UO‘K 81

O‘ZBEK-INGLIZ PARALLEL KORPUSINING TUZILISHI VA QO‘LLANISH SOHASI

Karimov Rustam Abdurasulovich,
Buxoro davlat universiteti
Fakultetlararo chet tillar kafedrasi dotsenti
ORCID: 0009-0000-0385-5256
rustamkarimov5594@gmail.com

Annotatsiya. *Ushbu maqolada o‘zbek-ingliz parallel korpusining tuzilishi, yaratish metodologiyasi, lingvistik va texnologik xususiyatlari hamda qo‘llanish sohalari tahlil qilinadi. Tadqiqotda matnlarni yig‘ish, tozalash, tokenizatsiya, gap va so‘z darajasida moslashtirish (alignment), metama’lumot hamda lingvistik annotatsiya jarayonlari yoritiladi. Shuningdek, parallel korpus yaratishda qo‘llaniladigan zamonaviy dasturiy vositalar va saqlash formatlarining imkoniyatlari ko‘rib chiqiladi. Tadqiqot natijalari o‘zbek-ingliz parallel korpusining tarjimashunoslik, korpus lingvistikasi, mashina tarjimasi, leksikografiya va til ta’limidagi ahamiyatini ko‘rsatadi.*

Kalit so‘zlar: *parallel korpus, o‘zbek-ingliz tillari, alignment, lingvistik annotatsiya, korpus lingvistikasi, mashina tarjimasi.*

СТРУКТУРА И СФЕРЫ ПРИМЕНЕНИЯ УЗБЕКСКО-АНГЛИЙСКОГО ПАРАЛЛЕЛЬНОГО КОРПУСА

Аннотация. *В статье рассматриваются структура, методология создания, лингвистические и технологические особенности узбекско-английского параллельного корпуса, а также основные сферы его применения. Освещаются процессы сбора и очистки текстов, выравнивания (alignment), добавления метаданных и лингвистической аннотации. Анализируются современные параллельные корпуса и программные инструменты, применяемые при их создании, а также вопросы подготовки корпусных данных для машинного обучения. Определяется значение узбекско-английского параллельного корпуса для языкознания, переводоведения, лексикографии, машинного перевода и образования.*

Ключевые слова: *параллельный корпус, узбекский и английский языки, выравнивание, лингвистическая аннотация, корпусная лингвистика, машинный перевод.*

THE STRUCTURE AND APPLICATION AREAS OF THE UZBEK-ENGLISH PARALLEL CORPUS

Abstract. *This article examines the structure, creation methodology, linguistic and technological features, and application areas of the Uzbek-English parallel corpus. It discusses the main stages of corpus development, including text collection, cleaning, alignment, metadata, and linguistic annotation. The study also considers modern parallel corpora and software tools used for corpus construction, as well as methods of preparing corpus data for machine learning. The findings highlight the importance of the Uzbek-English parallel corpus for linguistics, translation studies, lexicography, machine translation, and language education.*

Keywords: *parallel corpus, Uzbek-English languages, alignment, linguistic annotation, corpus linguistics, machine translation.*

Kirish. O‘zbek-ingliz parallel korpusining tuzilishi va uning ilmiy ahamiyati. Parallel korpuslar ikki yoki undan ortiq tilda bir xil mazmundagi matnlar juftligidan tashkil topgan lingvistik resurs hisoblanadi. O‘zbek-ingliz parallel korpusining samarali ishlashi va ilmiy qiymati uning to‘g‘ri tuzilishiga bevosita bog‘liq. Ushbu korpusning tuzilishi bir necha asosiy komponentlarni o‘z ichiga oladi: matnlarni yig‘ish, ularni tekislash (alignment), metadata qo‘shish va korpusga tegishli vositalarni yaratish.

Metodlar. Matnlarni yig‘ish. Korpusga kiritiladigan matnlar turli manbalardan olinishi mumkin: adabiy asar, ilmiy maqola, huquqiy hujjat, yangiliklar materiali, xalqaro tashkilotlarning hisoboti va texnik yo‘riqnomalar kiradi. Matnlarni tanlashda matn turi, janri va mavzusi bo‘yicha

muvozanat(reprezentativlik)ga erishish muhim. Masalan, umumiy leksikani qamrab olish uchun adabiy matnlar, terminologik xilma-xillikni ta’minlash uchun ilmiy va texnik matnlar kiritilishi kerak. Har bir matn juftligi bir tildan boshqasiga sifatli tarjima qilingan bo‘lishi parallel korpus matnlariga qo‘yilgan eng asosiy shartdir.

