O‘ZBEKISTON RESPUBLIKASI FANLAR AKADEMIYASI
MINTAQAVIY BO‘LIMI
XORAZM MA’MUN AKADEMIYASI

XORAZM MA’MUN AKADEMIYASI AXBOROTNOMASI

Axborotnoma OAK Rayosatining 2016-yil 29-dekabrdagi 223/4-son qarori bilan biologiya, qishloq xo‘jaligi, tarix, iqtisodiyot, filologiya va arxitektura fanlari bo‘yicha doktorlik dissertatsiyalari asosiy ilmiy natijalarini chop etish tavsiya etilgan ilmiy nashrlar ro‘yxatiga kiritilgan

2024-12/4
Xorazm Ma’mun akademiyasi axborotnomasi
2006 yildan boshlab chop qilinadi

Xiva-2024

Xorazm Ma’mun akademiyasi axborotnomasi: ilmiy jurnal.-№12/4 (121), Xorazm Ma’mun akademiyasi, 2024 y. – 636 b. – Bosma nashrning elektron varianti - http://mamun.uz/uz/page/56

ISSN 2091-573 X
Muassis: O‘zbekiston Respublikasi Fanlar akademiyasi mintaqaviy bo‘limi – Xorazm Ma’mun akademiyasi

МУНДАРИЖА

FILOLOGIYA FANLARI

UDK: 81:33

TIL KORPUSI TURLARI VA UMUMIY XUSUSIYATLARI

B.B.Elov, dotsent, Toshkent davlat o‘zbek tili va adabiyoti universiteti, Toshkent
A.U.Yuldashev, o'qituvchi, Toshkent davlat o‘zbek tili va adabiyoti universiteti, Toshkent
U.S.Yodgorov, o'qituvchi, Toshkent davlat o‘zbek tili va adabiyoti universiteti, Toshkent

Annotatsiya. *Tabiiy tilni qayta ishlash (NLP) uchun til korpuslari muhim ahamiyatga ega. Til korpusi - elektron shaklda taqdim etilgan, turli manbalardan olingan katta hajmli va strukturlangan matnlar to‘plami bo‘lib, NLP tizimining lingvistik tahlil va o‘rganish jarayonini ta’minlaydi. Ushbu maqolada til korpuslarining struktura va vazifalari, foydalanish sohalari hamda turlari tahlil qilinadi.*

Kalit so‘zlar: *Til korpusi, matnlarini qayta ishlash, lingvistik tahlil, mashinali o‘qitish modellari, korpus turlari, parallel korpuslar, teglangan korpuslar, POS teg.*

Аннотация. *Для обработки естественного языка (NLP) ключевую роль играют языковые корпусы. Языковой корпус представляет собой структурированный массив текстов в электронном формате, собранный из различных источников. Он обеспечивает процесс лингвистического анализа и изучения в системах NLP. В статье рассматриваются структура, функции, области применения и типы языковых корпусов.*

Ключевые слова: *языковой корпус, обработка текстов, лингвистический анализ, модели машинного обучения, типы корпусов, параллельные корпуса, размеченные корпуса, POS-тег.*

Abstract. *Language corpora play a key role in Natural Language Processing (NLP). A language corpus is a structured collection of texts in electronic form, gathered from various sources. It facilitates linguistic analysis and learning processes in NLP systems. This article examines the structure, functions, applications, and types of language corpora.*

Keywords: *language corpus, text processing, linguistic analysis, machine learning models, types of corpora, parallel corpora, annotated corpora, POS tagging.*

Kirish. NLPga asoslangan ilovalar(axborot tizimlari)ni ishlab chiqish uchun NLP tizimni mavjud ma'lumotlarni o'rganishini ta'minlash kerak. Ushbu amalni til korpusi vositasida amalga oshirish mumkin. **Til korpusi** elektron shaklda taqdim etiladigan *katta hajmli* va *strukturlangan matnlar to'plami* sifatida qaraladi. Til korpusi *yozma* yoki *og'zaki* materialni ifodalab, NLP tizimini mavjud resurslarni o'rganishi uchun **lingvistik tahlilni** amalga oshirish lozim.

Korpus – bu kompyuterlar o'qiy oladigan tabiiy tilda yozilgan matnlarning to'plami. Korpuslar *raqamli matn, audio transkript* va hatto *skanerlangan hujjatlar* kabi resurslardan iborat bo'ladi [1,2]. Til korpuslari, odamlar har kuni gapiradigan va yozganidek, tilning haqiqiy hayotda qanday ishlatilishini o'rganish va tushunish uchun juda muhimdir.

