

O‘ZBEK VA FORS TILLARI UCHUN PARALLEL KORPUS: LINGVISTIK VA TEXNOLOGIK YONDASHUVLAR

Xusainova Zilola Yuldashevna,
Filologiya fanlari bo‘yicha falsafa doktori
xusainovazilola@navoiy-uni.uz
ToshDO‘TAU

Raxmatxonova Nozimaxon Ismoilxon qizi,
Filologiya va tillarni o‘qitish yo‘nalishi talabasi
nozimaxonrahmatxonova@gmail.com
ToshDO‘TAU

Annotatsiya. Parallel korpuslar zamonaviy kompyuter lingvistikasi, mashina tarjimasini, til o‘rgatish va NLP sohalarida muhim manba hisoblanadi. O‘zbek va forstillari orasidagi madaniy, tarixiy va tilshunoslik aloqalari ushbu tillar uchun parallel korpus yaratish zaruratini yuzaga keltiradi. O‘zbek va forstillari uchun hali to‘liq va sifatli parallel korpus mavjud emas, mavjud manbalar esa lingvistik va texnologik jihatdan yetarlicha boy emas. Ushbu maqola O‘zbek-Fors parallel korpusini yaratish bosqichlari, duch kelinadigan lingvistik va texnologik muammolar hamda ularning yechimlarini yoritishga qaratilgan.

Abstract. Parallel corpora are an important resource in the fields of modern computational linguistics, machine translation, language teaching, and NLP. The cultural, historical, and linguistic connections between the Uzbek and Persian languages make it necessary to create a parallel corpora for these languages. There is no complete and high-quality parallel corpora for the Uzbek and Persian languages, and the existing sources are not sufficiently rich in linguistic and technological terms. This article aims to highlight the stages of creating an Uzbek-Persian parallel corpus, the linguistic and technological problems encountered, and their solutions.

Аннотация. Параллельные корпуса являются важным ресурсом в областях современной компьютерной лингвистики, машинного перевода, изучения языков и обработки естественного языка. Культурные, исторические и языковые связи между узбекским и персидскими языками обуславливают необходимость создания параллельного корпуса для этих языков. До сих пор не создан полный и качественный параллельный корпус для узбекского и персидского языков, а имеющиеся источники недостаточно богаты в лингвистическом и технологическом плане. Целью данной статьи является освещение этапов создания узбекско-персидского параллельного корпуса, возникших лингвистических и технологических проблем и путей их решения.

Kalit soʻzlar. *Korpus, parallel korpus, oʻzbek tili, fors tili, NLP, statistik tahlil, raqamlashtirish, mashinali oʻqitish, lingvistik muvofiqlik*

Kirish. Tabiiy tilni kompyuter yordamida tadqiq qilish va qayta ishlash toʻrt jihatni: formallashtirish, algoritmlash, dasturlash va amaliyotga qoʻllashni oʻz ichiga oladi[1]. Tabiiy tilni qayta ishlash (NLP)ga asoslangan dasturiy ilovalar(axborot tizimlari)ni ishlab chiqish uchun NLP tizimini mavjud maʼlumotlarni oʻrganishini (oʻqitishini) taʼminlash kerak. Ushbu amalni til korpusi vositasida amalga oshirish mumkin. Til korpusi elektron shaklda taqdim etiladigan katta hajmli va strukturlangan matnlar toʻplami sifatida qaraladi. Til korpusi yozma yoki ogʻzaki materialni ifodalab, NLP tizimini mavjud resurslarni oʻrganishi uchun lingvistik tahlilni amalga oshirish lozim. Til korpusi tabiiy tilni qayta ishlash tizimining asosidir. Unda gazeta, roman, retsept, radioeshittirishdan tortib telekoʻrsatuv, film va tvitlargacha turli shakldagi maʼlumotlar boʻlishi mumkin[2].

Korpus ikki yoki undan ortiq til doirasidagi maʼlumotlar bazasi boʻlib, undan koʻplab maqsadlarda foydalanish mumkin. Shuningdek, korpus tillar oʻrtasidagi muloqot vositasi hisoblanib, bu tizim yirik elektron koʻrinishdagi matnlar jamlanmasidir. Korpus lingvistikasining *statistik* va *hisoblash* usullari turli tillarning **grammatik**, **leksik** va **semantik** xususiyatlarini tizimli ravishda tahlil qilish imkonini beradi. U katta hajmdagi yozma yoki ogʻzaki matnlarni tahlil qilishga asoslanadi. Korpus aniq maqsad uchun **yigʻilgan** va **raqamlashtirilgan** matnlar toʻplami boʻlib, u turli xil lingvistik tahlillar uchun ishlatiladi[3]. Parallel korpuslar bir tildan ikkinchi tilga tarjima qilingan matnlarni oʻz ichiga olganligi sababli, ular **lingvistik va hisoblash tahlili uchun muhim manba hisoblanadi**.