Moslashtirish (Alignment). Matnlar yig‘ilgandan so‘ng ularni moslashtirish jarayoni amalga oshiriladi. Moslashtirish matnning segmentlari – gap, ibora yoki so‘zlarni bir-biriga moslashtirishni anglatadi.

N.Matyakubovaning ta’kidlashicha, mosl

ashtirish ilmiy jihatdan murakkab jarayon bo‘lib, asosan uch darajada amalga oshiriladi:

1. *Gap darajasida moslashtirish* eng keng tarqalgan usul bo‘lib, manba tildagi har bir gap tarjima tildagi mos keluvchi gap yoki gaplar guruhiga moslashtirishni o‘z ichiga oladi. Gap chegaralarini aniqlashda tinish belgilari, maxsus dasturlar yoki statistik modellardan foydalaniladi.

2. *So‘z darajasida moslashtirish* mashina tarjimasi tizimlari uchun muhim jarayon har bir so‘zni uning tarjima analogiga moslashtirishni talab qiladi. So‘z darajasidagi moslashtirish ko‘pincha lug‘atlar va statistik modellarga asoslanadi. Bu bosqich, ayniqsa, o‘zbek va ingliz tillari kabi sintaktik jihatdan farqli tillar uchun qiyin.

3. *Paragraf yoki qism darajasida moslashtirish*da matnning katta qismlari, masalan, bo‘lim yoki boblar o‘rtasidagi moslik aniqlanadi. Bu umumiy struktura va kontekstni saqlashga yordam beradi.

O‘zbek va ingliz tili korpusida mantlarni moslashtirish masalasi N.Matyakubovaning bir necha maqolalarida batafsil tahlil qilingan. Shu bois bu masalaga qisqa to‘xtalish yetarli deb hisoblaymiz.

Metama’lumot (ekstralingvistik) qo‘shish. Korpusning ilmiy qiymatini oshirish uchun har bir matn haqida batafsil metama’lumot kiritilishi shart. Bu ma’lumotlar matnning manbayi, muallifi, nashr qilingan sanasi, matnning janri (ilmiy, badiiy, publisistik va h.k.), mavzusi, tarjimonning ismi va tarjima sifati haqidagi ma’lumotlardan iborat bo‘ladi. Metama’lumot foydalanuvchiga ma’lumotlarni filtrlash va izlash imkonini beradi, tadqiqotchi uchun esa turli matn turlarini qiyosiy o‘rganishga zamin yaratadi.

Korpusga tegishli vositalar. Yuqorida aytib o‘tilgan komponentlar bilan bir qatorda, korpusdan samarali foydalanish uchun bir qator vositalar yaratilishi lozim. *Qidiruv tizimi* – foydalanuvchiga so‘z, ibora yoki grammatik konstruksiyalarni izlash imkonini beruvchi interfeys. Konkordans – tanlangan so‘zning turli kontekstlardagi qo‘llanilishini ko‘rsatib beruvchi vosita. Statistik analiz vositalari – so‘z chastotasi, n-gramm va boshqa statistik ma’lumotlarni hisoblaydigan dasturlar. Mazkur vositalar parallel korpusda qidiruv sifatini yaxshilashga asos bo‘ladi.

Zero, o‘zbek-ingliz parallel korpusining tuzilishi faqat matnlar yig‘ishdan iborat emas, balki ularni lingvistik talablarga mos ravishda moslashtirish, boyitish va foydalanuvchi uchun qulay vositalar bilan ta’minlash kabi murakkab jarayonlarni o‘z ichiga oladi. Bunday to‘g‘ri tuzilgan korpus tilshunoslik, mashina tarjimasi, leksikografiya va boshqa ko‘plab sohalarda fundamental tadqiqotlar uchun asosiy manba bo‘lib xizmat qiladi.