Korpus NLP tizimlarining asosidir. Ular AI va mashinali o'rgatish tizimlarini o'rgatish uchun ishlatiladi. Ular axborot hayotiy davrlarini modellashtirish va bashorat qilish uchun keng va xilma-xil ma'lumotlar to'plamini taqdim etadi. Korpus NLP tizimini tabiiy tilni boshqarish va sharhlash uchun tayyorlaydi, bu esa odamlar bilan tabiiy tilda oson muloqot qilish imkonini beradi.

Korpus – tabiiy tilni qayta ishlash (NLP) uchun muhim vosita bo'lib, asosiy manba bo'lib xizmat qiladi. Korpus – bu matn yoki audio ma'lumotlarning muhim, tashkil etilgan to'plami bo'lib, u ko'pincha bir yoki bir nechta aniq tillardagi keng doiradagi hujjatlar, matnlar yoki ovozlarni o'z ichiga oladi.

Til korpusi – bu tilning haqiqiy foydalanuvchilari tomonidan ishlab chiqarilgan va so'zlar, iboralar va, umuman, til qanday ishlatilishini tahlil qilish uchun ishlatiladigan juda katta matnlar to'plami. U tilshunoslar, leksikograflar, ijtimoiy olimlar, gumanitar fanlar, tabiiy tillarni qayta ishlash bo'yicha mutaxassislar va boshqa ko'plab sohalarda qo'llaniladi. Korpus shuningdek, dasturiy ta'minotni ishlab chiqishda ishlatiladigan turli til ma'lumotlar bazalarini yaratish uchun ishlatiladi, masalan, *bashoratli klaviaturalar, imloni tekshirish va tuzatish, matn/nutqni tushunish tizimlari, matndan nutqqa modullar, mashina tarjimasi tizimlari* va boshqalar.

NLPning ko'plab vazifalarini hal qilishda til korpusidan foydalaniladi [3,4,5]:

–**Mashinali o'qitish modellarini o'rgatish:** Turli NLP ilovalari, jumladan *hissiyotlarni tahlil qilish, matn tasnifi, mashina tarjimasi* va *nutqni tanib olish* uchun korpuslar mashinali o'qitish modellarini o'rgatish va takomillashtirish uchun ishlatiladi. Korpusdagi katta hajmdagi matn ma'lumotlar ushbu modellarga *shablonlar (patterns), korrelyatsiyalar (correlations)* va *murakkabliklar (complexities)ni* o'rgatish uchun ishlatiladi.

–**Tilni tushunish:** Til korpusi tabiiy tilning *tuzilishi, grammatikasi, lug'ati* va *qo'llanilishi* haqida to'liq tasavvur beradi. Kontekstda so'zlar va so'z birikmalarning qanday ishlatilishini o'rganish uchun NLP modellarida til korpusidan foydalaniladi. Chet tili o'qituvchilari va o'rganuvchilar uchun til korpusi tilning haqiqiy misollarini taqdim etadi va haqiqiy foydalanish shakllarini ochib beradi. U o'quv dasturlarini ishlab chiqish, materiallarni loyihalash va hatto dars mashg'ulotlarida foydalanish mumkin. Korpus ko'plab til resurslarini, jumladan, grammatika tekshiruvi, imlo tekshiruvi va nutqni aniqlash dasturlarini ishlab chiqishga yordam beradi. Ushbu NLP ilovalari bizning kundalik raqamli tajribamizda ajralmas rol o'ynaydi.

–**Diskurs tahlili:** Korpus nutqni tahlil qilish uchun ham ishlatilishi mumkin, bu tilning muayyan kontekstlarda qanday qo‘llanilishi haqida tushuncha beradi. Bu bizning siyosiy, ijtimoiy va madaniy nutqlar haqidagi tushunchamizni yanada oshirishi mumkin.

–**Tarjimashunoslik:** Til korpusi tarjimashunoslik uchun muhim vosita hisoblanadi. Ikki tilli yoki ko‘p tilli korpus yaratish tilshunoslarga tarjima me’yorlarini o‘rganish, ma’no jihatdan mos tarjimalarni aniqlash va samarali mashina tarjimasi modellarini yaratish imkonini beradi.

–**Qoidalarga asoslangan tizimlar:** Til korpuslari tilshunoslar va NLP mutaxassislari tomonidan til qoidalari va shablonlarini ishlab chiqish va sinab ko‘rish uchun ishlatiladi. Keyinchalik, POS teglash, grammatik qayta ishlash va NERlarni tanib olish kabi NLP vazifalari uchun ushbu qoidalarga asoslangan NLP tizimlarida qo‘llaniladi.