Shuningdek soʻz turkumlariga xos kategoriyalar, grammatik maʼnoning ifodalanishi, soʻz yasash tizimiga oid maʼlumotlar bera olishi bilan ham eʼtiborga molikdir. Parallel matnlar korpusi chogʻishtirma aspektidagi tadqiqotlarda oʻziga xos oʻrin tutadi. Parallel korpuslar yordamida bir soʻz, jumla, abzats, supersintaktik butunliklarning turli tillardagi variantlarini bilish mumkin. Parallel korpuslar madaniyatlararo muloqot keng yoyilgan hozirgi davr uchun muhim boʻlgan maʼlumotlar bazasidir. Parallel korpuslar orqali turli til muhitdagi, madaniyatlardagi universalialar hamda tillarning oʻziga xos mental xususiyatlari, realiya va lakunar birliklarni aniqlash mumkin boʻladi. Parallel matnlar korpusi avtomatik tarjima rivoji uchun ham xizmat qiladi, kompyuter leksikografiyasining taraqqiyotini taʼminlaydi[4]. Parallel matnlar korpusi turli tillar asosida xilma-xil madaniyatlarni

solishtirish, madaniy munosabatlarni ifodalovchi leksik birliklar mazmun-mohiyatini o'zlashtirish imkoniyatini beradi. Parallel korpusning NLPdagi roli turlicha, xususan, NLPning ko'plab vazifalarini hal qilishda til korpuslaridan foydalaniladi:

– **Mashinali o'qitish modellarini o'rgatish.** Turli NLP ilovalari, jumladan hissiyotlarni tahlil qilish, matn tasnifi, mashina tarjimai 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 tarjimai modellarini yaratish imkonini beradi.

– **Qoidalarga asoslangan tizimlar.** Til korpuslari tilshunoslar va NLP mutaxassisleri tomonidan til qoidalar 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 sohalarga oid bilimlar manbayi hisoblanadi. Chunki ular muayyan mavzular yoki sohalarga xos bo'lishi mumkin. Huquqiy hujjatlarni o'rganish, tibbiy ma'lumotlarni qayta ishlash va muayyan sohalar uchun yaratilgan chatbotlar kabi ilovalar bunga yaqqol misol bo'ladi[5].

O'zbek va fors tillarining ta'limiy parallel korpusi hozircha yaratilmagan bo'lib, biroq ikkala tilning ham boshqa tillar bilan aloqadorlikda yaratilgan korpuslari mavjud, masalan, o'zbek-ingliz, o'zbek-turk va fors-ingliz, fors-fransuz, fors-rus, fors-nemis, fors-arab kabilar. Ko'rinib turibdiki, fors tilidagi korpuslar o'zbek tiliga nisbatan son jihatdan ustun. O'zbek tili bilan bevosita aloqador korpuslar sekin-asta rivojlanib bormoqda. O'zbek-fors tillari o'rtasidagi korpus ham ana shu harakatlarning namunalaridandir. Bu shakllantirilayotgan korpusda ikkala tilning barcha xususiyatlari, ya'ni leksik, sintaktik, semantik, morfologik tomonlari parallel ravishda o'rganildi. Umuman olganda, fors tili IX-XIV asrlarda turkiy davlatlarda, ayniqsa, Xuroson, Samarqand, Buxoro, Xorazm kabi Markaziy Osiyoning yetakchi shahar va viloyatlarida keng iste'molda bo'lgan. Ana shu davrda turkiy tilga fors tilidan ko'plab so'zlar o'zlashgan bo'lib, bugungi kunda o'zbek tilining lug'at tarkibida forsha o'zlashmalar taxminan 30-40% atrofida bo'lishi mumkin.

Korpusni yaratish davomida bir nechta qiyinchiliklarga duch kelinadi. Chunki kam resursli tillar bilan bog'liq korpuslarda ma'lumotlar bazasini yig'ish biroz murakkab, tarjima jarayonida ko'proq inson omilidan foydalanishga to'g'ri keladi. Yana bir masalani sabab qilib ko'rsatish mumkinki, u ham bo'lsa, o'zbek tili agglyutinatив tillar oilasiga, fors tili esa flektiv tillar oilasiga kiradi. Quyidagi jadval orqali batafsilroq ma'lumotga ega bo'lamiz.

