



KORPUS LINGVISTIKASIDA PARALLEL KORPUSLARNING AHAMIYATI

Boyqobilova Shahriyo Shuhratovna

Surxondaryo viloyati Sherobod tumani 2-son politexnikumi ingliz tili va
informatika fani o'qituvchisi

shahriyoboyqobilova@gmail.com

+99899 422-16-18

Annotatsiya: Ushbu maqolada parallel korpuslarning nazariy va amaliy asoslari, ularning tarixiy taraqqiyoti hamda lingvistika, tarjimashunoslik va kompyuter tilshunosligidagi o'rni tahlil qilinadi. Rozettsk toshidan boshlab zamonaviy yirik loyihalargacha bo'lgan parallel korpuslarning rivojlanish bosqichlari ko'rsatib berilgan. Xususan, word alignment modellarining shakllanishi, Europarl, Opus, JRC-Acquis va ParaCrawl kabi yirik korpuslarning ahamiyati yoritilgan. O'zbekistonda korpus lingvistikasining rivoji, Uzbek National Corpusdagi parallel bo'lim va uning imkoniyatlari alohida tahlil qilingan. Shuningdek, parallel korpuslarning tarjima strategiyalarini aniqlash, kompyuter lingvistikasi, mashina tarjimasi, leksikografiya va madaniyatlararo muloqotdagi qo'llanish yo'nalishlari yoritilgan. Ilmiy maqolada parallel korpuslarning afzalliklari bilan birga, ularni yaratishdagi cheklovlar ham ko'rsatib o'tilgan.

Kalit so'zlar: parallel korpus, bitekst, word alignment, Europarl, OPUS, ParaCrawl, tarjimashunoslik, kompyuter lingvistikasi, mashina tarjimasi, Uzbek National Corpus

KIRISH

Parallel korpus ya'ni parallel corpora ham deb atalib, bir tildagi jami foydalaniladigan matnlarning boshqa bir nechta tillardagi tarjima matnlarining va ularning muqobil variantlarining elektron analogini o'zida jamlagan korpusdir. Bu korpus matnlari original matnlar va ularning turli tildagi tarjima variantlari aks etgan matnlar blokidan iborat. Qoida bo'yicha, bunday korpuslar faqat ikki tilda tayyorlanadi, shuning uchun ham undagi o'zaro matnlar bitekst ham deyiladi. Bimatn ya'ni bitekst deb original matn va uning boshqa tillardagi tarjimalari mavjud bo'lgan matnlarga aytiladi.



ADABIYOTLAR TAHLILI VA METODOLOGIYA

Parallel korpuslarning tarixiy kelib chiqishiga nazar soladigan bo'lsak, ilk parallel korpus na'munasi sifatida eramizdan avvalgi II asrda yaratilgan Rozettsk toshini keltira olamiz. Bu toshning tepa qismiga qadimgi Misr iyerogliflarida yozilgan matn, toshning pastki qismida esa shu matnning qadimgi yunon yozuvidagi ko'rinishi tasvirlangan. Ushbu yodgorlik keyinchalik parallel matnlar va ularning ilmiy o'rganilishi uchun tarixiy manba vazifasini bajargan.

Parallel korpuslar (yoki bitekstlar) nazariyasi va amaliyotining shakllanishi bir necha bosqichlarda kechgan bo'lib, olimlar izlanishlari mazmuniga ko'ra uch guruhga ajratilishi mumkin:

1-guruh. Nazariy-statistik asoslar va moslash modellarini yaratish:

1990–1993-yillarda Peter F. Brown va IBM tadqiqotchilari hamda Brown, Lai va Mercer parallel korpuslardan foydalanishning matematik-statistik tamoyillarini ishlab chiqib, word alignment modellarini taklif etdilar. Shu davrda William A. Gale va Kenneth W. Church gap uzunligiga asoslangan moslash algoritmini ishlab chiqdilar.

1992-yilda Michel Simard, George F. Foster va Pierre Isabelle kognatlarga asoslangan gap moslash yondashuvini ilgari surdilar. 2002-yilda Robert C. Moore samarali gap moslash algoritmini yaratdi. Shuningdek, Franz Josef Och va Hermann Ney so'z darajasida alignmentni dasturiy vositalar yordamida amalga oshirish bo'yicha muhim hissa qo'shdilar.