–**Lug‘at va semantika:** Turli lug‘atlar yoki so‘zlardan iborat to‘plamlar til korpuslari yordamida yaratiladi va kengaytiriladi. *Sinonimlar, antonimlar* va *so‘z birikmalari* kabi so‘z munosabatlarini ko‘rsatish orqali ular semantik tahlilga yordam beradi. Korpuslar asosida lug‘atlarni yaratish usulini tubdan o‘zgartirib, leksikografiya sohasini tubdan o‘zgartirdi. Katta hajmli strukturlangan matnlar to‘plamini tahlil qilib, leksikograflar yangi so‘zlarni, sezgilarni va foydalanish shakllarini yanada samaraliroq aniqlashlari mumkin.

–**Statistik tahlil:** Til korpuslari tabiiy tilning *statistik tahlili* uchun foydalidir. Ular yordamida *so‘z chastotasini taqsimlash, birgalikda paydo bo‘lish shablonlari* va boshqa statistik xususiyatlarni o‘rganish uchun ehtimollikka asoslangan NLP yondashuvlari uchun zarur bo‘lgan ma’lumotlarni taqdim etadi.

–**Sohaviy bilimlar:** Til korpuslari turli sohagalarga oid bilimlar manbayi hisoblanadi. Chunki ular muayyan mavzular yoki sohalarga xos bo‘lishi mumkin. *Huquqiy hujjatlarni o‘rganish, tibbiy ma’lumotlarni qayta ishlash* va muayyan sohalar uchun yaratilgan *chatbotlar* kabi ilovalar bunga yaqqol misol bo‘ladi.

Tabiiy til modellarini ishlab chiqish va uning samaradorligini oshirish uchun biz katta hajmdagi til korpusidan foydalanish lozim.

Korpus turlari

NLP tizimini ishlab chiqish uchun foydalanish mumkin bo‘lgan har xil turdagi korpuslar mavjud. NLPda til korpuslari mazmun, maqsad yoki manba kabi turli mezonlar asosida har xil turlarga bo‘linadi.

Matnli korpuslar:

–*Umumiy korpus (general corpora)* bo‘lib, u yozma va og‘zaki nutqdagi materiallardan iborat. Shuningdek, ushbu turdagi korpusda *turli yoshdagi, turli mintaqalardan* va *turli ijtimoiy qatlamlardan* bo‘lgan insonlarning turli shakldagi matnlarini o‘z ichiga oladi. Umumiy maqsadli korpuslar turli janr va sohalardagi turli matnlarni o‘z ichiga oladi. **Gutenberg** va **Braun** korpuslarini bu toidadagi korpuslarga misol sifatida keltirish mumkin [6,7].

–*Bir tilli korpus (Monolingual Corpora)* – bitta tildagi og‘zaki yoki yozma materiallar to‘plamidan iborat korpus. Bu ma’lum bir tildagi til shablonlari, tuzilmalari va qo‘llanilishini o‘rganishda foydalidir. Bir tilli korpus odatda POS teglanadi va turli xil NLP vazifalarini hal qilishda ishlatiladi. Masalan, so‘zning to‘g‘ri qo‘llanilishini tekshirish yoki so‘z birikmalarini izlash, ilmiy foydalanish uchun, tildagi shablonlarni yoki yangi tendensiyalarni aniqlash [8,9].

–*Ixtisoslashgan/maxsus korpus (specialized corpora)*lar – tilning ichki xususiyatlarini tushunish uchun olib boriladigan turli xil tadqiqotlar uchun foydalidir. Korpus va uning turlari haqida boshlang‘ich ma’lumotga ega bo‘lgandan so‘ng, korpus lingvistikasiga e’tibor qaratish kerak. Ixtisoslashgan korpuslar *ilmiy adabiyotlar, yuridik hujjatlar* yoki *tibbiy materiallarni* o‘z ichiga olgan muayyan sohalar yoki mavzularga qaratilgan.

–*Taqqoslanadigan korpus (Comparable Corpora)* – bu turli tillarda yoki turli manbalardan yozilgan o‘xshash ma’no (shakl)ga ega matnlar to‘plami. Bunday korpuslar tillararo yoki domenlararo tadqiqotlar uchun ular ishlatiladi.

–*Og‘zaki korpus (Spoken Corpora)* – nutqning transkripsiyalaridan iborat. Ular intervyular, dialoglar yoki og‘zaki nutqlar kabi turli manbalardan olingan bo‘lishi mumkin va tilning o‘z-o‘zidan foydalanishni o‘rganish maqsadida xizmat qiladi [2,10].

