlid qilish nutqning mazmuni (tarkib rejasi) uning shakli (ifoda rejasi) bilan.

Modelning “kirish” va “chiqish” dagi ma’lumot turiga qarab grammatik to‘g‘rilik tan oluvchi va generativga bo‘linadi.

Tan olish modeli: Masalan, K. Aydukevichning “kategorik grammatikasiga kirish” da tabiiy tilda yoki uning ichida matnning bir qismini oladi sun’iy tilda mavhum vakillik va “chiqish”da javob beradi, berilgan segment grammatik jihatdan to‘g‘ri yoki g‘ayritabiiy bo‘ladimi.

Generativ model: Masalan, N. Xomskiyning “Generativ grammatika” nasaziyasiga ko‘ra bu taniydiganga teskari. Birinchisini tanqidiy yengish orqali Xomskiyning “Generativ grammatika” versiyasi modelni yaratishga olib keldi. Generativ semantika (J. Lakoff) modellar bilan ko‘p o‘xshashliklarga ega gapirish yoki sintez qilish. Nutq faoliyatining qaysi jihati modellashtirilganiga

qarab – tinglash yoki gapirish-funksional modellar quyidagilarga bo‘linadi:

1. Analitik
2. Sintetik.

Ba’zi tillarning to‘liq analitik modeli kirish fikrlardan tushuniladiki, bunda matnning bir qismini oladi. Odatda, bayonotdan kam emas maxsus semantik meta-til (ya’ni uning talqini). To‘liq sintetik to‘liq analitik bilan teskari bo‘lgan ba’zi tillarning modelidir.

Yuqoridagi fikirlardan tushuniladiki, modellashtirish tili - bu izchil qoidalar to‘plami bilan belgilangan tuzilmadagi ma’lumot, bilim yoki tizimlarni ifodalash uchun ishlatilishi mumkin bo‘lgan har qanday sun’iy til. Qoidalar tarkibiy qismlarning ma’nosini izohlash uchun ishlatiladi. Modellashtirish tili grafik yoki matnli bo‘lishi mumkin. Grafik modellashtirish tillari cheklovlarni ifodalash uchun tushunchalar va belgilarni bog‘laydigan chiziqlarni, shuningdek munosabatlarni va boshqa turli xil grafik izohlarni ifodalovchi nomlangan belgilar bilan diagramma usullaridan foydalanadi. Matnni modellashtirish tillari odatda standartlashtirilgan kalit so‘zlardan, so‘ngra parametrlar bilan kompyuter tomonidan talqin qilinadigan ifodalarni yaratish uchun foydalanadi. Dasturlash, hisoblash tilshunosligining eng nufuzli maktablaridan biri N.Xomskiy yaratgan yo‘nalish - uning lingvo-falsafiy asoslaridir.

Soha doirasidagi bilimlar bu muallif tomonidan ham juda chuqur o‘rganilgan. Ushbu yo‘nalishda uning izdoshlari ham Raqobatbardosh yondashuvlarga asoslanib, daraja yoki tabaqalanish nazariyasi hamda, lingvistik modellar yuzasidan tadqiqotlar olib bordilar.

Tilshunoslikdagi bu ikki asosiy yo‘nalishning o‘ziga xos belgisitik modellashtirish [Шаляпина, 2007]. Ushbu monografiyada yuqoridagi tadqiqotlar yuzasidan fikr yuritilar ekan, muallif monografiya tadqiqotlarning nazariy jihatlarini ochib berishga qaratilganligini ta’kidlaydi. Til faoliyatining u yoki bu turi doirasida tilni rivojlantirish lingvistik aloqa predmetining harakati sifatida talqin qilinishi mumkin. Qanchalik ikki o‘lchovli va undan ham ko‘proq uch o‘lchovli fazoda bir nuqtadan ikkinchisiga o‘tish uchun turli yo‘llar ishlatilishi mumkin, bu lingvistik modellar doirasidagi asosiy o‘zgaruvchanlikni aks ettiradi

Monografiyada TmS modelining tavsifi yetti bobda tuzilgan. 1-bob umumiy “metodlar-lingvistik voqelikning modeli” bo‘lib, ular ichida to‘plangan lingvistik kompetensiya va lingvistika o‘rtasidagi qarama-qarshilik sifatida o‘z o‘rniga ega bo‘lar edi. Turli lingvistik

vaziyatlarda birining va ikkinchisining modellari orasida harakat qilish. Shu nuqtayi nazardan qaraganda, til kompetensiyasi modeliga asoslanishi kerak, ularning har birida o‘z tilshunoslik to‘plami darajalari. Bu qayta tiklash uchun asosiy imkoniyat yaratadi.

Bunday darajalarning bir o‘lchovli tizimidan ko‘p o‘lchovli va mos keladigan tizimga o‘tish maxsus til kompetensiyasi modelini ko‘p o‘lchovli sifatida qurish – ichida uch o‘lchovli til makonining minimal versiyasi. Chiziqli uch o‘qga nisbatan talqin qilish va tushuntirish imkoniyatlari bir qator umume’tirof etilgan lingvistik tushunchalarning gistik fazosi doirasida fikr yuritiladi.

Morka Pronchai tomonidan 2022-yil 13-dekabrda chop etilgan, “Til modellari va uning turlarini tushuntirish”ga bag‘ishlangan maqolasida, tilning evolyutsiyasi uni hayratda qoldirganligini ta’kidlar ekan, til modeli haqida shunday fikr yuritadi: “Til modeli so‘zlarga ehtimolliklarni belgilash uchun mashinani o‘rganishdan foydalanadi, so‘zlarni bashorat qilishda foydalaniladi. Oldingi yozuvga asoslangan jumladagi keyingi so‘z. Til modellari matndan o‘rganiladi va ular manba matnini yaratish, matndagi keyingi so‘zni bashorat qilish, nutqni aniqlash, optik belgilarni aniqlash va qo‘l yozuvini tanib olish uchun ishlatilishi mumkin”. Muallif mulohazalarini davom etdirib, Mashina qanday qilib jurnalistik maqolaga taqlid qiladigan maqola yaratishi mumkin? Til modeli nima? Shu kabi savollarga javob berishi bilan bir qatorda, til modellarining bir qancha turlari haqida to‘xtalib ularni tushuntirib o‘tadi: “Til modeli amalda, bu so‘zlarning ma’lum bir ketma-ketligi “haqiqiy” bo‘lish ehtimolini beradi. Ushbu kontekstdagi haqiqiylik grammatik aniqlikka ishora qilmaydi. Buning o‘rniga, bu odamlar qanday yozishga o‘xshashligini anglatadi va til modeli shunga o‘rgatilgan. Bu muhim nuqta. Til modelida boshqa mashina o‘rganish modellaridagi kabi sehr yo‘q, ayniqsa chuqur neyron tarmoqlari, bu shunchaki namunadan tashqari kontekstda qayta ishlatilishi mumkin bo‘lgan katta hajmdagi ma’lumotlarni siqilgan shaklga kiritish vositasidir.

Ehtimolli til modeli N-gramm ehtimolliklarni hisoblash yo‘li bilan oddiy ehtimolli til modeli quriladi. N-gramma n so‘zdan iborat ketma-ketlik bo‘lib, bu yerda n noldan katta butun sondir. N-gramm ehtimoli - n-grammning oxirgi so‘zi ma’lum bir n-1 grammdan keyin kelishining shartli ehtimolligi (oxirgi so‘zga e’tibor bermaslik). Bu oxirgi so‘zning n-1 grammdan keyin oxirgi so‘z bo‘lmagan holatlarining nisbati. Bu kontsepsiya Markovning taxminidir. Berilgan n-1 gramm (hozirgi), n-gramm (kelajak) ehtimolliklari n-2, n-3 va hokazo grammlarga (o‘tmish) bog‘liq emas. Bu yondashuvning

aniq kamchiliklari mavjud. Eng muhimi, faqat oldingi n so‘z keyingi so‘zning ehtimollik taqsimotiga ta’sir qiladi. Murakkab matnlar keyingi so‘zni tanlashga hal qiluvchi ta’sir ko‘rsatishi mumkin bo‘lgan chuqur kontekstga ega. Bu o‘rinda olim mavzu doirasidagi bir qancha muammollar va ularning yechimlari borasidagi qarashlarini keltirib o‘tadi.