Korpusning ilmiy-metodologik asoslari va sifati. Korpusni yaratishda metodologik yondashuvning o‘rni muhim. Bunga, birinchi navbatda, korpusning muvozanatlanganligi (rus. reprezentativligi) masalasi kiradi. Reprezentativlik (yoki mantlar mazmuni balansi) korpusga kiritiladigan matnlarning janr, mavzu va uslubiy jihatdan bir tekis taqsimlanishini anglatadi. Masalan, korpus faqat badiiy adabiyotdan iborat bo‘lsa, u texnik yoki ilmiy atamalarni aks ettirmaydi. Shuning uchun ilmiy maqola, huquqiy hujjat, ommaviy axborot vositalari matni kabi turli manbalarni kiritish talab etiladi. Bu korpusning umumiy lisoniy manzarasi haqida to‘liqroq ma’lumot berishiga xizmat qiladi. Ikkinchidan, korpusning annotatsiyasi juda ahamiyatli. Annotatsiya matnlarga qo‘shimcha lingvistik ma’lumotlarni kiritish jarayoni bo‘lib, Sh.Hamroyeva annotatsiyaning quyidagi ko‘rinishlarini farqlaydi:

Morfologik annotatsiya: har bir so‘zning qaysi so‘z turkumi ekanligi (masalan, ot, fe’l, sifat), uning leksema shaklini (lemma) va grammatik xususiyatlarini (son, egalik, kelishik, zamon, shaxs-son va boshqalar) belgilash. Bu ayniqsa, o‘zbek tili uchun muhim, chunki o‘zbek tili agglutinativ til bo‘lib, qo‘shimchalar so‘z ma’nosini sezilarli darajada o‘zgartiradi. O‘zbek tili birliklarini morfologik teglash masalasi Sh.Hamroyeva, N.Xudoyberganovning ishlarida atroflicha yoritilgan.

Sintaktik annotatsiya gaplarning sintaktik tuzilishini, so‘z birikmasi va so‘zlar orasidagi bog‘liqlikni ko‘rsatishdir. Bu mashina tarjimasi tizimlarini takomillashtirish va tilning sintaktik xususiyatlarini chuqur o‘rganish uchun ahamiyatli annotatsiya turidir. Sintaktik annotatsiya masalasi O.Abdullayevaning bir necha maqolalarida atroflicha tahlil qilingan.

Semantik annotatsiya so‘zning ma’nosini kontekstga qarab teglashdir. Masalan, annotatsiyaning bu turida so‘zning ko‘p ma’noliligi muammosini hal qilinishi mumkin. O‘zbek tili birliklarini teglash, ko‘p

ma’nolilik va polifunksionallikni aniqlash masalalari D.Axmedova, Sh.Gulyamova, X.Axmedovalarning tadqiqotlarida monografik planda yoritilgan. Shu sababli korpusni annotatsiyalash masalalariga batafsil to‘xtalmaymiz.

Bundan tashqari, korpusni yangilash va kengaytirish mexanizmlari ham doim diqqat markazda bo‘lishi lozim. Til doimiy ravishda o‘zgarib turadi, yangi so‘zlar va iboralar paydo bo‘ladi. Shuning uchun korpus turg‘un resurs emas, balki dinamik (doimiy rivojlanadigan) bo‘lishi kerak. Korpusni muntazam ravishda yangi, dolzarb matnlar bilan to‘ldirib borish uning ilmiy va amaliy ahamiyatini saqlab qolishga yordam beradi.

O‘zbek-ingliz parallel korpusining ahamiyati va qo‘llanish sohalari. O‘zbek-ingliz parallel korpusi tilshunoslik va kompyuter lingvistikasi sohalarida fundamental ahamiyatga ega bo‘lgan noyob resurs bo‘lib xizmat qilishi mumkin. U ikki xil til tizimiga ega bo‘lgan tillarni qiyosiy o‘rganish, til texnologiyalarini rivojlantirish va amaliy masalalarni yechish uchun mustahkam asos yaratadi. O‘zbek-ingliz parallel korpusining ahamiyati va qo‘llanish sohalari bir necha asosiy yo‘nalishlarda namoyon bo‘ladi. Korpus o‘zbek va ingliz tillarining leksik, grammatik va sintaktik xususiyatlarini qiyosiy tahlil qilish imkonini beradi. Bu orqali ikki tilning farqli va o‘xshash jihatlarini aniqlash, tarjima ekvivalentlari qanday shakllanishini o‘rganish mumkin. Masalan, o‘zbek tilidagi affiksatsiya (qo‘shimcha) yordamida hosil bo‘ladigan ma’nolarning ingliz tilida qanday sintaktik konstruksiya bilan ifodalanishini aniqlash misol bo‘ladi. Kuzatishlarimiz asosida umumlashtirish mumkinki, korpus tarjima nazariyasi va amaliyotini o‘rganish uchun bebaho manbadir. U tarjimonlarning ishini tahlil qilish, tarjima strategiyalarini aniqlash va turli matn turlari uchun eng mos keluvchi tarjima yechimlarini topishga yordam beradi. Ingliz-o‘zbek parallel korpuslari statistik va neyron mashina tarjimasi tizimlarini yaratishda asosiy manba bo‘lib xizmat qiladi. Ikki til o‘rtasidagi tarjima qoidalari va ehtimolliklari korpusdagi juft matnlar asosida o‘rganiladi. Qiyosiy tilshunoslik tadqiqotlarida morfosintaktik farqlar, leksik ekvivalentlik va tarjima strategiyalari tahlil qilinadi. Shuningdek, ko‘p tilli axborotni izlash va lug‘atshunoslik sohalarida ham bu korpuslar qimmatli ma’lumot bazasi vazifasini o‘taydi. Uning yordamida yangi lug‘at tuzish, mavjud lug‘atlarni boyitish va terminologik baza yaratish mumkin.