Multimodal korpuslar [11]**:**

–*Text-Image korpuslar* – bu korpuslar matnli va vizual ma'lumotlarni o'z ichiga oladi. Bu ularni rasmlarga sarlavha qo'yish va savollarga vizual javob berish kabi NLP vazifalarini bajarishga moslashtiradi.

–*Text-Speech korpuslar* – bu korpuslar og'zaki nutqni tanib olish va nutqni avtomatik aniqlash bo'yicha olib boriladigan tadqiqotlarni qo'llab-quvvatlash uchun matnli ma'lumotlarni tegishli audio yoki nutq yozuvlari bilan birlashtiradi.

Parallel korpuslar [12,13]**:**

–*Ikki tilli korpus (Bilingual Corpora)* – ikki yoki undan ortiq tabiiy tilda mavjud bo'lgan tarjima qilingan matnlarni o'z ichiga oladi. Tillararo tadqiqot, mashina tarjimasi kabi NLP vazifalarini bajarishda foydalaniladi. Ushbu turdagi korpuslar tarjimashunoslikda juda foydali. Ushbu turdagi korpuslarda ikkala tilni ham moslashtirish kerak, ya'ni mos keladigan segmentlar, odatda gaplar yoki paragraflar mos kelishi kerak. Foydalanuvchi bir tildagi so'z yoki birikmaning barcha misollarini qidirishi va natijalarni boshqa tildagi mos leksik birikmalar bilan birga ko'rsatiladi.

–*Taqqoslanadigan ikki tilli korpus (Comparable Bilingual Corpora)* – tillararo ma'lumot olish uchun foydalidir. Ushbu turdagi korpuslar parallel korpusga o'xshab, bir xil mavzu yoki domenga tegishli ko'plab tillardagi matnlarni o'z ichiga oladi.

Time-Series Corpora [14]**:**

–*Tarixiy korpus (historical corpora)* – oldingi davrdagi matnlardan iborat bo'lib, ular bir necha *o'n yillar* yoki *asrlarni* qamrab olgan [15]. Ushbu turdagi til korpuslari tabiiy tilning oldingi bosqichlari (variantlari)ni chuqur qamrab oladi. Ko'pgina tarixiy davrlardagi yozuvlarni o'z ichiga olgan ushbu korpuslar olimlarga til va tarixiy shablonlarning evolyutsiyasini ko'rib chiqishga imkon beradi.

–*Vaqtinchalik korpus (Temporal Corpora)* – vaqt o'tishi bilan matnlarni saqlab qolishga asoslangan bo'lib, bu ularni lingvistik evolyutsiyani kuzatish va tilning hozirgi holatini o'rganish uchun foydali resurs bo'lib xizmat qiladi.

–*Sinxron korpus (Synchronic Corpora)* – zamonaviy korpuslar hisoblanib, ma'lum bir davrdagi materiallar to'plamidir [16]. Bu turdagi korpuslar ma'lum bir davrda tildan foydalanishni o'rganish imkonini beradi.

–*Diaxronik korpus (Diachronic Corpora)* – vaqt o'tishi bilan tilning evolyutsiyasi haqida qimmatli fikrlarni taklif qiladi [17]. Ular turli davrlarga oid materiallar to'plamini o'z ichiga oladi va tarixiy tilshunoslikda muhim ahamiyatga ega.

Annotatsiyalangan/teglanlar korpuslar:

–*Lingvistik izohli korpus (Linguistically Annotated Corpora)* – POS teglash, grammatik tahlillar va qo'lda bajariladigan NERlarni izohlash kabi lingvistik izohlarni o'z ichiga oladi [18,19]. Ular NLP modellarini ishlab chiqish va sinovdan o'tkazish uchun zarurdir.

–*Semantik izohli korpus (Sentiment-Annotated Corpora)* – matnlaridagi hissiyotlar yoki his-tuyg'ularga oid ma'lumotlar etiketlash uchun foydalaniladi [20]. Bu esa hissiyotlarni tahlil qilish va his-tuyg'ularni aniqlash NLP vazifalarini amalga oshirish uchun muhim resurs hisoblanadi.

–*Ta'limiy korpus (learner corpora)* bo'lib, tabiiy tilni chet tili sifatida o'rganuvchilar uchun ishlab chiqarilgan ma'lumotlar to'plamidan iborat [21]. Ushbu turdagi korpuslar masalan, u *esse, yozma imtihon* yoki so'rovnoma shaklida shakllantirilishi mumkin.

Yuqorida keltirilgan toifalar NLPda keng qo'llaniladigan korpus turlariga bir nechta misollardir. Korpusni tanlash muayyan NLP vazifasiga, tadqiqot maqsadlariga va qo'llash sohasiga bog'liq. Tadqiqotchilar va amaliyotchilar ko'pincha turli NLP loyihalarini hal qilish uchun maxsus korpuslarni ishlab chiqadilar.