Keyingi so‘z oldingi n so‘zdan aniq bo‘lmasligi mumkin, garchi n 20 yoki 50 bo‘lsa ham. Bu atama oldingi so‘zni tanlashga ta’sir qiladi: “United” so‘zidan keyin Davlatlar bo‘lsa, ehtimol ancha yuqori. Bundan tashqari, Hajmi (n) oshgani sayin, mumkin bo‘lgan almashtirishlar soni keskin ortadi, garchi ularning aksariyati matnda hech qachon ko‘rinmaydi va barcha hosil bo‘lgan ehtimolliklarni (yoki N-grammlarning barcha raqamlarini) hisoblash va saqlash kerak. Bundan tashqari, yuzaga kelmaydigan N-grammalar kamlik muammosini keltirib chiqaradi, chunki ehtimollik taqsimotining granularligi juda past bo‘lishi mumkin. So‘z ehtimollari bir necha xil ma’noga ega, shuning uchun ko‘pchilik so‘zlar bir xil ehtimolga ega.

Neyral tarmoqqa asoslangan til modellari: Neyron tarmoqqa asoslangan til modellari kirishlarni kodlash yo‘li bilan siyraklik muammosini osonlashtiradi. So‘zni joylashtirish qatlamlari har bir so‘zning semantik munosabatlarini o‘z ichiga olgan ixtiyoriy o‘lchamdagi vektorini yaratadi. Ushbu uzluksiz vektorlar keyingi so‘zning ehtimollik taqsimotida juda zarur bo‘lgan granularlikni yaratadi. Bundan tashqari, til modeli funksiyadir, chunki barcha neyron tarmoqlar juda ko‘p matritsali hisob-kitoblarga ega, shuning uchun keyingi so‘zning ehtimollik taqsimotini ishlab chiqarish uchun barcha n-gramm sonlarini saqlash shart emas. Viktor Lavrenko Til modellarining evolyutsiyasi Neyron tarmoqlar siyraklik muammosini hal qilsa ham, kontekst muammosi saqlanib qolmoqda. Birinchidan, kontekst muammosini yanada samaraliroq hal qilish uchun til modellari ishlab chiqildi - ehtimollik taqsimotiga ta’sir qilish uchun tobora ko‘proq kontekstli so‘zlarni keltirdi. Ikkinchidan, maqsad modelga qaysi kontekstdagi so‘zlar boshqalardan ko‘ra muhimroq ekanligini o‘rganish imkoniyatini beruvchi arxitekturani yaratish edi. Til modeli nima qila oladi? Kontekstdan so‘z ehtimoli haqida xulosa chiqarish uchun zarur bo‘lgan tabiiy tilni mavhum tushunish bir qator muammolarni hal qilish uchun ishlatilishi mumkin. Lemmatizatsiya yoki stemming qaratilgan so‘zni eng asosiy shakliga qisqartirish va shu bilan tokenlar sonini keskin kamaytirish. Agar so‘zning nutq qismi ma’lum bo‘lsa, bu algoritmlar yaxshi ishlaydi. Fe’l postfikslari ot postfikslaridan farq qilishi mumkin, shuning

uchun til modeli uchun umumiy vazifa bo‘lgan nutq qismini teglash (yoki POS teglari) uchun mantiqiy asos. Yaxshi til modeli bilan biz matnlarni ekstraktiv yoki mavhum umumlashtirishni amalga oshirishimiz mumkin. Agar bizda turli tillar uchun modellar mavjud bo‘lsa, mashina tarjimasi tizimini osongina qurish mumkin. Kamroq oddiy foydalanish holatlari savollarga javob berishni o‘z ichiga oladi (kontekstli yoki kontekstsiz)

Til modellari nutqni aniqlash, optik belgilarni aniqlash, qo‘l yozuvini aniqlash va boshqalar uchun ham ishlatilishi mumkin. Imkoniyatlarning butun doirasi mavjud. Men ilgari aytib o‘tgan birinchi model zich (yoki yashirin) qatlam va tepada joylashgan chiqish qatlamidir. Uzluksiz so‘zlar sumkasi (CBOW) Word2Vec modeli. CBOW Word2Vec modeli so‘zni kontekstdan taxmin qilishga o‘rgatilgan. Skip-Gram Word2Vec modeli buning aksini qiladi, so‘zdan kontekstni taxmin qiladi. Amalda, CBOW Word2Vec modeli uni o‘rgatish uchun quyidagi strukturaning ko‘plab misollarini talab qiladi: kirishlar n so‘zdan oldin va/yoki so‘zdan keyin, ya’ni chiqishdir. Ko‘rishimiz mumkinki, kontekst muammosi hali ham o‘zgarmagan.

Taktirilgan neyr tarmoqlari (RNN) Takroriy neyron tarmoqlari (RNN) bu masala bo‘yicha yaxshilanishdir. RNNlar uzoq qisqa muddatli xotira (LSTM) yoki gated recurrent unit (GRU) uyali tarmoq bo‘lishi mumkinligi sababli, keyingi so‘zni tanlashda oldingi barcha so‘zlarni hisobga oladi. AllenNLP ning ELMo dasturi bu tushunchani bir qadam oldinga olib boradi va ikki tomonlama LSTM dan foydalanadi, u so‘zlarni hisoblashdan oldin va keyin kontekstni hisobga oladi.

Transformers rnn-ga asoslangan arxitekturalarning asosiy kamchiligi ularning ketma-ketligidan kelib chiqadi. Natijada, mashg‘ulot vaqtlari uzoq ketma-ketliklar uchun ko‘tariladi, chunki parallellashtirish imkoniyati yo‘q. Bu muammoning yechimi transformator arxitekturasidir. OpenAI va Google BERT ning GPT modellari transformator arxitekturasidan ham foydalanadi. Ushbu modellar, shuningdek, “Diqqat” deb nomlangan mexanizmni qo‘llaydi, uning yordamida model ma’lum holatlarda qaysi kirishlar boshqalarga qaraganda ko‘proq e’tiborga loyiqligini bilib oladi. Model arxitekturasi nuqtayi nazaridan, asosiy kvant sakrashlari birinchi navbatda RNNlar, xususan, LSTM va GRU edi. siyraklik muammosi va disk maydonini kamaytiradigan til modellari, keyinchalik transformator arxitekturasi, parallellashtirishni mumkin va diqqat mexanizmlarini yaratadi. Ammo arxitektura til modeli ustun bo‘lishi

mumkin bo‘lgan yagona jihat emas. GPT-1 arxitekturasi bilan solishtirganda, GPT-3 deyarli hech qanday yangilikka ega emas. Lekin bu juda katta. U 175 milliard parametrga ega va u umumiy sudralib yurishda o‘rgatilgan eng katta korpusda o‘qitilgan. Bu qisman til modelining yarim nazorat ostida o‘qitish strategiyasi tufayli mumkin. Ba’zi so‘zlar tushirib qoldirilgan matndan o‘quv namunasi sifatida foydalanish mumkin. GPT-3ning aqlbovar qilmaydigan kuchi shundaki, u so‘nggi yillarda internetda paydo bo‘lgan barcha matnlarni ko‘proq yoki kamroq o‘qigan va u o‘z ichiga olgan tabiiy tilning murakkabligini aks ettirish qobiliyatiga ega. Ilgari, til modellari standart NLP vazifalari uchun ishlatilgan, masalan, nutq qismini (POS) teglash yoki engil o‘zgartirishlar bilan mashina tarjimasi. Bir oz qayta tayyorlash bilan BERT tabiiy tilning asosiy tuzilishini tushunishning mavhum qobiliyati tufayli POS-tegger bo‘lishi mumkin. T5 bilan NLP vazifalari uchun hech qanday o‘zgartirishga hojat yo‘q. Agar u <M> tokenlari bo‘lgan matnni olsa, u bu tokenlar tegishli so‘zlar bilan to‘ldirish uchun bo‘shliqlar ekanligini biladi. Shuningdek, u savollarga javob berishi mumkin. Agar u savollardan keyin qandaydir kontekstni olsa, u javob uchun kontekstni qidiradi. Aks holda, u o‘z bilimidan javob beradi. Qiziqarli fakt: u trivia viktorinasida o‘z ijodkorlarini mag‘lub etdi.