Til tizimi xususiyati	O'zbek tili	Fors tili
Alifbo	Lotin (ba'zi manbalarda Kirill)	Arab yozuvi

Fonetika	Unli tovushlar: 6 ta: a, e, i, o, u, o'. Undosh tovushlar: 24 ta.	Unli tovushlar: 6 ta. Yozuvda qisqa unlilar ifodalanmaydi (e, a, o'): کتاب, شتر, بهار kabi. Undosh tovushlar 23 tadir.
Morfologiya	Agglyutativ	Flektiv
Fe'l zamonlari	3 ta (Hozirgi, O'tgan, Kelasi)	Asosan 2 ta: o'tgan zamon va hozirgi-kelasi zamon. (O'tgan zamonning o'zi ham ingliz tili kabi bir nechta ichki turlarga bo'linadi, masalan, o'tgan zamon davom fe'li, o'tgan zamon natijali fe'li, uzoq o'tgan zamon fe'li kabi)
So'z tartibi	SOV: Men ertaga universitetga boraman.	SOV: من فردا به دانشگاه می روم
Fe'l shakllari	Fe'l + yordamchi so'zlar: sifatdosh(kelgan, ishlaydigan), ravishdosh(kulib, borgach), harakat nomi(o'qimoq, salomlashish)	Fe'l va uning o'zgaruvchan shakllari: sifatdosh(نشسته, آمده), ravishdosh(maxsus qo'shimchasi yo'q) – sifatdosh shaklidagi fe'l bilan ifodalanadi, gapda ifodalangan mazmun orqali ajratib olinadi
So'z yasash	Qo'shimchalar va kompozitsiya usuli orqali: tekislamoq, terimchi, hozirjavob	Qo'shimchalar va birikish usuli orqali: کتابخانه، همراه، کارگر، دوربین
Kelishiklar	6 ta (bosh kelishik shakli yo'q, qaratqich, tushum, jo'nalish, o'rin-payt va chiqish)	Kelishik kategoriyasi yo'q, ma'lum bir predloglar orqali ifodalanadi: به، در، از، با، تا، برای، بر، را
Predlog/yordamchi birliklar	Pregloglar ham yordamchi birliklar ham ko'p	Predloglar ko'p
Yozuv tizimi	Chapdan o'ngga: → Mening uyim Pushkin ko'chasida joylashgan.	O'ngdan chapga: ← خانه ی ما در خیابان پوشکین است

Til korpusini yig'ishdagi eng muhim omillardan biri sifatida manba, ya'ni ma'lumotlar bazasini to'plashni olishimiz mumkin. Korpusning hajmi qancha yirik bo'lsa, undan foydalanish koeffitsiyenti shuncha baland bo'ladi. Faqatgina hajm bilan cheklanib qolmaslik zarur, chunki korpusga kiritiladigan ma'lumotlar turli xil bazalardan to'planadi, masalan, rasmiy qonun yoki hujjatlar, badiiy asarlar, ilmiy tadqiqotlar va hokazo. Manbalarni to'plash jarayonida tarjimaga ham e'tibor berib boriladi, matnlar sifatining yuqori bo'lishi ishning tizimli ravishda o'sishiga yordam

beradi. Yetarlicha manba to'planib bo'lgandan so'ng, ishning ikkinchi bosqichiga o'tiladi. Tozalash (dataset) bosqichi yordamida matnlar UTF-8 formatida shakllantiriladi. Ma'lumotlar to'plamini juda ko'p xatolari bo'ladi va qayta-qayta tozalashdan o'tadi. Avtomatik tarjima yordamida yig'ilgan ma'lumotlar inson omili yordamida ham bir necha bor tekshiriladi. Shundan so'ng matnlar keraksiz belgi va teglardan tozalanadi. Bu jarayonni Python kodlari yordamida avtomatik ravishda amalga oshirish mumkin. Keyingi muhim qadamlardan biri parallel bo'lmagan jummalarni transliteratsiya qilishdir. Korpusdagi har bir jummalarni kerakli tilga transliteratsiya qilish lozim bo'ladi. Maxsus dastur va algoritmlar yordamida avtomatik transliteratsiya usulidan foydalaniladi, chunki qo'lda transliteratsiya qilish juda ko'p vaqtni oladi. Sifat nazoratini tekshirish orqali korpusimizning qiymatini belgilab olamiz. Ba'zida keraksiz belgilar qolib ketgan bo'lishi mumkin yoki tarjima jarayonida matnlar bir-biriga aralashib ketish hollari kuzatiladi. Bunday holda keraksiz belgilar o'chiriladi, aralash matnni ajratib olinadi va to'g'rilanadi. Bir-biriga ma'nosi to'g'ri kelmaydigan yoki umuman xato tarjimalar bor yoki yo'qligini ko'rib chiqish kerak. Masalan: من مدرسه رفتم