2-guruh. Amaliy yirik korpuslar va alignment vositalari:

Keyingi bosqichda katta hajmdagi parallel korpuslarni shakllantirishga e'tibor qaratildi. Philipp Koehn Yevropa Parlamenti matnlariga asoslangan Europarl korpusini taqdim etdi. Jörg Tiedemann tomonidan ishlab chiqilgan OPUS loyihasi internetdagi parallel matnlarni tizimlashtirishga xizmat qildi. Ralf Steinberger boshchiligidagi guruh esa JRC-Acquis (Yevropa Ittifoqi huquqiy matnlari)ni korpusga aylantirdi. Shuningdek, Michal Ziemski, Marcin Junczys-Dowmunt va Bruno Pouliquen tomonidan BMT Parallel Korpusi ishlab chiqildi.

Keyinchalik ParaCrawl loyihasi 300 dan ortiq til bo'yicha parallel resurslarni taqdim etdi. Bu korpuslar zamonaviy neyron mashina tarjimasini (NMT) texnologiyalarining rivojlanishiga asos bo'ldi. Texnik jihatdan korpuslarni samarali tuzish uchun Dániel Varga Hunalign gap-moslash vositasini ishlab chiqdi.

3-guruh. Internet davridagi parallel matnlarni avtomatik yig'ish:

1998-yilda Philip Resnik internetdan parallel matnlarni avtomatik tarzda aniqlash va yig'ish g'oyasini ilgari surdi (Parallel Strands). Ushbu konsepsiya



keyinchalik ParaCrawl, OPUS va JW300 kabi yirik loyihalar uchun nazariy asos bo'lib xizmat qildi.

O'zbekistonda korpus lingvistikasi bo'yicha nazariy tadqiqotlar 2018 yilda boshlangan va 2021 yilda O'zbek tilining ta'limiy korpusi¹ yaratilishi bilan rivojlandi.

O'zbek olimlaridan korpus lingvistikasi xususida B.Mengliyev, Sh.Shahobiddinova, Z.Xolmanova, S.Karimov, N.Abduraxmonova, L.Raupova, Sh.Hamroyeva, M.Abjalova, G.Toirova, G.Ikromova, J.Djumabayeva, G.Ergasheva, A.Eshmo'minovlar ilmiy ishlar qilishgan. O'zbek tili milliy korpusining konseptologiyasi B.Mengliyev rahbarligidagi olimlar jamoasi hamorlikda ishlab chiqilyapti².

MUHOKAMA VA NATIJALAR

Korpus lingvistikasida parallel matnlar korpusi muhim ahamiyatga ega. Bundan tashqari parallel korpuslar ilm-fanning ko'plab sohalarida keng qo'llaniladi. Ayniqsa, uning nazariy va amaliy ahamiyati zamonaviy ilm-fan – lingvistika, tarjimashunoslik va kompyuter tilshunosligining rivojiga salmoqli hissa qo'shib kelmoqda.

Parallel korpuslar nazariy-metodologik jihatdan tilshunoslik, kontrastiv va chog'ishtirma tilshunoslikda nazariy manba vazifasini o'taydi, tahlillar yaratish imkonini beradi, til universallari va tipologik farqlarni topishda, so'zlarning ma'nolarini aniqlashda -semantik, polisemantik hodisalarni aniqlashda, kognitiv va madaniy konsepsiyalarni tadqiq qilishda keng qo'llaniladi.

Parallel korpuslar tarjima va tarjimashunoslikda amaliy tarjima faoliyatini rivojlantirishda, tarjima strategiyalarni empirik aniqlashda qo'llaniladi. Shu bilan birga tarjimonlarga ta'lim berishda, bilim va ko'nikmalarini oshirishda, ularning bilim olish samaradorligini oshiruvchi mashq va materiallar tayyorlashda foydalaniladi. Tarjima sifatini tekshirish, post-editing jarayonlarida va tarjima xatolarini statistik tahlil qilishda ham keng qo'llaniladi.

Parallel korpuslar kompyuter tilshunosligi va tabiiy tilni qayta ishlashda (NLP – Natural Language Processing) tilni matematik modellarda ifodalab, elektron mashinalarda elektron ko'rinishda tasvirlashda qo'llaniladi. Ular yana mashina tarjimasi modellarini o'qitishda, korpusdagi belgilangan tillar o'rtasidagi so'z va fraza moslashuvi algoritmlaridan foydalanishda, terminologiya va bilingvistik resurslarni yaratishda va kam resursli tillar uchun baza vazifalarida keng qo'llaniladi.