Yangi boshlanuvchilar uchun NLP. Til modellarining kelajagi bo‘yicha to‘liq qo‘llanma. Ushbu soha sun’iy intellektni yaratishga eng yaqin bo‘lgan sohamiz. AI atrofida juda ko‘p shov-shuvlar mavjud va ko‘plab oddiy qaror tizimlari va deyarli har qanday neyron tarmoq AI deb ataladi, ammo bu asosan marketing. Ta’rifga ko‘ra, sun’iy intellekt mashina tomonidan amalga oshiriladigan insonga o‘xshash aql qobiliyatlarini o‘z ichiga oladi. Transfer o‘rganish kompyuterni ko‘rish sohasida porlaydi va AI tizimi uchun uzatishni o‘rganish tushunchasi juda muhim bo‘lsa-da, bir xil model NLP vazifalarining keng doirasini bajarishi va kiritilgan ma’lumotlardan nima qilish kerakligi haqida xulosa chiqarishi haqiqatdir. Yuqoridagi fikirlarni tahlil qilish natijasida shuni aytish mumkinki, demak, til modeli gapdagi kontekst va atrofdagi so‘zlarga qarab, til model kontekstli yoki kontekstsiz savollarga javoblaringizni o‘zgartirish mumkin. Ular javoblarni turli yo‘llar bilan taqdim etishlari mumkin, masalan, aniq iboralarni olish, javobni tarjima qilish yoki ro‘yxatga olish variantlarini tanlang. *Matnni umumlashtirish*. Til modellari hujjatlar, maqolalar, podkastlar, videolar va boshqalarni avtomatik ravishda eng muhim qismlarga yuklash uchun ishlatish mumkin. Modellar ikki usulda yaratish mumkin: manba matnidan eng muhim ma’lumotlarni

ajratib olish yoki asl tilni takrorlamaydigan yukni taqdim etish. *Hissiyot tahlili*. Tilni modellashtirishni hissiyotlarni tahlil qilish vositalari uchun yaxshi imkoniyatlardir, chunki u matnlarni ovoz ohangini va semantik yo‘nalishini aniqlay oladi.

Suhbatdosh sun’iy intellekt tizimlarining bir qismi sifatida til modellari rivojlanishiga mos matn javoblarini taqdim etishi mumkin. *Mashina tarjimasi*. Mashinani o‘rganishga samarali til modellarining uzoq kontekstlarni samarali umumlashtirish uchun mashina tarjimasini yaxshilash berdi. Matnni so‘zma-so‘z tarjima qilish o‘rniga, til modellari kirish va chiqish ketma-ketliklarining tasvirlarini o‘rganishi va ishonchli tuzatishni berishi mumkin.

Kodni to‘ldirish. Yaqinda keng ko‘lamli til modellari kodni belgilab, tahrirlash va joriyning ta’sirchanligini namoyish qildi. Ular faqat oddiy dasturlash vositalarini bajarishlari mumkin, ko‘pincha kodga tarjima qilishlari yoki xatoliklarni bajarishlari mumkin. Til modellarining turlari til modellari turli xil bo‘lib, ularning ikki darajaga bo‘lishi mumkin: statistik modellar va chuqur neuron tarmoqlarga barqaror modellar. Statistik til modellari

Statistik til modellari so‘zlarning ma’lum ketma-ketligini bashorat qilish uchun ma’lumotlardagi statistik naqshlardan qaror topgan model turidir. Ehtimoliy til modelini olishning asosiy belgilarini N-gramm bilib olishlarini aytishdir.

N-gramma so‘zlar ketma-ketligi bo‘lib, n noldan katta sondir. Tilning oddiy probabilistik modelini kiritish uchun siz matnda turli xil N-grammlar (so‘z birikmalari) paydo bo‘lishi mumkinligini hisoblaysiz. Bu har bir necha marta kelganini va oldingi so‘zning paydo bo‘lishi soniga bo‘lish orqali amalga oshirish natijasida bajariladi. Bu fikr Markov kontseptsiyasiga parallel ravishda unda so‘z birikmasi (kelajak) bundan boshqa (o‘tmish) so‘zlarga emas, balki faqat oldingi so‘zga (hozirgi) bog‘liqligini bildiradi. N grammalar nisbatan sodda va samarali, ular ketma-ketlikda so‘zlarning uzoq muddatli kontekstini yuklab olish mumkin shu bilan birga til modelining quydagi turi haqida fikr yuritmoqchimiz.

Neyron tili modellari

Neyron tili modellari, nomidan ko‘rinib turibdiki, so‘zlar ketma-ketligi ehtimolini bashorat qilish uchun neyron tarmoqlardan foydalanadi. Ushbu modellar katta hajmdagi matnli ma’lumotlarga o‘rgatilgan va tilning asosiy tuzilishini o‘rganishga qodir. Ikkita yashirin qatlamli neyron tarmoq arxitekturasi.

Ular katta lug‘atlarni boshqarishi va taqsimlangan vakilliklardan foydalangan holda noyob yoki noma’lum so‘zlar bilan

shug‘ullanishi mumkin. NLP vazifalari uchun eng ko‘p qo‘llaniladigan neyron tarmoq arxitekturalari takrorlanuvchi neyron tarmoqlari (RNN) va transformator tarmoqlaridir (bularni keyingi bo‘limda ko‘rib chiqamiz).

Neyron tili modellari kontekstni an’anaviy statistik modellarga qaraganda yaxshiroq ishlay oladi. Bundan tashqari, ular murakkabroq til tuzilmalarini va so‘zlar orasidagi uzoqroq bog‘liqlikni boshqarishi mumkin. Keling, RNN va transformator kabi neyron tili modellari buni qanday qilishini aniq tushunamiz. Tabiiy tilni qayta ishlash kontekstida statistik model oddiyroq til tuzilmalarini boshqarish uchun etarli bo‘lishi mumkin. Biroq, murakkablik oshgani sayin, bu yondashuv samarasiz bo‘ladi. Misol uchun, juda uzun statistik model matnlari bilan ishlaganda, aniq bashorat qilish uchun zarur bo‘lgan barcha ehtimollik taqsimotlarini eslab qolish qiyin bo‘lishi mumkin.

Buning sababi, 100 000 so‘zdan iborat matnda model 100 000 ehtimollik taqsimotini eslab qolishi kerak. Va agar model ikki so‘zni orqaga qaytarishi kerak bo‘lsa, eslashi kerak bo‘lgan taqsimotlar soni 100 000 kvadratgacha oshadi. Bu yerda RNN kabi murakkabroq modellar rol o‘ynaydi. Takroriy neyron tarmoqlari (RNN) keyingi kirishlarni qabul qilishda oldingi natijalarni eslay oladigan neyron tarmoq turidir. Bu kirish va chiqishlar bir-biridan mustaqil bo‘lgan an’anaviy neyron tarmoqlardan farqli o‘laroq. RNNlar, ayniqsa, jumladagi keyingi so‘zni taxmin qilish kerak bo‘lganda foydalidir, chunki ular jumladagi oldingi so‘zlarni hisobga olishlari mumkin. RNN ning asosiy xususiyati yashirin holat vektori bo‘lib, ketma-ketlik ma’lumotlarini eslab qoladi. Ushbu “xotira” RNN ga barcha hisoblangan ma’lumotlarni kuzatib borish va bu ma’lumotlardan bashorat qilish uchun foydalanish imkonini beradi. Yashirin holat tarmoqdagi yashirin qatlam tomonidan saqlanadi. Biroq, RNN hisoblash qimmat bo‘lishi mumkin va juda uzoq kirish ketma-ketligi uchun yaxshi o‘lchovga ega emas. Gap uzunroq bo‘lganda, boshlang‘ich so‘zlardan olingan ma’lumotlar ko‘chiriladi va gapning qolgan qismi bilan birga uzatiladi. RNN jumlaning oxirgi so‘ziga yetib borgunga qadar, birinchi so‘zdan olingan ma’lumotlar nusxaning nusxasiga aylandi va ko‘p marta o‘zgartiriladi.

