

O‘ZBEKCHA–QORAQALPOQCHA PARALLEL KORPUS: LOKAL QIDIRUV TIZIMI

Axmedova Aynura Quinishbaevna,
Kompyuter lingvistikasi mutaxassisligi
1-kurs magistranti
aynuraaxmedova345@gmail.com
ToshDO‘TAU

Annotatsiya. Ushbu maqolada o‘zbekcha-qoraqalpoqcha parallel korpus asosida lokal qidiruv tizimini yaratish tajribasi yoritiladi. Tadqiqot doirasida 1081 ta gap juftligidan iborat korpus bazasi shakllantirilib, u HTML, CSS va JavaScript texnologiyalari asosida ishlovchi qidiruv interfeysiga integratsiya qilindi. Tizim o‘zbekcha va qoraqalpoqcha matn birliklarini qidirish, solishtirish hamda qiyosiy tahlil qilish imkonini beradi. Mazkur resurs tarjimashunoslik, qiyosiy tilshunoslik, til ta’limi va tabiiy tilni qayta ishlash sohalari uchun amaliy ahamiyatga ega.

Kalit so‘zlar: *parallel korpus, o‘zbek tili, qoraqalpoq tili, lokal qidiruv, gap juftligi, korpus lingvistikasi, tarjima tahlili, NLP.*

Abstract. This article describes the development of a local search system based on an Uzbek-Karakalpak parallel corpus. Within the study, a corpus database consisting of 1,081 sentence pairs was compiled and integrated into a search interface developed using HTML, CSS, and JavaScript. The system enables users to search, compare, and analyze Uzbek and Karakalpak text units. The resource is of practical importance for translation studies, comparative linguistics, language education, and natural language processing.

Keywords: *parallel corpus, Uzbek language, Karakalpak language, local search, sentence pair, corpus linguistics, translation analysis, NLP.*

Kirish. Zamonaviy kompyuter lingvistikasi va tarjimashunoslikda elektron til resurslarini yaratish muhim ilmiy-amaliy yo‘nalishlardan biri hisoblanadi. Ayniqsa, kam resursli tillar uchun matnli korpuslar, parallel korpuslar, elektron lug‘atlar va

qidiruv tizimlarini shakllantirish tabiiy tilni qayta ishlash texnologiyalarining rivojlanishida muhim o‘rin tutadi.

Parallel korpus bir tildagi matn birliklarini boshqa tildagi mazmunan mos birliklar bilan bog‘lab saqlovchi elektron resursdir. Bunday korpuslar tarjima jarayonini o‘rganish, ikki tilli lug‘atlarni boyitish, qiyosiy lingvistik tahlil olib borish va mashina tarjimasi tizimlari uchun ma’lumot tayyorlashda qo‘llanadi. Korpusga asoslangan tarjimashunoslikda real tarjima matnlarini tahlil qilish nazariy xulosalarning ishonchliligini oshirishi ta’kidlanadi [1:175].

O‘zbek va qoraqalpoq tillari turkiy tillar oilasiga mansub bo‘lib, ular o‘rtasida leksik, morfologik va sintaktik jihatdan yaqinlik mavjud. Shu bilan birga, har ikki tilda imloviy, grammatik va uslubiy tafovutlar ham kuzatiladi. Bu o‘xshashlik va farqlarni tizimli o‘rganish uchun maxsus parallel korpus yaratish zarur.

Asosiy qism. *Parallel korpuslarning nazariy asoslari.* Parallel korpus ikki yoki undan ortiq tildagi matn birliklarini o‘zaro moslashtirilgan holda saqlaydigan elektron ma’lumotlar bazasidir. Bunday korpuslarda moslashtirish gap, abzats yoki segment darajasida amalga oshiriladi. Gap darajasidagi moslashtirish amaliy jihatdan qulay bo‘lib, foydalanuvchiga manba til va tarjima tilidagi birliklarni bevosita solishtirish imkonini beradi.

Ikki tilli korpuslarda gaplarni moslashtirish masalasi kompyuter lingvistikasida muhim tadqiqot yo‘nalishlaridan biri hisoblanadi. Gale va Church parallel matnlardagi gaplarni moslashtirishda statistik yondashuvdan foydalanish imkoniyatini ko‘rsatgan. Ularning tadqiqotida gaplarni moslashtirish xarakterlar uzunligiga asoslangan statistik model orqali amalga oshirilishi mumkinligi asoslangan [4:80].