¹ <http://uzschoolcorpora.uz/>

² Abduraxmonova N. O'zbek tili elektron korpusining kompyuter modellari (monografiya) /Toshkent, Muharrir, 2021, 202b



Leksikografiya va terminologik tadqiqotlarda lugʻat tuzish va terminlarni ilmiy asoslashda parallel korpuslar katta roʻl oʻynaydi. Bu jarayonlarda parallel korpuslar asosida lugʻatlar yaratiladi, kollokatsiyalarni aniqlash va mavjud real foydalaniladigan misollarni toʻplash, terminologiyani standartlashtirish va ekvivalentlarini belgilash amallari bajariladi.

Madaniyatlararo muloqot va diskurs tadqiqotlarida parallel korpuslar realiya va lakunalarni aniqlashda, pragmatik strategiyalarni tahlil qilishda, diskurs va janrlar tahlillarida yaʼni argumentatsiya qilishda, retorika va stilistik farqlarni aniqlashda foydalaniladi.

Oxirgi yillarda parallel korpuslar yangi ilmiy yoʻnalishlarning rivoji uchun hissa qoʻshyapti va ilmiy tadqiqotlar olib borishda ham qoʻllanilib kelinmoqda. Masalan, antropologiya va madaniyatshunoslik sohalarida parallel korpuslar yordamida diaxron tadqiqotlar – tarixiy parallel matnlar asosida olib borilyapti.

Uzbek national corpus yordamida tilshunoslikda empirik dalillarga asoslangan tadqiqotlar olib borildi, til birliklarining real kontekstda qoʻllanilishini kuzatish imkoniyati yaratildi va tarjimalarda semantik moslikni taʼminlash boʻyicha bir qancha muammolarga yechim topildi.

Shu bilan birga hozirgi kunda ommalashib borayotgan Chatgpt va LLM modellari ham – zamonaviy tendensiyalar asosida parallel korpuslarga tayangan holda maʼlumot izlash va tarjima qilishda keng foydalanadi.

XULOSA

Bugungi vaqtga kelib deyarli barcha tillar boʻyicha korpuslar yaratilib boʻlindi va foydalanib kelinmoqda. Yillar davomida korpus yaratish rivojlantirilib, takomillashtirilib borildi. Shunga qaramasdan korpus yaratish tajribalari shuni koʻrsatadiki – mukammal korpus yaratishning iloji yoʻq. Foydalanish maqsadiga koʻra korpus imkoniyatlari cheklanadi va har bir vazifani alohida korpuslar bajaradi.

Korpuslar yaratishda bir qancha cheklovlar va muammolar bor. Bular:

- reprezentativlik va umumlashtirishda yuzaga keladigan muammolar;
- matnlarni moslashdagi muammolar;
- maʼnolarning toʻliq mos kelmasligi va tafovutlarning yuzaga kelishi muammolari;
- mualliflik asosida yaratilgan korpuslarda mualliflik huquqi va huquqiy cheklovlarning mavjudligi.

Oʻzbek-ingliz parallel matnlar korpusi – Uzbek National Corpus yaratilgan boʻlib, lekin qisman cheklangan yaʼni parallel matnlar toʻliq qamrab olinmagan. Uzbek



National Corpus da parallel korpus bo‘limi mavjud. Uning rasmiy saytida “Parallel corpus” bo‘limi ko‘rsatilib, foydalanuvchi foydalanish interfeysi orqali qidiruv amalga oshiradi. Hali bu korpus ustida amaliy ishlar olib borilyapti va uning hajmi, allignatsiya darajasi va ommaga to‘liq ochiq emas.

Kelajakda o‘zbek-ingliz parallel korpusini kengaytirish, ko‘p tilli korpuslar yaratilib va ochiq manba sifatida taqdim etilsa, mamlakatimiz ilm-fanida muhim yangiliklarga qadam bo‘lar edi.

FOYDALANILGAN ADABIYOTLAR

1. Abduraxmonova N. O‘zbek tili elektron korpusining kompyuter modellari (monografiya) /Toshkent, Muharrir, 2021
2. M.A. Abjalova. Korpus lingvistikasi. [Matn] : uslubiy qo‘llanma /Toshkent, Bookmany print, 2022
3. N.Sobirova, M.Akramjanova, Sh.Sobirov. Parallel korpus xususiyati, tuzilishi, interfeysi. Academic Research in Educational Sciences. 2022;5(5-2022):819-5
4. <http://uzschoolcorpora.uz/>