Korpusning xususiyatlari

NLPdagi til korpusning xususiyatlari uni til bilan bog'liq barcha turdagi vazifalar va tadqiqotlar uchun juda foydali qiladi. Til korpusining muhim xususiyatlaridan ba'zilari quyidagilar:

–*Korpus hajmi:* Umuman olganda, til korpusi hajmi imkon qadar katta bo'lishi kerak. Katta hajmdagi maxsus ma'lumotlar to'plamlari hissiyotlarni tahlil qilish kabi NLP vazifalarini hal qilishda amalga oshiradigan algoritmlarni o'rgatish uchun zarurdir.

–*Yuqori sifatli ma'lumotlar (High-Quality Data):* Korpusdagi ma'lumotlarning yuqori sifatli ekanligi muhim ahamiyatga ega. O'quv ma'lumotlaridagi eng kichik noaniqlik (xato)lar ham mashinali o'rganish tizimining natijasiga jiddiy ta'sir ko'rsatadi.

–*Tozalangan ma'lumotlar:* Yuqori sifatli korpusni yaratish va qo'llab- quvvatlash toza ma'lumotlarga bog'liq. NLP ulovalari uchun yanada ishonchli korpusni yaratish uchun ma'lumotlarni tozalash juda muhim, chunki bu jarayonda har qanday xatolar yoki takroriy ma'lumotlar aniqlanadi va bartaraf qilinadi.

–*Xilma-xillik:* Turli toifalar, yozuvlar, tillar va mavzular – bularning barchasi korpus taqdim etayotgan til xilma-xilligining bir qismidir. Ushbu xususiyat orqali NLP modellari va algoritmlari lingvistik variantlarning keng doirasini boshqarishga qodir.

–*Annotatsiya:* Tilga xos izohlar, masalan, POS teglash, grammatik tahlillar, NERlar, his-tuyg'u yorliqlari yoki semantik izohlar ko'plab korpuslarning tarkibiy qismi hisoblanadi. Ushbu annotatsiyalar mashinali o'rganish va NLP vazifalarini nazorat qilishga yordam beradi.

–*Metadata:* Muallif ismlari, nashr sanalari, manba tafsilotlari va hujjat nomlari kabi matnlar haqidagi umumiy ma'lumotlar ko'pincha korpusda mavjud. Kontekst va uning kelib chiqish asosini ifodalash uchun metama'lumotlar muhim ahamiyatga ega.

Korpusni shakllantirish

Tabiiy tilni qayta ishlash uchun til korpusini yaratishda ko'plab omillarni hisobga olish kerak. Muayyan NLP ilovasi uchun til modelini yaratish uchun til korpusini yaratishga to'g'ri keladi. Til korpusini shakllantirishda qanday turdagi hujjatlarni to'plash kerakligini aniqlab olish lozim. Bu korpusdan foydalanmoqchi bo'lgan NLP vazifalariga bog'liq. Misol uchun, agar sizning asosiy maqsadingiz Markaziy Osiyoda o'rta asrlarda qaysi ismlar va joy nomlaridan foydalanganligini aniqlash bo'lsa, siz asosan o'rta asrlar davri hujjatlaridan (adabiyot, yangiliklar, yuridik va boshqalar) iborat korpusni shaklantirishingiz lozim. Boshqa tomondan, agar siz ma'lum vaqt oralig'ida O'rta Osiyodagi urushlar haqidagi o'zgaruvchan his-tuyg'ularni aniqlash uchun model yaratmoqchi bo'lsangiz, siz yangiliklar, blog nashrlari, tvitlar, siyosiy chiqishlar, huquqiy qarorlar va boshqalardan iborat korpusni yaratishingiz kerak. Yoki, agar siz tibbiyot sohasiga tegishli Word2Vec modelini yaratmoqchi bo'lsangiz, Google tomonidan taqdim etilgan maqolalar asosida yaratilgan Word2Vec modeli sizning ehtiyojlaringiz uchun yaxshi ishlamaydi. Shu sababli tibbiy maqolalardan iborat korpusni ishlab chiqishga ehtiyoj tug'iladi.

Yuqorida keltirilgan fikr-mulohazalar natijasida qanday va qancha hajmdagi hujjatlar kerakligini hal qilish lozim. Ba'zi NLP ilovalari uchun, masalan, *so'zlarni joylashtirish modelini (word embedding model)* yaratish uchun, masalan, Naive Bayes tasniflagichini ishlab chiqishdan ko'ra ko'proq ma'lumotlar kerak bo'lishi mumkin. Katta hajmdagi korpusni qurish va saqlash uchun katta xarajatlar talab etiladi. Til modelining korpusga juda yaqin sozlangan bo'lsa, ortiqcha moslashish xavfi ham mavjud. Korpus qanchalik katta bo'lsa, korpusga mos til modelini qayta o'qitish xavfi yuqori bo'ladi.