Zamonaviy korpus texnologiyalarida parallel korpuslar keng qo‘llanadi. Masalan, OPUS loyihasi turli tillar uchun parallel ma’lumotlar, vositalar va interfeyslarni taqdim etuvchi ochiq resurs sifatida tanilgan [7:2230]. Shuningdek,

Sketch Engine kabi korpus platformalari konkordans, soʻz eskizlari va korpus yaratish imkoniyatlari orqali lingvistik tadqiqotlarda muhim vosita boʻlib xizmat qiladi [6:27]. Ushbu tajribalar oʻzbek va qoraqalpoq tillari uchun ham elektron parallel resurslar yaratish zarurligini koʻrsatadi.

Tadqiqot material. Tadqiqot uchun asosiy maʼlumot manbasi sifatida parallel_corpus_dataset.xlsx faylidan foydalanildi. Ushbu faylda oʻzbekcha va qoraqalpoqcha gap juftliklari ikki tematik boʻlimga ajratilgan. Birinchi boʻlim geografiya darsligi materiallaridan, ikkinchi boʻlim esa xabar matnlaridan tashkil topgan. Korpus tarkibi quyidagicha:

№	Boʻlim nomi	Gap juftliklari soni
1	geografiya_darslik	286
2	xabarlar	795
	Jami	1081

Geografiya darsligi materiallari ilmiy-oʻquv uslubiga xos birliklarni, xabar matnlari esa axborot uslubiga oid gap qurilmalarini qamrab oladi. Bu jihat korpusning tematik va uslubiy xilma-xilligini taʼminlaydi. Har bir qator bitta parallel birlik sifatida qaraladi: oʻzbek tilidagi gap va uning qoraqalpoq tilidagi mos varianti bir jadval qatorida joylashtiriladi.

Maʼlumotlar modeli. Parallel korpusning maʼlumotlar modeli gap juftliklarini tartibli, izchil va qidiruvga mos shaklda saqlashga asoslanadi. Korpusdagi asosiy birlik - parallel gap juftligidir. Har bir gap juftligi oʻzbekcha matn va unga mos qoraqalpoqcha matndan iborat. Tavsiya etiladigan maʼlumotlar modeli quyidagi maydonlarni oʻz ichiga oladi:

Maydon nomi	Tavsifi
id	Gap juftligining tartib raqami
section	Matn tegishli boʻlgan tematik boʻlim
uzbek_text	Oʻzbek tilidagi gap
karakalpak_text	Qoraqalpoq tilidagi mos gap
source_type	Matn turi: darslik yoki xabar

note

Qo'shimcha izoh yoki tahrir belgisi

Bunday model korpusni keyingi bosqichlarda kengaytirish uchun qulaydir. Jumladan, har bir gap juftligiga mavzu teglari, lemmalar, so'z turkumlari, morfologik belgilar yoki terminologik izohlar qo'shilishi mumkin. Natijada korpus oddiy qidiruv bazasidan lingvistik tahlilga yo'naltirilgan ko'p qatlamli resursga aylanishi mumkin.

Lokal qidiruv tizimi. Tadqiqot doirasida korpusdan foydalanishni qulaylashtirish maqsadida HTML, CSS va JavaScript texnologiyalari asosida lokal qidiruv tizimi ishlab chiqildi. Tizimning asosiy xususiyati shundaki, u server yoki internet aloqasini talab qilmaydi. Foydalanuvchi HTML faylni brauzer orqali ochadi va qidiruv tizimidan lokal rejimda foydalanadi.



1-rasm. Parallel korpus interfeysi

Qidiruv jarayoni JavaScript orqali amalga oshiriladi. Dastur ishga tushirilganda korpusdagi gap juftliklari ma'lumotlar massivi sifatida xotirada saqlanadi. Foydalanuvchi qidiruv maydoniga so'z, ibora yoki gap kiritganda, tizim o'zbekcha va qoraqalpoqcha ustunlar bo'yicha mos birliklarni aniqlaydi va natijalarni jadval shaklida ko'rsatadi.