Parallel korpus ikki tilli lug‘atlar yaratish yoki mavjudlarini takomillashtirish uchun zarur. Korpus yordamida so‘zlarning turli kontekstlardagi ma’nolari, iboralarning tarjimasi va so‘z birikmalarining ekvivalenti aniqlanadi. Bu lug‘atlarning aniqligi va dolzarbligini oshiradi. Bundan tashqari, ilmiy va texnik matnlar asosida yaratilgan korpuslar soha terminlarining o‘zbekcha va inglizcha ekvivalentlarini aniqlash va tartibga solishda ahamiyatli. Bu texnik hujjatlarni tarjima qilishda yagona terminologik bazadan foydalanishga yordam beradi.

Parallel korpus ta’limda ham ahamiyati vosita sanaladi. Korpus talaba va til o‘rganuvchilari uchun qimmatli manba bo‘lib, u real matnlarda so‘zlar va grammatik qoidalar qanday qo‘llanishini ko‘rsatib beradi. O‘quvchilar o‘zbek va ingliz tilidagi gaplarni solishtirib, tarjima amaliyotini yaxshilashi mumkin. Tarjimon tarjima qiyinchiliklariga duch kelganda, korpusda shunga o‘xshash gaplar qanday tarjima qilinganini ko‘rib, o‘z ishi uchun eng yaxshi yechimni topishi mumkin. Bu jarayon tarjima sifatini oshirishga va vaqtni tejashga yordam beradi. Xulosa qilib aytganda, o‘zbek-ingliz parallel korpusi nafaqat nazariy tilshunoslikning muhim vositasi, balki amaliy til texnologiyalari va ta’lim sohasini rivojlantirishga xizmat qiluvchi strategik resursdir. Uning mavjudligi o‘zbek tilining raqamli makonda o‘z o‘rnini mustahkamlashi va xalqaro miqyosdagi ma’lumot almashinuvida faol ishtirok etishi uchun zamin yaratadi.

Hozirda ingliz-turk parallel korpuslari mavjud va ular lingvistik tadqiqotlar hamda til texnologiyalari uchun muhim resurs vazifasini bajarib kelmoqda. Ikki til o‘rtasidagi tipologik farqlar (ingliz tili analitik til, turk tili esa agglyutinativ til) sababli bu korpuslar tarjima hodisalarini o‘rganish va tabiiy tilni qayta ishlash (NLP)ning ilg‘or ilovalarini yaratish uchun ahamiyatli. Bir qancha loyiha va tadqiqotlar bu korpuslarni yaratish va ularning imkoniyatlariga qaratilgan. Quyida ba’zi korpuslarni tahlil qilamiz.

MaCoCu-tr-en katta hajmli veb-korpus, u turk va kipr yuqori darajadagi domenlarni skanerlash va ikkala tilda ham mazmuni mavjud bo‘lgan veb-saytlarni aniqlash orqali yaratilgan. parallel gaplarni ajratib olish va moslashtirish uchun ishlatilgan *bitextor*, yuqori sifatni ta’minlash uchun ma’lumotlarni tozalash va dublikatlarni olib tashlash uchun esa BicleanerAI qo‘llanilgan. Bu turdagi korpuslar ko‘pincha zamonaviy neyron mashina tarjimasi modellarini o‘qitish uchun ishlatiladi.

Akademik va maxsus soha korpuslari tadqiqotchilar ma’lum sohalar uchun kichikroq, ixtisoslashgan korpuslarni yaratishgan. Masalan, dissertatsiya annotatsiyalaridan tuzilgan korpus akademik tarjimalarga e’tibor qaratadi. Bu umumiy maqsaddagi ma’lumotlarning maxsus qo‘llanish uchun yetarli emasligini hal etishga yordam beradi; leksikografiya va terminologiyani boshqarishda ahamiyatli.