✓ Men maktabga bordim.

✗ Men universitetdan qaytdim.

Sifat nazorati to'liq tekshirilib bo'lingandan so'ng, korpusga biroz qo'shimchalar kiritish mumkin, masalan, bironta gap keltirilganda uning aynan tarjimasi bilan birgalikda unga teng keluvchi muqobil variantlarni ham kiritiladi va shu tariqa korpusning ichki manbasi boyishiga sezilarli hissa qo'shiladi. Buning uchun esa ko'proq badiiy asarlardan namunalar kiritilishi lozim.

Xulosa. Parallel korpus tilshunoslik, tarjima va sun'iy intellekt sohalarida muhim ahamiyat kasb etadi. Tarjimonlar yoki til o'rganmoqchi bo'lgan havaskorlar uchun manfaatli bo'lib, ular tarjima jarayonida nafaqat gap tarjimalarini balki o'sha so'zga to'g'ri kelishi mumkin bo'lgan muqobil variantlarni ham ko'rib chiqish imkoniga ega bo'ladilar. Tilshunoslar esa o'zbek va fors tillari o'rtasidagi grammatik va leksik munosabatlarni chuqur tahlil qila olishadi. Dasturchilar yoki AI muhandislari uchun avtomatik tarjima tizimlarini rivojlantirishda asosiy ma'lumotlar bazasi sifatida ishlatiladi. Keyinchalik uni rivojlantirib borishimiz natijasida avtomatlashtirilgan tarjima tizimi yaratilib, forsha kontentlar uchun o'zbekcha subtitrlar ishlab chiqilishi ham maqsad qilingan. Parallel ko'rinishda taqdim etiladigan subtitrlar fors tilini o'rganayotganlar uchun qulaylik yaratadi.

Foydalanilgan adabiyotlar:

1. Elov B.B., Xusainova Z.Y. Katta til modellari / “O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi”. 2024. Vol 2(6). – P. 78-92.
2. Xusainova, Z., & Yangibayeva, S. (2024). Til korpusi turlari. Uzbekistan: Language and Culture, 2024. Vol 2(6). – P. 54-64.
3. Elov, B., & Xusainova, Z. (2024). Til korpuslarini lingvistik teglash bosqichlari. // Computer Linguistics: problems, solutions, prospects, 1(1).
4. Akramdjanova, M. (2024). Parallel matnlar korpusini yaratishning amaliy va nazariy ahamiyati. *Наука и инновация*, 2(34), 34-37.
5. .B.Elov, A.U.Yuldashev, U.S.Yodgorov. Til korpusi turlari va umumiy xususiyatlari / Xorazm Ma'mun akademiyasi axborotnomasi: ilmiy jurnal.- 2024. №12/4 (121). – P.118-125.
6. Elov, B. B., Hamroyeva, S., & Xusainova, Z. (2022). NLP (tabiiy tilga ishlov berish) ning vazifalari va zamonaviy yondashuvlar. TerDU, filologik tadqiqotlar: til, adabiyot, ta'lim, 5-6.
7. B. Elov, Sh. Hamroyeva, R. Alayev, Z. Xusainova, U. Yodgorov – O‘zbek tili korpusi matnlarini qayta ishlash usullari – Raqamli Transformatsiya va Sun'iy Intellect ilmiy jurnali, October 2023 – Volume 1, Issue 3.
8. Ro‘Ziyev H. B., Xodjayeva N. T. Matnlarning parallel korpuslarini tuzish muammolari //Academic research in educational sciences. – 2021. – T. 2. – №. 9. – C. 515-522.
9. Sobirova N. G. Q. Korpus lingvistikasi va parallel korpuslar tavsifi //Academic research in educational sciences. – 2022. – T. 3. – №. 5. – C. 1442-1450.