Tizimning ishlash ketma-ketligi quyidagicha:

1. lokal HTML fayl brauzerda ochiladi;
2. korpus ma'lumotlari JavaScript orqali yuklanadi;
3. foydalanuvchi qidiruv so'rovini kiritadi;
4. tizim o'zbekcha va qoraqalpoqcha matnlar bo'yicha moslikni tekshiradi;
5. aniqlangan natijalar parallel jadval shaklida chiqariladi;
6. tanlangan tilga qarab ustunlar tartibi moslashtiriladi.

Ushbu yondashuv kichik va o'rta hajmdagi korpuslar uchun samarali hisoblanadi. Chunki barcha qidiruv amallari foydalanuvchi qurilmasining o'zida bajariladi va murakkab server infratuzilmasi talab qilinmaydi.

Interfeys imkoniyatlari. Lokal qidiruv tizimi foydalanuvchiga ikki tildagi parallel birliklar bilan ishlash imkonini beradi. Interfeysda o'zbek va qoraqalpoq tillarini tanlash funksiyasi mavjud. Til tanlanganda natijalar ustunlarining joylashuvi mos ravishda o'zgaradi. Masalan, o'zbek tili tanlanganda natijalar “O'zbekcha → Qoraqalpoqcha”, qoraqalpoq tili tanlanganda esa “Qoraqalpoqcha → O'zbekcha” tartibida ko'rsatiladi.

Tizim so'z, ibora yoki gap darajasida qidiruvni qo'llab-quvvatlaydi. Bu imkoniyat foydalanuvchiga ma'lum birlikning ikki tildagi qo'llanishini real kontekst asosida kuzatish imkonini beradi. Natijalar sonini cheklash funksiyasi esa ko'p natija qaytgan holatlarda interfeys bilan ishlashni tartibga soladi.

Mazkur interfeysning muhim jihati shundaki, u parallel birliklarni yonma-yon ko'rsatadi. Bu esa tarjima ekvivalentlarini aniqlash, terminlarning ikki tildagi ifodasini solishtirish va gap qurilishidagi farqlarni tahlil qilish uchun qulay sharoit yaratadi.

Lingvistik va amaliy ahamiyati. O'zbekcha-qoraqalpoqcha parallel korpus bir necha yo'nalishda ahamiyatga ega. Avvalo, u qiyosiy tilshunoslik uchun manba vazifasini bajaradi. O'zbek va qoraqalpoq tillaridagi leksik birliklar, grammatik shakllar va sintaktik qurilmalarni real matnlar asosida solishtirish imkonini beradi.

Ikkinchidan, korpus tarjimashunoslik tadqiqotlari uchun muhimdir. Tarjima birliklarini real kontekstda o'rganish orqali ekvivalentlik, uslubiy moslik va terminologik izchillik masalalarini tahlil qilish mumkin. Korpusga asoslangan tarjimashunoslikda tarjima hodisalarini empirik materiallar asosida o'rganish muhim metodologik tamoyil sifatida qaraladi [1:17].

Uchinchidan, ushbu korpus ta'lim jarayonida qo'llanishi mumkin. Talabalar ikki tildagi gap qurilishi, terminlar va iboralarni solishtirib o'rganishlari mumkin. Ayniqsa, o'zbek va qoraqalpoq tillarini qiyosiy o'qitishda bunday elektron resurslar amaliy samaradorlikni oshiradi.

To'rtinchidan, mazkur korpus tabiiy tilni qayta ishlash vazifalari uchun dastlabki resurs bo'lib xizmat qilishi mumkin. Parallel ma'lumotlar mashina tarjimasi, avtomatik termin ajratish, ikki tilli lug'at yaratish va semantik moslikni baholash kabi vazifalarda qo'llanadi. OPUS loyihasi tajribasi parallel ma'lumotlarning kompyuter lingvistikasi va tarjimashunoslikdagi ahamiyatini ko'rsatadi [7:2214].