Lingvistika hamda kompyutor-texnologiyalari sohasida 2020-yilda yana bir ajoyib sun’iy intellekt yaratildi GPT-3 deb nomlangan va San-Fransiskoda OpenAI tomonidan ishlab chiqilgan bo‘lib, u kitoblar, maqolalar va veb-saytlardan milliardlab so‘zlarni o‘zlashtirgandan so‘ng ravon matn yaratishga qodir “katta til modeli”. GPT-3 shu qadar rivojlanganki, ko‘pchilik model tomonidan

yaratilgan yangiliklarni inson mualliflari yozgan yangiliklardan farqlashda qiynalgan. GPT-3-da suhbat vazifari uchun maxsus sozlangan ChatGPT bolalar versiyasi mavjud. Ushbu yutuqlar bilan tilni modellashtirish kontsepsiyasi butunlay yangi davrga kirdi. OpenAI ChatGPT Til modellari tabiiy tilni qayta ishlashning (NLP) asosiy komponentidir, chunki ular mashinalarga inson tilini tushunish, yaratish va tahlil qilish imkonini beradi. Ular asosan kitoblar yoki maqolalar to‘plamlari kabi matnli ma’lumotlarning katta to‘plamidan foydalangan holda o‘qitiladi. Keyin modellar ushbu o‘quv ma’lumotlaridan olingan belgilardan jumladagi keyingi so‘zni bashorat qilish yoki grammatik jihatdan to‘g‘ri va semantik jihatdan mos keladigan yangi matn yaratish uchun foydalanadilar.

Tilshunoslik hamda kompyuter texnologiyalarining integrallashuvi jahon tilshunosligida ko‘plab amaliy va nazariy tadqiqotlarning olib borilishiga asos bo‘lib xizmat qilmoqda. Mana shunday tadqiqotlar sohaning yana-da rivojlanishini ta’minlaydi. Linvistik modellar tilshunoslikka doir masalalarni kompyuter dasturlari yordamida hal etilishida muhim vositadir ana shunday lingvistik dasturlar esa tilni modellashtirish asosida yaratiladi.

Til modellarini qo‘llash natijasida Google Gboard va Microsoft SwiftKey klaviaturalaridagi matnli xabarlarni yozishda jumlalarni yakunlash bo‘yicha avtomatik takliflarni taqdim etuvchi aqlli funksiyalar Ishlab chiqildi. Bundan tashqari kontent yaratish ham shular jumlasidandir. Til belgilarining aniq namoyon bo‘ladigan sohalardan biri kontent yaratishdir. Bu odamlar tomonidan taqdim etilgan ma’lumotlar va atamalardan to‘liq matnlar yoki matn qismlarini yaratishni o‘z ichiga oladi. Kontent yangiliklar maqolalari, press-relizlar va blog postlaridan tortib onlayn-do‘kon mahsulot tavsiflarini qamrab oladi. Shu bilan birga til modellari POS yorlig‘i vazifalarida eng so‘nggi natijalarga erishish uchun keng qo‘llaniladi. POS teglash - bu matndagi har bir so‘zni ot, fe’l, sifat va boshqalar kabi nutqning tegishli qismi bilan belgilash jarayonidir.

Tabiiy tilni qayta ishlash til modellarini kishlab chiqish jahon tilshunosligida barcha rivojlangan tillar qatorida fransuz tilshunosligini ham tadqiqotlar bilan boyitish zaruriyatini tug‘dirdi.

Mana shunday amaliy hamda nazariy tadqiqotlar natijasida avtoregressiv til modellarining hajmini oshirish va tayyorlash nol va past bosqichli o‘rganish yordamida tabiiy tilni qayta ishlash muammolarini hal qilishning yangi usullarini topishga imkon berdi. GPT-3 kabi ekstremal miqyosdagi til modellari ko‘p tilli imkoniyatlarni taklif qilsa-da, ingliz tilidan boshqa tillar deyarli o‘rganilmagan.

Fransuzlar tomonidan fransuz tili uchun maxsus tayyorlangan katta ochiq manbali avtoregressiv til modeli taqdim etiladi. Natijalar shuni koʻrsatadiki, Cedille mavjud fransuz modellaridan ustundir va GPT-3 bilan bir qator fransuz koʻrsatkichlari boʻyicha raqobatlashadi. Bundan tashqari, ushbu modellar tomonidan koʻrsatilgan toksiklikni chuqur taqqoslashni taqdim etadi, bu Sedille ma'lumotlar toʻplamini filtrlash tufayli til modellarining xavfsizligi yaxshilanishini koʻrmoqda. Camembert Fransuz til modeli CamemBERT - yangi koʻp tilli OSCAR korpusining fransuz sub-korpusida oldindan tayyorlangan RoBERTa arxitekturasiga asoslangan fransuz tili uchun eng zamonaviy til modeli. Bu CamemBERTni fransuz tili uchun toʻrt xil quyi oqim vazifasi boʻyicha baholaydi. Nutqning bir qismini (POS) teglash, bogʻliqlik tahlili, nomli obyektlarni aniqlash (NER) va tabiiy til xulosasi (NLI); oldingi monolingual va koʻp tilli yondashuvlar bilan solishtirganda koʻpgina vazifalar boʻyicha eng zamonaviy texnologiyalarni yaxshilash, fransuz tili uchun oldindan tayyorlangan katta til modellarining samaradorligini tasdiqlaydi.

CamemBERT Lui Martin, Benjamin Myuller, Pedro Ortiz Suares, Yoan Dupon, Loran Romari, Erik Villemonte de la Klerjeri, Jamet Sedda va Benoit Sago tomonidan oʻrganilgan va baholangan.

Tilshunoslikda modellashtirishning asosiy maqsadi shaxsning yaxlit lingvistik qobiliyatini modellashtirishdir. Hozirgi tilshunoslikda "model" atamasining mazmuni asosan "nazariya" atamasi bilan ilgari yoritilgan, ayniqsa, Yelmslev tomonidan. Faqatgina bunday nazariya model nomiga loyiq deb hisoblanadi. Bu juda aniq ifodalangan va yetarlicha rasmiylashtirilgan (ideal holda, har bir model kompyuterda amalga oshirish imkonini berishi kerak) Modelni loyihalash nafaqat lingvistik hodisalarni aks ettirish vositalaridan biri, balki til hodisalari haqidagi bilimlarning haqiqatini tekshirishning obyektiv amaliy mezonidir. Til oʻrganishning boshqa usullari bilan birlikda modellashtirish nutq faoliyatining yashirin mexanizmlari, uning nisbatan ibtidoiy modellardan tilning mohiyatini toʻliqroq ochib beruvchi yanada mazmunli modellarga oʻtishi haqidagi bilimlarni chuqurlashtirish vositasi boʻlib xizmat qiladi. Tizim sifatida til tarkibida mavjud.

Modellashtirish prinsipi: uning ayrim quyi tizimlari boshqalarni modellashtiradi, Modellash usuli odatda ishora tizimlariga tayanadi, lekin tilning oʻzi ishora tizimidir, yani. Soʻzlarni soʻzlar yordamida modellashtiriladi. Har qanday model, jumladan, lingvistik model ham aniq boʻlishi kerak. Model, agar u ularning bayonotlarini bogʻlaydigan boshlangʻich obyektlarni

va ular bilan ishlash qoidalarini (yangi obyektlar va bayonotlarni shakllantirish yoki identifikatsiya qilish qoidalari) aniq koʻrsatilishi lozim. Formallik, toʻgʻrilik, noaniqlik nazariya taqdim etilayotgan tilning xossasidir. Bu xususiyat oʻz-oʻzidan rasmiy nazariyaning bashoratlari obyektiv eksperimental ma'lumotlar bilan mos kelishini ta'minlamaydi. Formal model eksperimental ma'lumotlar bilan bitta yoki boshqa talqin. Modelni talqin qilish deganda model obyektlari (ramzlari) oʻrniga ma'lum bir predmet sohasi obyektlarini, masalan, tilni almashtirish uchun ehtimollik yoki qat'iy qoidalarni koʻrsatish tushuniladi.

Tilshunoslikdagi modellar tizimi tilshunoslikda tilni va uning alohida tomonlarini (fonologik, grammatik, leksik va boshqa tizimlar) tavsiflash uchun lingvistik tushunchalar va ular oʻrtasidagi aloqalarni aniqlashtirish uchun qoʻllaniladi, bu til hodisalarining cheksiz xilma-xilligi asosida yotgan tuzilmalarni aniqlashga yordam beradi. Qoʻllash sohasiga qarab, modellar fonologik, morfologik, sintaktik va semantiklarga boʻlinadi. Matematikani qurishda matematik tilshunoslikning vositalari va usullaridan foydalaniladi. Har qanday Modelda quyidagilar qayd etiladi: bevosita kuzatish ma'lumotlariga mos keladigan obyektlar — tovushlar, soʻzlar, gaplar toʻplami; tadqiqotchi tomonidan tavsiflash uchun qurilgan obyektlar (konstruksiyalar) qa'tiy cheklangan toifalar, xususiyatlar, elementar semantik tuzilmalar va boshqalar toʻplamidir.