Parallel daraxt banklari (treyebanks): ba’zi ilg‘or tadqiqotlar faqat gapga moslashtirilgan matnlarni emas, balki parallel daraxt bankini yaratishga ham qaratilgan. Bu gaplar uchun parallel sintaktik daraxtlarni qo‘lda yaratishni o‘z ichiga oladi, sintaksisga asoslangan mashina tarjimasi tizimlarini ishlab chiqish va chuqur qiyosiy lingvistik tahlilda amaliy ahamiyat kasb etadi.

KazParC asosan qozoq tiliga yo‘naltirilgan korpus bo‘lsa-da, KazParC loyihasi ingliz va turk, shuningdek, rus tillari uchun ham parallel matnlarni o‘z ichiga olgan ko‘p tilli korpusga misol bo‘la oladi. Bu turli turkiy va Yevropa tillari o‘rtasida mashina tarjimasiga qaratilgan yirik loyihalarda turk tilining tez-tez kiritilishini ko‘rsatadi.

Sanab o‘tilgan korpuslar mashina tarjimasi tizimlarini o‘qitish, qiyosiy lingvistik tadqiqotlar o‘tkazish, ikki tilli lug‘atlarni ishlab chiqish va til o‘rganish uchun vositalarni yaratish kabi turli xil ilovalar uchun zaruriy manba vazifasini bajaradi. Bunday korpuslarni yaratish dinamik faoliyat bo‘lib, turli soha ehtiyojlarini qondirish, ilg‘or tozalash (text cleaning) hamda annotatsiya texnikalari orqali ma’lumot sifatini yaxshilash maqsadida yangi resurslar ishlab chiqilmoqda.

Parallel korpusining lingvistik va texnologik xususiyatlari. Parallel korpus ikki yoki undan ortiq tildagi bir xil mazmundagi matnlarni o‘z ichiga olgan lingvistik resurs bo‘lib, u bir qator lingvistik va texnologik xususiyatlarga ega. Bu xususiyatlar korpusning tuzilishi, ishlanishi va qo‘llanilishi uchun muhim ahamiyat kasb etadi.

Xulosa shuki, o‘zbek-ingliz parallel korpusi lingvistik va texnologik jihatdan murakkab, ammo beqiyos ahamiyatga ega bo‘lgan resursdir. Uning to‘g‘ri tuzilishi, muvozanatlangan va reprezentativ matnlar bazasiga ega bo‘lishi, sifatli moslashtirish hamda annotatsiya jarayonlari nafaqat nazariy tilshunoslik, balki mashina tarjimasi, lug‘atshunoslik va ta’lim sohalari uchun ham keng imkoniyatlar yaratadi. Korpusning mavjudligi o‘zbek tilining raqamli makondagi mavqeyini mustahkamlash, uni ingliz tili bilan qiyosiy o‘rganish va zamonaviy til texnologiyalarini rivojlantirishga xizmat qiladi. Shu bois parallel korpusni muntazam yangilash, boyitish va texnologik jihatdan takomillashtirish ilmiy va amaliy faoliyatning dolzarb yo‘nalishlaridan biri sanaladi.

ADABIYOTLAR:

1. Matyakubova, Noila & Khamroeva, Shahlo & Adali, Eşref & Rajabovich, Mengliyev. (2023). Algorithm of Creating the “Uzbek-English Aligner” Program. 01-05. 10.1109/UBMK59864.2023.10286703;

2. Khamroeva, Shahlo & Abdullayeva, Oqila & Matyakubova, Noila & Rakhmanova, Azizakhon. (2024). Alignment of Phrases in Uzbek-English Languages. 1830-1833. 10.1109/PIERE62470.2024.10805057.

3. Karimov Rustam Abdurasulovich, Mengliev Bakhtiyor Rajabovich (2019). The Role of the Parallel Corpus in Linguistics, the Importance and the Possibilities of Interpretation. International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-5S3 July 2019.

4. Elov, Botir & Abdullayeva, Oqila. (2024). O‘zbek tili korpusini sintaktik teglash masalasi. – 1-son.

5. Gulyamova, Shakhnoza. (2020). The problem of polyfunctionality in the traditional and computational linguistics // Middle European Scientific Bulletin. – Vol. 5. 10.47494/mesb.2020.5.72.

6. Giza++. URL: https://aclanthology.org/2022.amta-research.20.pdf