O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2024, [1/7/1]
ISSN 2181-7324

FILOLOGIYA
http://journals.nuu.uz
Social sciences

UDK:81’33

Kamola VASLIDDINOVA,
O‘zbekiston davlat jahon tillari universiteti tayanch doktoranti
E-mail: kvasliddinova12@mail.ru

O‘zMU professori, f.f.d N.Abduraxmonova taqrizi asosida

FUNCTIONAL CAPABILITIES OF THE PARALLEL CORPUS

Annotation

In this article, the evolution of the formation of the parallel corpus, its importance as an object in translation and lexicographic research is justified with examples. Functional possibilities of parallel corpora put into practice based on various theories created by linguists were analyzed and synthesized. Based on the functional capabilities of the parallel corpus, its functional category is enriched with new hypotheses.

Key words: Corpus, bilingual corpus, parallel corpus, automatic translation, active language, dead language, statistical machine translation, text alignment theory.

ФУНКЦИОНАЛЬНЫЕ ВОЗМОЖНОСТИ ПАРАЛЛЕЛЬНОГО КОРПУСА

Аннотация

В данной статье на примерах обосновывается эволюция формирования параллельного корпуса, его значение как объекта переводческих и лексикографических исследований. Проанализированы и синтезированы функциональные возможности параллельных корпусов, реализованные на практике на основе различных теорий, созданных лингвистами. На основе функциональных возможностей параллельного корпуса его функциональная категория обогащается новыми гипотезами.

Ключевые слова: Корпус, двуязычный корпус, параллельный корпус, автоматический перевод, активный язык, мертвый язык, статистический машинный перевод, теория выравнивания текста.

PARALLEL KORPUSNING FUNKSIONAL IMKONIYATLARI

Annotatsiya

Bu maqolada parallel korpusning shakllanish evolyutsiyasi, uning tarjimada va leksikografik tadqiqotlarda obyekt sifatidagi ahamiyati misollar bilan asoslab berilgan. Tilshunoslar tomonidan yaratilgan turli nazariyalar asosida amaliyotga joriy qilingan parallel korpuslarning funksional imkoniyatlari analiz va sintez qilingan. Parallel korpusning funksional imkoniyatlari asosida uning vazifaviy kategoriyasi yangi gipotezalar bilan boyitilgan.

Kalit so‘zlar: Korpus, bilingval korpus, parallel korpus, avtomatik tarjima, faol til, o‘lik til, statistik mashina tarjimasi, text alignment nazariyasi.

Kirish. Korpusning bilingval turi ikki turga bo‘linadi: Parallel va Comporable korpus. Parallel korpus bu korpusning funksional turlaridan biri hisoblanib, so‘nggi yillarda mashina tarjimasi, avtomatik tarjima, NLP, til texnologiyalari, bir tildagi matnning boshqa tildagi tarjimasi bilan yonma-yon kelishi natijasida lingvistik birliklarni qiyosiy tahlil qilish kabi zamonaviy tilshunoslikda olib borilayotgan tadqiqotlarning obyekti sifatida keng qo‘llanilmoqda. Parallel so‘zi lotin tilidan “parallelus”- para va grek tilidan “allelois” yonida so‘zlarining birikishidan hosil bo‘lgan so‘z birikmasi hisoblanadi. Korpusda matnlarning muqobili bilan yonma-yon turishiga aynan parallel korpuslar deya nom berilishi parallel so‘zining ma’nosiga bevosita bog‘lanadi. Ma’lumki, parallel korpus - bu bir tildagi matnning asl to‘plami va uning boshqa tildagi muqobilining yonma-yon kelishidir.

Mavzuga oid adabiyotlar tahlili. Parallel korpuslar sohasida izlanishlar olib borgan o‘zbek tadqiqotchisi Anorxon Axmedova tadqiqot ishida dastlabki parallel korpus sifatida 1822-yilda Jan-Franyois Shampollion tomonidan topilgan Rozetta bitiktoshida miloddan avvalgi 196-yilda Misr ibodatxonalari tomonidan qirol Ptolomey V ni sharafiga ikki tilda ya’ni yunon va misr tillarida bitilgan parallel matnlarni misol tariqasida keltiradi. Keyinchalik bu bitiktosh iyerogliflar ustida olib borilayotgan minglab gipotezalar va miflarga javob toppish uchun kalit vazifasini bajaradi [1].