Har bir model butun tilni emas, balki uning ma'lum bir sohasini yoki hatto alohida toifasini tavsiflaganligi sababli, tilning aniq tavsifi tilning bir sohasiga tegishli turli xil modellardan bir vaqtning oʻzida foydalanishni oʻz ichiga oladi [Советов, Яковлев, 2001]. Tilshunoslarni insonning til qobiliyati qanday tuzilganligi, tillar qanday rivojlanishi va ular oʻzaro qanday munosabatlarda mavjudligi, shuningdek, tabiiy tillardagi matnlar qanday xususiyatlarga ega ekanligi koʻproq qiziqtiradi. Ushbu savollarni oʻrganish uchun tilshunoslar koʻpincha matematik usullardan foydalanishlari kerak.

1950-1960-yillarda matematik tilshunoslikning boshlanishida, matematik mantiq usullari eng ommabop boʻlgan va hozirda, kompyuterlarning sohaga kirib kelishi hamda kompyuter va lingvistikaning integrallashuvi natijasida, turli xil statistik algoritmlar tobora muhim rol oʻynamoqda. Matematik mantiqning gʻoyalari va usullari til *hodisalarining* tavsifini yanada formal, shuning uchun yanada qat'iyroq qilish imkonini beradi va miqdoriy usullarni katta hajmdagi lingvistik ma'lumotlarga qoʻllash bizga birinchi qarashda

ko‘rinmaydigan aloqalar va belgilarni aniqlash imkonini beradi. Shuningdek, matnlarni tabiiy tilda qayta ishlash: masalan, matn terish xatolarini tuzatish yoki matnlarni bir tildan boshqa tilga tarjima qilish. Lingvistik model hamda modellashtirish yuzasidan soha doirasida zamonaviy tilshunoslikning turli yo‘nalishlaridagi muammolar va ularni hal qilinilishida matematik usullar asos bo‘lib xizmat qiladi.

Bir so‘z bilan aytganda tabiiy tilni qayta ishlashda lingvistik modellar asos sifatida qaralar ekan, jahon tilshunosligida tilni modellashtirish masalasiga bo‘lgan ilmiy nazariy hamda amaliy yondoshuvlari kompyuter lingvistikasi doirasida o‘rganilayotgan har bir sohasi dolzarb ekanligini belgilash bilan bir qatorda uning yanada rivojlanishiga zamin yaratadi.

Foydalanilgan adabiyotlar

Баранов А.Н. Введение в прикладную лингвистику. - М.: Изд-во ЛКИ, 2007. 360 – с.

Лосев А.Ф. Введение в общую теорию языковых моделей: учебное пособие. - М., 2004.

Советов Б.Я., Яковлев С.А. Моделирование систем. - М.: Высшая школа, 2001.

Лихачев Д. С. Концептосфера русского языка // Русская словесность: антология. – М.: Academia, 1997. – С. 280–287.

Попова З. Д., Стернин И. А. Понятие «концепт» в лингвистических исследованиях. – Воронеж, 1999. – С. 144.

Бабушкин А. П. Типы концептов в лексико-фразеологической системе языка. – Воронеж: Изд-во ВГУ, 1996. – 104 с.

Арутюнова Н. Д. Язык и мир человека. – М.: Языки русской культуры, 1999. – 896 с.

Алефиренко Н. Ф. Проблемы вербализации концепта: теоретическое исследование. – Волгоград: Перемена, 2003. – 96 с.

THEORY AND PRACTICE OF NATURAL LANGUAGE MODELING IN WORLD LINGUISTICS

Sabura Xudayarova[1]

Abstract. Several decades have passed since the advent of technology, which brought to life the formation and rapid development of computer technology. The development of such a non-traditional field of linguistics as computer has created the need to model linguistic knowledge and processes. This article focuses on the differences and similarities of the research conducted in the world linguistics on natural language processing and linguistic modeling.

Key words: *Model, modeling, deductive, analytical, synthetic, meta-language, N-gram, ELM program, Open AT, Neuron.*

References

Baranov A.N. Vvedenie v prikladnuyu lingvistiku. - M.: Izd-vo LKI, 2007. 360 – s.

Losev A.F. Vvedenie v obщuyu teoriyu yazыkovыx modeley: uchebnoe posobie. - M., 2004.

Sovetov B.Ya., Yakovlev S.A. Modelirovanie sistem. - M.: Vыsshaya shkola, 2001.

Lixachev D. S. Kontseptosfera russkogo yazыka // Russkaya slovesnostь: antologiya. – M.: Academia, 1997. – S. 280–287.

Popova Z. D., Sternin I. A. Ponyatie «kontsept» v lingvisticheskix issledovaniyax. – Voronej, 1999. – S. 144.

Babushkin A. P. Tipы kontseptov v leksiko-frazeologicheskoy sisteme yazыka. – Voronej: Izd-vo VGU, 1996. – 104 s.

Arutyunova N. D. Yazыk i mir cheloveka. – M.: Yazыki russkoy kulьturы, 1999. – 896 s.

Alefirenko N. F. Problemы verbalizatsii kontsepta: teoreticheskoe issledovanie. – Volgograd: Peremena, 2003. – 96 s.

[1]*Xudayarova Sabura Shuxrat qizi* - Master of degree. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: *xudayorovasabura@navoiy-uni.uz*
ORCID: 0000-0001-7102-0824

ISSN 2181-922X 63-73

TENSORFLOW KUTUBXONASINING IMKONIYATLARI

Jahongir Berdiyev[1]

Annotatsiya. Hozirgi axborotlashgan davrda ma'lumotlardan foydalanish juda muhim ahamiyat kasb etmoqda. Ma'lumotlarni qayta ishlash, tasniflash, tahlil qilish, tahrir qilish kabi masalalarni hal qilishda mashinaga bo'lgan ehtiyoj ortib bormoqda. Bu vazifalarni mashina orqali bajarish insoniyatni qo'l mehnatidan va ortiqcha vaqt yo'qotishdan qutqaradi. Buning uchun esa mashinali o'rganishda kerak bo'ladigan dasturiy ta'minoit kutubxonasidan foydalaniladi. Shunday dasturiy ta'minot kutubxonalaridan biri TensorFlowdir.

TensorFlow mashinali o'qitish chuqur o'rganish modellarini ishlab chiqish va qo'llash uchun asos bo'lib xizmat qiladi. Ushbu maqolada TensorFlow kutubxonasining asosiy tamoyillari, xususiyatlari va bir qancha ilovalari yoritiladi. Uning asosiy tushunchalari, jumladan, tensorlar, hisoblash grafiklari va operatsiyalari bilan tanishish, TensorFlowning asosiy arxitekturasi haqida ham ma'lumotga ega bo'lish mumkin. Bundan tashqari, TensorFlowning turli xil platformalar uchun bir qancha turlari haqida tushunchaga ega bo'lish mumkin. Tabiiy tilni qayta ishlash va shu kabi turli sohalarni qamrab olgan aniq misollar va amaliy tadqiqotlar orqali maqola TensorFlowning ko'p qirraliligini namoyish etadi. Qolaversa, maqolada TensorFlowning boshqa vositalar bilan integratsiyalashuvi keltirilgan, bu esa mashinani o'rganish hamjamiyatida o'zaro hamkorlik qilishni yuzaga keltirishi mumkin. Shuningdek, maqolada TensorFlowning rivojlanayotgan shakli va uning sun'iy intellekt tadqiqotlari hamda sanoat ilovalaridagi yutuqlarni boshqarishdagi hal qiluvchi roli haqida fikr yuritiladi. TensorFlow kutubxonasi haqida atroflicha ma'lumot berib, ushbu maqola mashinali o'rganishni yangi boshlovchilar uchun kerakli manba bo'lib xizmat qiladi.