Cheklovlar va takomillashtirish yo'nalishlari. Yaratilgan korpus dastlabki tajriba resursi bo'lgani sababli ayrim cheklovlarga ega. Birinchidan, korpus hajmi 1081 ta gap juftligi bilan chegaralangan. Bu lokal qidiruv tizimini sinovdan o'tkazish uchun yetarli bo'lsa-da, statistik tahlil yoki mashina tarjimasi modellarini o'qitish uchun yetarli hajm hisoblanmaydi.

Ikkinchidan, korpus hozircha ikki turdagi matn manbasini qamrab oladi: geografiya darsligi va xabar matnlari. Kelgusida badiiy matnlar, rasmiy hujjatlar, ilmiy-ommabop materiallar va ta'limiy matnlarni qo'shish orqali korpusning janr va uslubiy qamrovini kengaytirish lozim.

Uchinchidan, qidiruv tizimi oddiy matn mosligi asosida ishlaydi. O'zbek va qoraqalpoq tillari agglyutinatativ tillar bo'lgani sababli so'zlar ko'plab qo'shimchalar

bilan qo'llanadi. Shu bois kelgusida lemmatizatsiya, morfologik normalizatsiya va imlo variantlarini hisobga oluvchi qidiruv mexanizmlarini joriy etish zarur.

To'rtinchidan, tizim hozircha lokal rejimda ishlaydi. Bu dastlabki bosqich uchun qulay bo'lsa-da, korpus hajmi kengaygach, onlayn platforma, ma'lumotlar bazasi va foydalanuvchi boshqaruv tizimini ishlab chiqish maqsadga muvofiq bo'ladi.

Xulosa. Ushbu maqolada o'zbekcha-qoraqalpoqcha parallel korpus asosida lokal qidiruv tizimini yaratish tajribasi yoritildi. Tadqiqot doirasida 1081 ta gap juftligidan iborat dastlabki korpus bazasi shakllantirildi. Mazkur baza geografiya darsligi materiallari va xabar matnlaridan iborat bo'lib, ikki turkiy til o'rtasidagi leksik, grammatik va uslubiy mosliklarni o'rganish uchun muhim manba hisoblanadi.

HTML, CSS va JavaScript texnologiyalari asosida yaratilgan lokal qidiruv tizimi foydalanuvchiga parallel gap juftliklarini qidirish, ikki til kesimida ko'rish va qiyosiy tahlil qilish imkonini beradi. Tizimning lokal ishlashi uni kichik hajmdagi korpuslar bilan ishlashda texnik jihatdan qulay yechimga aylantiradi.

Kelgusida korpus hajmini kengaytirish, morfologik qidiruvni joriy etish, ma'lumotlarni standart formatlarga o'tkazish va onlayn platforma yaratish orqali ushbu resursning ilmiy-amaliy imkoniyatlarini yanada oshirish mumkin.

Foydalanilgan adabiyotlar ro'yxati

1. Baker M. Corpus-based Translation Studies: The Challenges that Lie Ahead // Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. Amsterdam: John Benjamins Publishing Company, 1996. P. 175-188.
2. Elov B., Alayev R. O'zbek tili korpusi va uning imkoniyatlari // O'zbekiston Informatika va energetika muammolari jurnali. Toshkent, 2023. № 2. 89-100-b.

3. Elov B., Amirkulov M. Uzbek-English Parallel Corpus Algorithm and Alignment Problem // Central Asian Journal of Literature, Philosophy and Culture. 2023. Vol. 4, Issue 6. P. 71-78.
4. Gale W. A., Church K. W. A Program for Aligning Sentences in Bilingual Corpora // Computational Linguistics. 1993. Vol. 19, No. 1. P. 75-102.
5. Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari. Filologiya fanlari bo'yicha falsafa doktori dissertatsiyasi. Buxoro, 2018. 165 b.
6. Kilgarrieff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. The Sketch Engine: Ten Years On // Lexicography. 2014. Vol. 1. P. 7–36.
7. Tiedemann J. Parallel Data, Tools and Interfaces in OPUS // Proceedings of the Eighth International Conference on Language Resources and Evaluation - LREC12. Istanbul: European Language Resources Association, 2012. P. 2214-2218.
8. Zaxarov V. P., Mengliyev B. R., Hamroyeva Sh. M. Korpus lingvistikasi. O'quv qo'llanma. Toshkent, 2021.