Yuqorida keltirib o‘tilgan parallelizmning tarixi bugungi kundagi yo‘qalib borayotgan yoki aynan shu muammo havfi mavjud bo‘lgan barcha tillar uchun parallel korpus ustida chuqur izlanishlar olib borish hamda dunyoning eng rivojlangan tillarida tarjima muqobillari mavjud mahalliy materyallarni parallel korpusini yaratishi keyingi avlodlarga tabiiy tilni yetkazib berish imkonini beradi. Quyidagi 2-rasmda Einar H.Dyvik tomonidan The most spoken languages worldwide 2023 | Statista saytida 2024-yil uchun eng ommabop tillar ro‘yhati batafsil keltirib o‘tilgan.

Bugungi kunda dunyo tillari, ularning ishlatilish darajasi, lingvistik xilma xillikni barcha imkoniyatlarini keng ochib berish maqsadida UNESCO tomonidan “UNESCO WAL ya’ni THE WORLD ATLAS OF LANGUAGES” platformasi tashkil etilgan bo‘lib, bu platforma dunyo miqyosida tillar ustidan keng nazorat olib borish imkonini beradi. Butunjahon tillar atlasi metodologiyasiga [14] ko‘ra, hukumatlar, jamoat institutlari va akademik hamjamiyatlar tomonidan hujjatlashtirilgan 8324 ta til mavjud bo‘lib, ularda gaplashadigan yoki imzolangan; 8324 ta tildan 7000 ga yaqini hali ham ishlatilmoqda.

Galgary universiteti tadqiqotchisi Awil Hashi izlanishlari asosida 6700 ta tillarning 50% bugungi kunda yo‘qalib ketish ehtimoli mavjud ekanligini globallashuv va madaniy xilma xillikni yo‘qalib borishi bilan ilmiy asoslab berishga harakat qiladi va yo‘qolib ketish ehtimoli yuqori bo‘lgan tillar uchun Somali tillari asosida korpus modeli yaratish nazariyasini ilgari suradi [2].

Madaniyatlararo bioxilma-xillikning yo‘qolib borishi, globallashuv va modernizatsiya tillarning yo‘qolib borishining eng asosiy sababi ekanligi “the IUCN Red List criteria” asosida dunyo olimlari tomonidan ilmiy asoslab berilgan.

Masalan lotin tili o‘lik til hisoblanib, jamiyatning biror qatlami tomonidan og‘zaki nutq yoki hukumat hujjatlarida foydalanilmaydi. Lekin lotin tilida yozilgan ilmiy, badiiy, tarixiy asarlar har doim tadqiqotlar markazidan oliy o‘rin egallagan. Shu bois jahonning rivojlangan tillari hisoblangan ingliz, fransuz va nemis tillarida lotin tili juftligida parallel korpuslar yaratilgan.

The Persues Digital Library tomonidan 10.5 million lotin tilidagi so‘zdan iborat “Perseus Latin Word Study Tool” parallel korpusi yaratilib [16], Cicero. Caesar, Vergil kabi yozuvchilarning lotin tilida yozilgan asarlari ingliz tilidagi tarjimasi bilan parallellikda korpusga kiritilgan. Natijada lotin tili klassik til sifatida Gretsiyada o‘rta va oliy ta’lim muassalarida o‘qitilmoqda va Katolik cherkovi diniy tili sifatida saqlab qolindi.

Tadqiqot metodologiyasi. Tadqiqotimiz davomida analitik, analiz, sintez, qiyosiy, statistik metodlardan foydalangan holda parallel korpusning funksianal imkoniyatlarini yoritib berdik. Yuqorida batafsil keltirib o‘tilgan omillar ham parallel korpusning bugungi kundagi NLP til texnologiyalari uchun zaruratini asoslashga hizmat qiladi. Ayniqsa statistik ma’lumotlar asosida keltirib o‘tilgan faol tillar bilan tarjima muqobillari mavjud parallel matnlar korpusi yo‘qolib borayotgan tillar yoki tilshunos olimlar uchun muhim manba vazifasini bajaradi.

Mashina tarjimasida dastlab parallel korpuslardan foydalanish 1954-yil Amerikaning New York shahrida amalga oshirilgan Georgetown-IBM tajribasi ya’ni 49 ta ruscha gaplarining ingliz tilidagi muqobili bilan yonma-yon berilgan korpusi hisoblanadi [7]. Garchi bu korpus sig‘im jihatidan kichik hajmli bo‘lsada keyinchalik, boshqa yirik hajmli parallel korpuslar uchun na’muna sifatida asos bo‘ldi.