Kalit so'zlar: *TensorFlow, tensor, sun'iy intellekt, mashinali o'qitish, chuqur o'rganish.*

[1]*Berdiyev Jahongir Botir o'g'li* -Alisher Navoiy nomidagi Toshkent davlat o'zbek tili va adabiyoti universiteti kompyuter lingvistikasi mutaxassisligi 1-kurs magistranti
E-pochta: *berdiyevjahongir94@gmail.com*
ORCID: 0009-0002-3756-5681

Kirish

Jahonda TensorFlow dasturiy ta'minot kutubxonasidan foydalanish, uning amaliyotda keng qo'llanilishini juda ko'p uchratish mumkin. Hozirgi zamonda sun'iy intellektning tobora bilan rivojlanayotgani va unga bo'lgan talab kun sayin ortib borayotgani TensorFlow kutubxonasining o'rni naqadar ahamiyatli ekanligini isbotlab turibdi deyish mumkin.

Sun'iy intellekt (AI) va mashinali o'qitish (ML) tez rivojlanayotgan zamonda TensorFlow kutubxonasi ko'plab soha vakillariga chuqur o'rganishni amalga oshirish uchun asosiy vosita bo'lib turibdi. Google kompaniyasining Brain jamoasi tomonidan ishlab chiqilgan TensorFlow kutubxonasi turli xil ilovalarda neyron tarmoqlarni qurish, o'qitish va joylashtirish uchun standart vosita sifatida yuzaga chiqdi. Yaratilganidan beri TensorFlow bir qancha jarayonlardan o'tib, sezilarli evolyutsiyani boshdan kechirdi. Uning ko'p qirraliligi, kengaytirilishi va mukammalligi uni chuqur o'rganish inqilobining oldingi safiga olib chiqdi va kompyuterni ko'rish, tabiiy tillarni qayta ishlash, robototexnika, sog'liqni saqlash va boshqa sohalardagi murakkab muammolarni hal qilish usulini shakllantirdi. Mashinali o'qitish va chuqur o'rganish jarayonlarida muhim o'rin tuta boshladi va bu dasturchilar uchun yana-da yaxshilangan natijalarga erishishga imkon berdi.

Jahonda TensorFlow kutubxonasidan foydalanish bo'yicha ko'plab tadqiqotlar amalga oshirilgan. Xususan, B.V.ElsevierTensorFlowda mashina o'rganish loyihalarini omborlarda saqlashdagi muammolarni aniqlash [Elsevier, 2020] bo'yicha tadqiqot o'tkazgan. A.Shener va boshqalarning tadqiqotida esa temiryo'l liniyalaridagi nosozliklarni aniqlash [Sheyner, 2022] uchun TensorFlow kutubxonasi yordamida yaratilgan konvolyutsion neyron tarmoqdan iborat sun'iy intellektga asoslangan modelni taklif qiladi va bunda ma'lumotlar bazasi sifatida nosoz temiryo'l liniyalari rasmlaridan foydalanadi. N.Shukla va K.Friklas "TensorFlow yordamida mashinani o'rganish" [Shukla, Kfriklas, 2018] deb nomlangan qo'llanma ishlab chiqishgan.

Asosiy qism

Eng ilg'or sun'iy intellekt modellarini ishlab chiqish, unumdorlikni optimallashtirish, keng miqyosda yechimlarni ishlab chiqish yoki chuqur o'rganishning imkoniyatlarini o'rganishda TensorFlow foydalanuvchiga ko'plab vositalar va resurslarni taklif etadi. Google Brain jamoasi tomonidan ishlab chiqilgan TensorFlow

2015-yil noyabr oyida ochiq manbali mashinani o‘rganish tizimi sifatida chiqarildi. Biroq uning tarixi va rivojlanishi Google va boshqa kompaniyalarning mashinani o‘rganishdagi avvalgi sa’y-harakatlariga borib taqaladi. Jumladan, TensorFlowning yaratilishi uchun quyidagi ikki bosqich muhim ahamiyat kasb etdi.

1. Ilk tadqiqot va ishlanmalar (2000 – 2010-yillar). Google doim mashinani o‘rganish va sun’iy intellektga qiziqish bildirib kelgan va buning uchun doim izlanishda bo‘lib kelgan. Mashinani o‘rganish bo‘yicha ilk amaliy ishlar XX asrning 50-yillaridan boshlangan bo‘lsa-da, 2000-yillardan boshlab mashinani o‘rganishga bo‘lgan qiziqish yana-da kuchaya boshladi. Ayniqsa, Google jamoasi tadqiqotchilari, jumladan Jeffri Xinton, Jeff Din va boshqalar neyron tarmoqlar va chuqur o‘rganish usullarini rivojlantirishga hissa qo‘shdilar. Bu usullar tilni aniqlash, tasvirni tanish, tarjima qilish kabi turli xil vazifalarda birinchi o‘rinda turgan. Ushbu sa’y-harakatlar TensorFlowning paydo bo‘lishi uchun asos yaratdi.

2. DistBelief. TensorFlowdan oldin Google kompaniyasining Brain jamoasi 2011-yildan boshlab DistBeliefni chuqur o‘rganish neyron tarmoqlariga asoslangan xususiy mashinani o‘rganish tizimi sifatida yaratdi. Google kompaniyasi DistBelief kod bazasini soddalashtirish va qayta tiklash uchun bir nechta kompyuter olimlarini, jumladan, mashhur Jeff Dinni jalb qildi [Wikipedia, 2023]. DistBelief chuqur o‘rganish landshaftini shakllantirishda katta ta’sir ko‘rsatdi, ammo uning moslashuvchanligi va kengaytirilishi nuqtayi nazaridan cheklovlar mavjud edi. Buning natijasida esa TensorFlowning yaratilishiga zamin payti bo‘ldi.

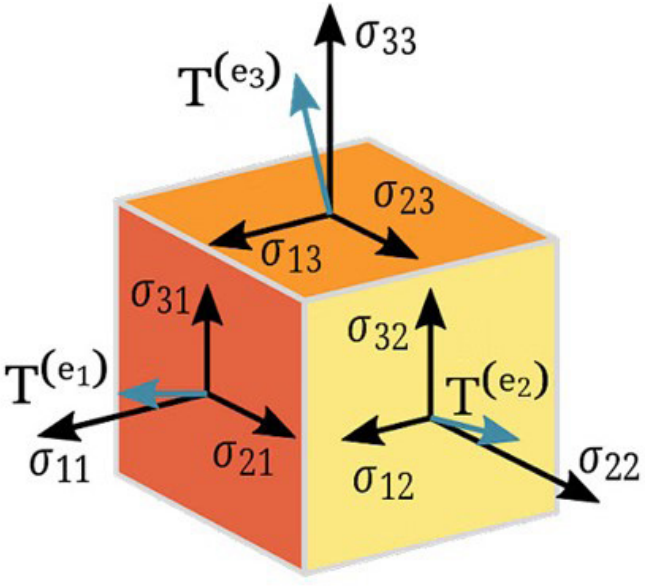
2015-yil noyabr oyida Google TensorFlow-ni Apache 2.0 litsenziyasi ostida ochiq manbali loyiha sifatida chiqardi. Ushbu harakat eng zamonaviy mashinalarni o‘rganish vositalariga kirishni yengillashtirdi va bu sohada innovatsiyalarni tezlashtirdi. TensorFlow tadqiqotchilar, ishlab chiquvchilar va sanoat amaliyotchilari orasida foydalanish qulayligi, keng ko‘lamli hujjatlar va mashinalarni o‘rganish bo‘yicha keng ko‘lamli vazifalarni qo‘llab-quvvatlashi tufayli tezda mashhur bo‘ldi. Uning faol hamjamiyati kutubxonalar, o‘quv qo‘llanmalari va oldindan tayyorlangan modellarni ishlab chiqish orqali uning o‘sishiga hissa qo‘shdi. TensorFlow yordamida mashinani o‘rganishning turli usullari, jumladan, tasvirlarni aniqlash, matnlar ustidagi turli amallar, ovozni tanib olish kabilar turli xil sohalar uchun qulayliklar yaratdi. Masalan, Abdullah Shener va boshqalar tarafidan amalga oshirilgan tadqiqot temiryo‘l liniyalaridagi nosozlikni aniqlashdan iborat. Ma’lumotlar to‘plami

sifatida esa temiryo‘l liniyalarining soz holatdagi va nosoz holatdagi tasvirlaridan foydalangan. Mualliflar buni shunday tushuntiradi: “Ma’lumotlar to‘plami temir yo‘l liniyalarini tashkil etuvchi temiryo‘l tasvirlaridan iborat. Ma’lumotlar to‘plami ommaviy va ikkita sinfdan iborat. Sinflar nuqsonli va nuqsonsiz toifalardan iborat. Har bir tasvirning o‘lchamlari aniq emas va piksellar nisbati, odatda, yaxshi sifatga ega. Rasm hajmi 24 bit va fayl kengaytmalari JPG. nosoz sinfga bo‘shashgan mahkamlagichlar, yetishmayotgan qismlar, yetishmayotgan mahkamlagichlar, ulanish nuqtalaridagi bo‘shliqlar va boshqalar kiradi [Abdullah Shener va b., 2022].