Keyingi qadamda til korpusini qurish uchun hujjatlarni qanday olishni aniqlab olish kerak. Ma'lumotlarni Vikipediya va yangilik saytlaridan jamlash mumkin. Biroq bunday turdagi ma'lumotlarni korpusga qo'shishdan avval ulardan matnni ajratib olish, tozalash kerak bo'ladi. Sayt hujjatlardan zarur matnlarni olishning 2 ta usuli mavjud:

–*webhose.io, Octoparse, diffbot servislari, Bing yoki Google kabi qidiruv tizimlari;*

–*tabiiy tilga moslashtirilgan dasturiy ta'minot (o'zimiz ishlab chiqamiz).*

Webhose.io (Octoparse, diffbot va boshqalar) kabi xizmatlar Internetdan hujjatlarni olish va ularning mazmunini qirib tashlash va ularni toifalarga ajratish uchun juda yaxshi. Biroq, bunday servislar asosan yangiliklar tasmalari va bloglarga e'tibor qaratishgani sababli, ilmiy maqolalar korpusini yaratmoqchi bo'lsangiz, bu xizmatlar yetarli bo'lmaydi. Agar yuqorida keltirilgan pullik xizmatlardan foydalanmaslikka qaror qilsangiz, Scrapy kabi tizim yordamida o'zingizning skanerlashni amalga oshirishingiz yoki keyingi bosqichda aytib o'tilgan vositalar yordamida skanerlash va qirqish uchun Bing tomonidan taqdim etilgan APIdan foydalanishingiz mumkin. Qabul qilingan hujjatlarning mazmunini chiqarib olish uchun ularning barchasi bir xil formatda, matnli fayllarda bo'lishi lozim. HTML hujjatlari mazmunini chiqarish uchun Beautiful Soup kabi Python

kutubxonalaridan yoki PDF hujjatlari mazmunini chiqarish uchun PDFMiner paketidan foydalanish mumkin.

Til korpusi ma'lum bir tildagi hujjatlardan, masalan, o'zbek tilidagi hujjatlardan iborat bo'lishini istasangiz, o'zbek tilidagi bo'lmagan hujjatlarni olib tashlash uchun tilni aniqlashni amalga oshirishingiz kerak bo'ladi. Ba'zi o'zbek hujjatlarida o'zbekcha bo'lmagan qismlar bo'lishi mumkin va siz nima qilmoqchi ekanligingizga qarab ularni aniqlab, olib tashlashingiz kerak bo'lishi mumkin. Shuningdek, NLP vazifasiga qarab, korpusdan dublikatlarni va deyarli takroriy nusxalarni olib tashlashga to'g'ri keladi. Bu jarayon **de-duping** deb ataladi va buning uchun xeshlash usullaridan foydalanish mumkin.

Yuqoridagi amallarni qo'llash jarayonida ba'zi hujjatlarning ayrim qismlari buzilgan bo'lishi mumkin. Shu sababli ularni aniqlash va tuzatish yoki olib tashlash usulini ishlab chiqish kerak. Buzilgan belgilarni aniqlashning usullaridan biri bu har bir hujjatni tokenizatsiya qilish va har bir belgi lug'atda mavjudligini tekshirishdir.

Qo'yilgan NLP masalasiga qarab til korpusining asosiy tarkibiy qismlari quyidagilar:

Matnlarni jamlash. Korpusni yaratishdagi qiyinchiliklardan biri bu xilma-xil ma'lumotlar to'plamini yig'ishdir. Ma'lumotlar maqsadli domenni to'g'ri qamrab olishi va chuqur tahlilni qo'llab-quvvatlash uchun yetarlicha katta hajmda bo'lishi kerak. Bu mualliflik huquqi cheklovlarini yengib o'tishni, kelishuvlarni muzokaralar olib borishni va maxfiylik muammolarini hal qilishni talab qilishi mumkin.

–Korpus til xilma-xilligini aks ettirishiga ishonch hosil qilish uchun tizimli tanlash usulidan foydalanish;

–Matnlarni tasodifiy, maqsadli yoki tabaqalashtirilgan tanlab olish orqali tanlash.