Tahlil va natijalar (Analysis and results). Ma’lumki, parallel korpusning yuqorida keltirib o‘tilgan xususiyatlarini o‘zida namayon etgan eng dastlabki korpuslardan biri bu – ikki tilda ya’ni ingliz va fransuz tilida Kanada Parlamenti hujjatlari asosida tuzilgan HANSARD KORPUSI hisoblanadi. Tilshunoslar Church va Gale konkordansing funksiyasi asosida parallel matnlarda “text alignmet” nazariyasiga asos soladi [10]. Lekin tilshunoslar tomonidan bu ilmiy nazariyaga faqatgina nazariy konsept emas, shuningdek operativ konsept deya ta’rif berishni joiz deb topadilar.

“Text alignmet” nazariyasining mohiyati “Corresponding units” tamoyili asosida ish ko‘radi [11], ya’ni parallel berilayotgan asliyatdagi matn va uning ikkinchi tilda berilgan tarjima muqobili so‘z, gap va leksik birliklar (chunk, kollokatsiya, frazema, formula gaplar, leksik qolipga aylangan gaplar) asosida ikki tildagi segment birliklarini aniqlaydi.

Yuqorida keltirilgan Church va Gale tomonidan yaratilgan Hansard korpusini 2014 va 2016-yillarda SAMUELS loyihasi doirasida bajarilgan https://www.english-corpora.org/hansard/ korpusi bilan taqqoslash mumkin [15]. Bu korpus ham Hansard nomini olgan va “The UK Arts and Humanities Research Group” tomonidan moliyalashtirilgan. Hansard korpusining ahamiyatli jihati shundaki, 1803-2005 yillar mobaynidagi Britaniya Parlamentida so‘zlangan har bir rasmiy nutqlar kiritilib, nafaqat leksik balki, semantik jihatdan ham teglangan hamda nafaqat leksikografik va og‘zaki nutq sentiment analizida muhim parallel korpus turlaridan biri hisoblanadi.

Perugia universiteti professori Federico Zanettin tarjima imkoniyatlarini kengaytirish maqsadida bilingval ingliz-italian parallel korpusida konkordansing software dasturining imkoniyatlari xususida izlanishlar olib brogan [5]. Buyuk Britaniyaning Sheffield universiteti professori Elke St.John nemis tilida so‘zlashuvchi talabalarga ingliz tilini ikkinchi til sifatida o‘qitishda nemis-ingliz parallel korpusi (INTERSECT) va Multiconcard deb nomlangan nemis-ingliz konkordansidan foydalangan holda 800 ta so‘zdan iborat olti turdagi matnlarni tahlil qiladi:

1.Hoechst, BASF, Siemens banklarining yillik bank xisobotlari matni,

2.”German News” internet saytidan olingan publitsistik uslubga xos matnlar,

3.Yevrofa ittifoqi matnlari,

4.Birlashgan millatlar tashkiloti hujjatlari matni,

5.Germaniyaning sobiq prezidenti Herzogning og‘zaki uslubdagi nutqi transkribsiyasi,

6.Germaniya, Shvetsariya va Avstriya davlatlari konstitutsion matnlari.

Tarjima muqobillari asosida yeg‘ilgan yuqoridagi matnlarning tarjima muqobillari asosida lingvistik xususiyatlar tahlilga tortilgan va dars jarayoniga bevosita tadbiq qilingan. Elke St.John olingan samarali natijalar asosida tarjimani o‘qitishda korpusdan foydalanishning istiqbolli mezonlari xususida tavsiyalar berib o‘tadi [6].

Portugaliyaning Surrey universiteti tadqiqotchisi Ana Frankenberg Garcia parallel korpusning strukturaviy jihatdan uchta turgan bo‘ladi: unidirectional, bidirectional, combination of both types [4]. Uning nazariyasi asosida parallel korpusning ikki tomonlama konfiguratsiyali ingliz-portugal parallel korpusi yaratildi. COMPORA deb atalgan ingliz - portugal parallel korpusi ishlash quvvat jihatidan 7.04 versiyali bo‘lib, Portugaliya, Braziliya, Mozambika, Angola, Birlashgan Qirollik(the UK), AQSH, Janubiy Afrika kabi davlatlarning 33 ta turli yozuvchilari tomonida 1837 va 2000-yillar oralig‘ida nashr etilgan asarlarining 69 tasini saralab olib, 3 million so‘zdan ziyod so‘zlarni parallel korpusga joylaydi [4]

1.Bir tomonlama konfiguratsiyali (unidirectional)

2.Ikki tomonlama konfiguratsiyali (bidirectional)

3.Kombinatsiyali (a combination of unidirectional and bidirectional configuration).