TensorFlow yordamida ifodalangan hisob-kitoblar telefonlar va planshetlar kabi mobil qurilmalardan, yuzlab mashinalarning keng miqyosli taqsimlangan tizimlariga va GPU kartalari kabi turli xil hisoblash qurilmalarigacha bo‘lgan turli xil tizimlarning keng assortimentida juda kam yoki hech qanday o‘zgartirishlarsiz amalga oshirilishi mumkin [Fatih Erdam, Galip aydin, 2017].

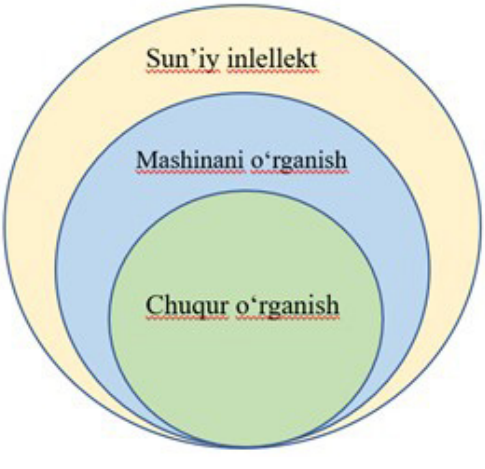
Chuqur o‘rganish dasturlari juda murakkab, o‘quv jarayoni juda ko‘p hisoblashni talab qiladi. Ma’lumotlar hajmi katta bo‘lgani uchun bu uzoq vaqt talab etadi va u bir nechta iteraktiv jarayonlarni, matematik hisoblarni, matritsalarni ko‘paytirishni va hokazolarni o‘z ichiga oladi. Agar bu amallar oddiy markaziy protsessorda (CPU) bajarilsa, bu ancha uzoq davom etadi. TensorFlowning asosiy afzalliklaridan biri shundaki, u GPUlarni, shuningdek, markaziy protsessorlarni, hatto mobil qurilmalarni ham qo‘llab-quvvatlaydi. Shuningdek, u Keras va Torch kabi boshqa chuqur o‘rganish kutubxonalariga qaraganda tezroq kompilyatsiya vaqtiga ega.

Tensorlar. TensorFlowning asosiy qismini tensorlar tashkil qiladi. TensorFlowdagi barcha hisob-kitoblar tensorlarni o‘z ichiga oladi. Bu ma’lumotlarning bir nechta turlarini ifodalovchi n-o‘lchovli matritsadir. Neyron tarmog‘iga kirish sifatida beriladigan o‘lchamlari har xil bo‘lgan ma’lumotlar massivlari tensorlar deyiladi. Tensor hisoblash natijasi bo‘lishi mumkin yoki u kiritilgan ma’lumotlardan kelib chiqishi mumkin.



1-rasm. Tensorning ko‘rinishi

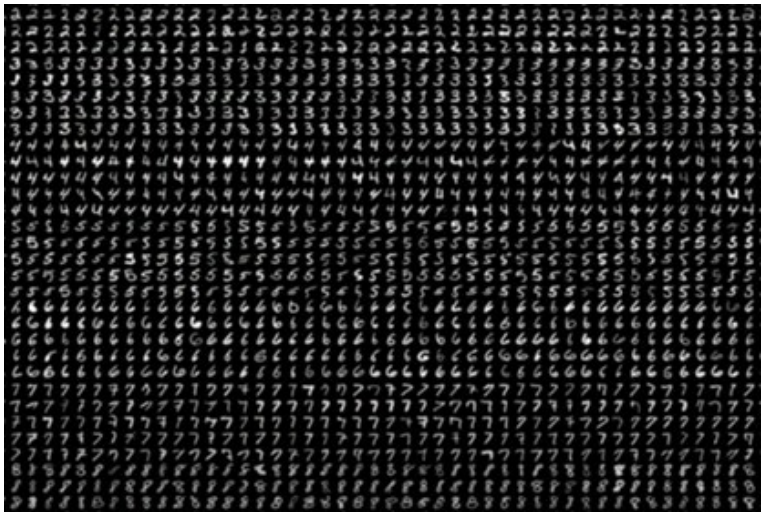
TensorFlow sun’iy intellekt uchun muhim ahamiyatga ega. Chunki u sun’iy intellektga bog‘liq holda ishlaydi va o‘z-o‘zini o‘qitadi. Umuman olganda, sun’iy intellekt mashinani o‘rganish va chuqur o‘rganishni o‘z ichiga oladi. Uni quyidagicha tasavvur qilish mumkin:



2-rasm. Sun’iy intellektning tuzilishi

TensorFlow ma’lumotlarni qayta ishlashda uch bosqichni o‘z ichiga oladi. Bular kirish qatlam, yashirin qatlam hamda chiqish qatlamlardir. Kirish qatlamda matnlar, ovozli ma’lumotlar, tasvirlar kabi ma’lumotlar qabul qilinadi. Kirish qatlam qabul qilingan

ma’lumotlarni yashirin qatlamga uzatadi. Yashirin qatlamda murakkab hisob-kitoblarni amalga oshirish orqali kiritilgan ma’lumotlar qayta ishlanadi va ma’lumotlarning o‘ziga xos xususiyatlari ajratib olinadi [https://www.simplilearn.com/tutorials/deep-learning-tutorial/what-is-tensorflow, 2023]. Masalan, kiritilgan ma’lumotlar turli xil tasvirlar bo‘lsa, ushbu tasvirlar nimaning tasviri ekanligi ajratib olinadi. Yoki qo‘lda yozilgan raqamlarni aniqlash uchun turli xil shaklda yozilgan raqamlar bazasi kiritiladi va qaysi shakllar qaysi raqamga to‘g‘ri kelishi aniqlab olinadi.



3-rasm. Qo‘lda yozilgan turli raqamlar bazasidan namuna

Chiqish qatlamda qayta ishlangan ma’lumotlar qaysi sinfga tegishli ekanligi bashorat qilinadi. Masalan, u tasvir bo‘lsa, bizga kiritilgan tasvir qaysi sinfga tegishli ekanligini aytadi. Yoki yuqorida qo‘lda yozilgan raqamlar kiritilgan bo‘lsa, bizga chiqish qatlamda ushbu raqamlarni bashorat qilib beradi.

TensorFlow arxitekturasi. TensorFlow arxitekturasini qatlamli yondashuv orqali tushunish mumkin:

1. Asosiy (pastki darajali):

Ushbu qatlam asosiy e’tiborni hisoblash uchun asosiy qurilish bloklariga qaratadi. Bunga quyidagilar kiradi:

1. **Operatsiyalar (Ops):** Bular ma’lumotlar bo‘yicha hisob-kitoblarni amalga oshiradigan asosiy matematik birliklardir. Misollar qo‘shish, ko‘paytirish va neyron tarmoqlardagi konvolyutsiya kabi murakkabroq funksiyalarni o‘z ichiga oladi.

2. **Tensorlar:** Yuqorida aytib o‘tilganidek, tensorlar ma’lumotlarni ifodalovchi ko‘p o‘lchovli massivlardir. Ular

TensorFlowda ishlatiladigan asosiy ma'lumotlar tuzilmalari.

3. **Ma'lumotlar oqimi grafigi:** Bu qatlam modeldagi hisob-kitoblar oqimini belgilaydi. Operatsiyalar tarmoq orqali ma'lumotlar qanday oqishini ko'rsatuvchi grafikni shakllantirish uchun ulanadi .

2. O'rta daraja (ixtiyoriy):

Ushbu qatlam ma'lumotlar oqimi grafigini yaratish va boshqarish uchun vositalarni taqdim etadi. Bunga quyidagilar kiradi:

1. **Ramziy dasturlash:** Bu hisoblash grafigini bajarishdan oldin ramziy ifodalar yordamida aniqlash imkonini beradi. Bu modellarni yaratish va o'zgartirishda moslashuvchanlikni ta'minlaydi.