Korpus strukturasi va hajmi. Korpusning strukturasi to'g'riligi va katta hajmliligi, uning izchilligi saqlash va aniq tahlil qilish uchun juda muhimdir. Standart ko'rsatmalarni ishlab chiqish va ma'lumotlarni tashkil etish bo'yicha eng yaxshi yechimlarni o'rnatish yordam beradi. Ammo korpusning barcha qismlarida izchillikni ta'minlash murakkab vazifa hisoblanadi.

–Hisoblash imkoniyatlari va tadqiqot maqsadlarini hisobga olgan holda korpus hajmini aniqlash;

–Korpusda turli xil til atributlari, jumladan, kamdan-kam uchraydigan yoki kam uchraydigan hodisalar mavjudligiga ishonch hosil qilish.

Ma'lumotlarni annotatsiyalash. Korpusdagi ma'lumotlarni annotatsiyalash, ayniqsa, katta hajmdagi ma'lumotlarni etiketlashda ko'p mehnat va vaqt talab qilishi mumkin. Annotatsiyalarning sifati inson omillariga yoki teglash bo'yicha qoidalarga qarab ham farq qilishi mumkin.

–POS teglash, grammatik tartiblash, NER obyektlarini aniqlash, hissiyotlarni tahlil qilish yoki semantik izohni o'z ichiga olishi mumkin bo'lgan lingvistik annotatsiyalashning tegishli darajasini tanlash;

–Annotatsiyalashni avtomatik, yarim avtomatik yoki qo'lda amalga oshirilishini hal qilish.

Til va domenning o'ziga xosligi. Tilga xos xususiyatlar va domenga xos jargon korpus yaratishda qiyinchiliklarga olib kelishi mumkin. Til yoki domenning o'ziga xos xususiyatlarini tushunish va foydali korpusni yaratishda juda muhimdir. Shuningdek, kam o'rganilgan tillar uchun korpus yaratish kamroq resurslar va cheklangan tadqiqotlarni o'z ichiga olishi mumkin.

Xulosa. Til korpusi tabiiy tilni qayta ishlash tizimining asosidir. Unda gazetalar, romanlar, retseptlar, radio eshittirishlardan tortib teleko'rsatuvlar, filmlar va tvitlargacha bo'lgan turli shakldagi ma'lumotlar bo'lishi mumkin. Til korpuslari mashinada o'qiladigan matnlar yoki nutqlar to'plamlarini o'z ichiga oladi. Korpus odatda *milliardlab so'zlardan* iborat bo'lib, *lingvistik* yoki tabiiy tilida so'zlashuvchilar tomonidan ixtiro qilingan misollardan iborat bo'lmaydi. Korpus tabiiy ravishda yuzaga keladigan *og'zaki* yoki *yozma* manbaalardan foydalanishga asoslangan. Korpus - tilshunoslar, leksikograflar, ijtimoiy olimlar, gumanitar fanlar, tabiiy tillarni qayta ishlash bo'yicha mutaxassislar va boshqa ko'plab sohalarda qo'llaniladi. Shuningdek, dasturiy ta'minotni ishlab chiqishda ishlatiladigan turli til ma'lumotlar bazalarini yaratishda til korpusidan foydalaniladi.

Mashinali o'qitish modellarini o'rgatish, tilni tushunish, diskurs tahlil, tarjimashunoslik, lug'at va semantikaga asoslangan tizimlarni ishlab chiqish kabi NLP vazifalarini hal qilishda til korpusidan

foydalaniladi. Til korpuslari mazmun, maqsad yoki manba kabi turli mezonlar asosida *matnli korpuslar, multimodal korpuslar, parallel korpuslar, Time-Series Corpora* va *annotatsiyalangan/teglangan korpuslar* kabi turlarga ajratiladi. Korpusni tanlash muayyan NLP vazifasiga, tadqiqot maqsadlariga va qo'llash sohasiga bog'liq. Tadqiqotchilar va amaliyotchilar ko'pincha turli NLP loyihalarini hal qilish uchun maxsus korpuslarni ishlab chiqishlari mumkin.

NLPda til korpusining xususiyatlari uni til bilan bog'liq barcha turdagi vazifalar va tadqiqotlar uchun juda foydali vositaga aylantiradi. Til korpusining muhim xususiyatlariga *korpus hajmi, yuqori sifatli ma'lumotlar, tozalangan ma'lumotlar, xilma-xillik, annotatsiya* va *metama'lumotlar* kabilarni keltirish mumkin. Qo'yilgan NLP masalasiga qarab til korpusini ishlab chiqishning asosiy bosqichlari quyidagilar: *matnlarni jamlash, korpus strukturasini ishlab chiqish, ma'lumotlarni annotatsiyalash yoki teglash* va *korpusni yangilash hamda kengaytirish.* Xulosa sifatida til korpuslari NLPdagi muhim vazifalarni hal qilishda muhim ahamiyat kasb etishini qayd etish mumkin.