2. **Avtomatik farqlash:** Mashinani o'rganish modellarini o'rgatish uchun muhim xususiyat. U mashg'ulot vaqtida model og'irliklarini optimallashtirish uchun foydalaniladigan gradientlarni (parametrlarga nisbatan yo'qotish funksiyasining o'zgarishi) avtomatik ravishda hisoblab chiqadi.

3. Yuqori daraja:

Ushbu qatlam modellarni yaratish va o'qitish uchun foydalanuvchilarga qulay interfeyslarni taklif etadi:

1. **Eager Execution:** Bu pythondagi kabi operatsiyalarni to'g'ridan-to'g'ri bajarishga imkon beradi. U ramziy dasturlash bilan solishtirganda tezroq va interaktiv tajribani taqdim etadi.

2. **Keras:** TensorFlow ustiga qurilgan yuqori darajadagi API, neyron tarmoqlarni qurishning sodda va intuitiv usulini taklif qiladi . U quyi darajadagi ba'zi murakkabliklarni mavhumlashtiradi.

4. Taqsimlangan ijro (ixtiyoriy):

TensorFlow bir nechta mashinalar yoki GPUlarda taqsimlangan trening uchun ishlatilishi mumkin. Ushbu qatlam katta ma'lumotlar to'plamlari bo'yicha o'qitishni tezlashtirish uchun turli qurilmalar bo'ylab hisoblashlarni taqsimlash va muvofiqlashtirish bilan shug'ullanadi.

Ushbu arxitekturaning afzalliklari:

Moslashuvchanlik: Qatlamli arxitektura yangi boshlanuvchilar va tajribali foydalanuvchilar uchun turli darajadagi mavhumlikni ta'minlaydi.

Masshtablilik: TensorFlow murakkab modellar va katta ma'lumotlar to'plamlarini taqsimlangan bajarish orqali samarali boshqarishi mumkin.

Ishlash: optimallashtirilgan operatsiyalar va ma'lumotlar tuzilmalari mashinani o'rganish vazifalari uchun samarali hisoblash imkonini beradi.

Ochiq manba: Ochiq manba bo‘lish faol rivojlanishga va vositalar va resurslarning keng ekotizimiga olib keladigan katta hamjamiyatni rivojlantiradi. Umuman olganda, TensorFlow arxitekturasi mashinani o‘rganish modellarining keng doirasini yaratish va joylashtirish uchun kuchli va ko‘p qirrali platformani taqdim etadi.

Xulosa

TensorFlowdan foydalanish orqali tabiiy tilga ishlov berish, neyron tarmoqlardan foydalanish va mashinani o‘rganishda TensorFlowning ahamiyati muhim ekanligiga guvoh bo‘lishimiz mumkin.

TensorFlow boy xususiyatlar to‘plami bilan mashinani o‘rganish imkonini beradi. Uning yadrosi tensorlar (ko‘p o‘lchovli ma’lumotlar) va samarali bajarilishi uchun ma’lumotlar oqimi grafigini tashkil etuvchi operatsiyalar (hisoblashlar) atrofida aylanadi. Ramziy dasturlash modelni moslashuvchan aniqlash imkonini beradi, avtomatik farqlash esa gradientlarni hisoblash orqali o‘qitishni soddalashtiradi. Keras kabi yuqori darajadagi APIlar foydalanuvchilarga qulaylikni ta’minlaydi. Katta ma’lumotlar to‘plamlari uchun taqsimlangan bajarish bir nechta mashinalar – GPUlardan foydalanadi. Muxtasar qilib aytganda, u mashinani o‘rganish modellarini yaratish va joylashtirish uchun kuchli va ko‘p qirrali platformani taklif qiladi.

Foydalanilgan adabiyotlar

A Tour of TensorFlow. https://arxiv.org/abs/1610.01178, 2016.

Abdullah Şener, Burhan Ergen , Mesut Toğaçar. Fault Detection from Images of Railroad Lines Using the Deep Learning Model Built with the Tensorflow Library. Turkish Journal of Science & Technology 17(1), 2022, 47-53.https://doi.org/10.55525/tjst.1056283

Fatih Ertam, Galip Aydın. Data Classification with Deep Learning using Tensorflow. (UBMK, 17) and International Conference on Computer Science and Engineering. 2017.

Janardhanan. Project repositories for machine learning with TensorFlow. Third International Conference on Computing and Network Communications (CoCoNet’19). 2020.

Kai Staats, Edward Pantridge, Marco Cavaglia. TensorFlow Enabled Genetic Programming. https://doi.org/10.48550/arXiv.1708.03157, 2017.

Nicolas Knudde, Joachim van der Herten. GPflowOpt: A Bayesian Optimization Library using TensorFlow. https://doi.org/10.48550/arXiv.1708.03157, 2017.

Nishant Shukla Kenneth Fricklas, Senior Technical Editor. Machine Learning with TensorFlow. Manning Publications, 2018.

https://blog-tensorflow-org.translate.goog/2024/02/graph-neural-networks-in-tensorflow.html?_x_tr_sl=en&_x_tr_tl=uz&_x_tr_hl=uz&_x_tr_pto=sc

https://blog-tensorflow-org.translate.goog/2023/11/half-precision-inference-doubles-on-device-inference-performance.html?_x_tr_sl=en&_x_tr_tl=uz&_x_tr_hl=uz&_x_tr_pto=sc

TENSORFLOW LIBRARY CAPABILITIES

Jahongir Berdiyev[1]

Abstract. In today's information age, the use of information is becoming very important. The need for a machine in solving problems such as data processing, classification, analysis, and editing is increasing. Carrying out these tasks by machines saves mankind from manual labor and excessive time loss. For this purpose, a software library necessary for machine learning is used. One such software library is TensorFlow.

TensorFlow provides a framework for developing and deploying machine learning deep learning models. This article covers the basic principles of the TensorFlow library, its main features, and several applications. You can learn about its basic concepts, including tensors, computational graphs and operations, and the basic architecture of TensorFlow. In addition, it is possible to gain an understanding of several variants of TensorFlow for different platforms. Through concrete examples and case studies covering natural language processing and similar fields, the article demonstrates the versatility of TensorFlow. Additionally, TensorFlow's integration with other tools is presented, which can enable interoperability and collaboration within the machine learning community. The article also reflects on the evolving landscape of TensorFlow and its critical role in driving advances in AI research and industrial applications. Providing an in-depth overview of the TensorFlow library, this article is a must-have resource for machine learning beginners.

Key words: *TensorFlow, tensor, artificial intelligence, machine learning, deep learning.*

References

A Tour of TensorFlow. https://arxiv.org/abs/1610.01178, 2016.

Abdullah Şener, Burhan Ergen , Mesut Toğaçar. Fault Detection from Images of Railroad Lines Using the Deep Learning Model Built with the Tensorflow Library. Turkish Journal of Science & Technology 17(1), 2022, 47-53.https://doi.org/10.55525/tjst.1056283

Fatih Ertam, Galip Aydın. Data Classification with Deep Learning using Tensorflow. (UBMK, 17) and International Conference on Computer Science and Engineering. 2017.

Janardhanan. Project repositories for machine learning with TensorFlow. Third International Conference on Computing and Network Communications (CoCoNet'19). 2020.

[1]*Berdiyev Jahongir Botir o'g'li* - Master of degree. Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: *berdiyevjahongir94@gmail.com*
ORCID: 0009-0002-3756-5681

Kai Staats, Edward Pantridge, Marco Cavaglia. TensorFlow Enabled Genetic Programming. https://doi.org/10.48550/arXiv.1708.03157, 2017.

Nicolas Knudde, Joachim van der Herten. GPflowOpt: A Bayesian Optimization Library using TensorFlow. https://doi.org/10.48550/arXiv.1708.03157, 2017.

Nishant Shukla Kenneth Fricklas, Senior Technical Editor. Machine Learning with TensorFlow. Manning Publications, 2018.

https://blog-tensorflow-org.translate.goog/2024/02/graph-neural-networks-in-tensorflow.html?_x_tr_sl=en&_x_tr_tl=uz&_x_tr_hl=uz&_x_tr_pto=sc

https://blog-tensorflow-org.translate.goog/2023/11/half-precision-inference-doubles-on-device-inference-performance.html?_x_tr_sl=en&_x_tr_tl=uz&_x_tr_hl=uz&_x_tr_pto=sc

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 25.12.2023-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.