FOYDALANILGAN ADABIYOTLAR RO'YXATI:

1. Biber, D. (1993). Representativeness in corpus design. Literary and Linguistic Computing, 8(4). https://doi.org/10.1093/llc/8.4.243

2. Lin, P., & Adolphs, S. (2023). Corpus linguistics. In The Routledge Handbook of Applied Linguistics. https://doi.org/10.4324/9781003082644-25

3. Paquot, M., & Gries, S. T. (2021). A Practical Handbook of Corpus Linguistics. In A Practical Handbook of Corpus Linguistics. https://doi.org/10.1007/978-3-030-46216-1

4. B.Elov, R.Alayev (2023). O'zbek tili korpusi va uning imkoniyatlari. O'zbekiston Informatika va energetika mummolari jurnali. O'zbekiston Jurnali. - Toshkent, 2023, - № 2.

5. B.Elov, Sh.Hamroyeva, R.Alayev, Z.Xusainova, U.Yodgorov (2023). 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

6. Shira Wein, Lucia Donatelli, Ethan Ricker, Calvin Engstrom, Alex Nelson, Leonie Harter, & Nathan Schneider. (2022). Spanish Abstract Meaning Representation: Annotation of a General Corpus. Northern European Journal of Language Technology, 8(1). https://doi.org/10.3384/nejlt.2000-1533.2022.4462

7. Brooke, J., Hammond, A., & Hirst, G. (2015). GutenTag: An NLP-driven Tool for Digital Humanities Research in the Project Gutenberg Corpus. NAACL HLT 2015 - 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the 4th Workshop on Computational Linguistics for Literature, CLFL 2015. https://doi.org/10.3115/v1/w15-0705

8. Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings.

9. Markantonatou, S., Sofianopoulos, S., Spilioti, V., Tambouratzis, G., Vassiliou, M., Yannoutsou, O., & Ioannou, N. (2000). Monolingual Corpus-based MT using Chunks. Machine Translation.

10. Genç-Yöntem, E., & Eveyik-Aydın, E. (2022). The compilation of a developmental spoken English corpus of Turkish EFL learners. Research in Corpus Linguistics, 10(1). https://doi.org/10.32714/ricl.10.01.03

11. Knight, D., & Adolphs, S. (2021). Multimodal Corpora. In A Practical Handbook of Corpus Linguistics. https://doi.org/10.1007/978-3-030-46216-1_16

12. Zeroual, I., & Lakhouaja, A. (2022). MulTed: a multilingual aligned and tagged parallel corpus. Applied Computing and Informatics, 18(1–2). https://doi.org/10.1016/j.aci.2018.12.003

13. B.Elov, M.Amirqulov (2022). O'zbek-ingliz tillarining teglangan parallel korpusini yaratish bosqichlari. O'zbekiston: til va madaniyat (Amaliy filologiya), 2022, 5(4).

14. Kobayashi, H., & Saga, R. (2016). Finding division points for a time series corpus based on structural change point detection. Artificial Life and Robotics, 21(2). https://doi.org/10.1007/s10015-016-0271-z

15. Lyu, L., Koutraki, M., Krickl, M., & Fetahu, B. (2021). Neural ocr post-hoc correction of historical corpora. Transactions of the Association for Computational Linguistics, 9. https://doi.org/10.1162/tacl_a_00379

16. Gries, S. Th. (2022). Managing Synchronic Corpus Data with the British National Corpus (BNC). In The Open Handbook of Linguistic Data Management. https://doi.org/10.7551/mitpress/12200.003.0043

17. Hofmann, K., Marakasova, A., Baumann, A., Neidhardt, J., & Wissik, T. (2020). Comparing Lexical Usage in Political Discourse across Diachronic Corpora. Proceedings of the Second ParlaCLARIN Workshop, 11(6).

18. Rodríguez-Fuentes, R. A. (2015). Review of corpus linguistics and linguistically annotated corpora. In Language Learning and Technology (Vol. 19, Issue 3).

19. Agić, Ž., & Ljubešić, N. (2014). The SETIMES.HR linguistically annotated corpus of Croatian. Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014.

20. Mohanty, G., Mishra, P., & Mamidi, R. (2020). Annotated corpus for sentiment analysis in odia language. LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings.

21. Iurato, A. (2022). Learner Corpus Research Meets Chinese as a Second Language Acquisition: Achievements and Challenges. Annali Di Ca Foscari Serie Orientale, 58(1). https://doi.org/10.30687/AnnOr/2385-3042/2022/01/024