Xulosa va takliflar. Parallel korpus ba’zi manbalarda tarjima korpuslari ham deb ham atalib, ikki yoki undan ortiq monolingval korpuslardan iborat bo‘ladi. Maryland universiteti tadqiqotchisi Philip Resnik va Jons Hopkins universiteti professori Noah Smith Jons Hopkins nomli maqolasida parallel korpuslarning mashina tarjimasi va NLP sohasidagi tadqiqotlarda muhim ahamiyat kasb etgani bois ularga “bitexts” deya ta’rif beradi.

Tilshunos olimlar Gale va Church (1991) avtomatik til o‘rganish jarayonlarida parallel korpus resurs vazifasini bajaradi deya ta’kidlasa, Brown 1990, Melamed 2000, Och va Ney 2002 parallel korpuslarning statistik tarjima modellari uchun zaruriy ma’lumotlar bazasi deya ta’riflaydi. Davis va Dunning 1995, Landauer va Littman 1990, Oard 1997 parallel korpusning tillararo so‘zlarning lingvistik xususiyatlarini solishtirish imkoniyatlari haqida fikr yuritadi [9].

Rus olimi V.Zaxarov parallel korpusning ma’noviy xususiyati jihatidan shartli ravishda ikki turga ajratadi:

muayyan tildagi asliyat matnning ikkkinchi tildagi aynan tarjimasi;

biror sohaga oid mavzulashtirilgan guruhlar bo‘yicha bir tasnifga kiritilgan matnlar [12].

Parallel matnlarning mos juftliklarini tuzish uchun inson tomonidan tahrir qilingan matnlar kiritilib, lug‘at orqali matnda uchragan segment birliklarga qarab avtomatik tarzda tarjima birliklari aniqlanadi. Buning uchun ikki tilga mos leksik va grammatik jihatdan ishlab chiqilgan variantlar lug‘atga kiritilgan bo‘lishi kerak. Parallel korpus quyidagi maqsadlarda qo‘llanilishi mumkin [13].

• ikki yoki ko‘p tilli tarjima lug‘atlar tuzishda;

• mashina tarjima tizimlari uchun lug‘at yaratish va uni doimiy tarzda to‘ldirib borish;

• kontekstda uchraydigan ko‘p ma’noli so‘zlarning kompyuter tahlili orqali leksik birliklarni polisemiyaga oidligini aniqlash;

• matndagi terminologik va frazeologik birliklarni tarjima qilish;

• korpusdan foydalangan holda tarjimalarning mos variantlarini kompyuter xotirasiga yuklash va shu orqali mashina tarjimasi tizimi uchun to‘liq avtomatik tarjimani amalga oshirish.

Demak, parallel korpusga berilgan turli ta’riflardan kelib chiqib, uning quyidagicha funksional xususiyatlar kategoriyasini shakllantirish mumkin.

Statistik mashina tarjimasi uchun database;

Tillararo leksik-semantik xususiyatlarni chog‘ishtirish uchun manba;

Tarjima modellari uchun platforma;

Tarjimani o‘qitishda metodik yondashuv;

Yo‘qalib ketish havfi mavjud tillar uchun eletron SOS arxiv vazifasini bajaradi.

Tillararo tarixiy va sotsio-madaniy jihatdan tarjimada til normalari aks etgan qo‘llanma.

Tarjimada korpusdan foydalanish haqidagi dastlabki nazariya ingliz tilshunosi Mona Baker tomonidan ilgari surilib, olim “Corpus linguistics and translation studies: implications and applications” asarida korpus va tarjimani quyidagicha bog‘laydi “Ma’lum bir tildagi asliy matn va uning aynan tarjima variantining yirik hajmli korpusiga yuklanishi natijasida corpus-driven metoddan foydalanib tilning turli xususiyatlarini aniqlash imkoni paydo bo‘ladi” [8].

Korpusning funksional xususiyatlari sohasida olib borilayotgan tadqiqotlar korpusning imkoniyatlarini tobora kengligini isbotlab bermoqda. Masalan dastlab kopusning konkordans funksiyasi til o‘rganish yoki o‘qitishdagi instrument sifatida baholangan bo‘lsa, keyinchalik metodika sohasida korpus materiallarining autentik xususiyatlari borasida olib borilgan izlanishlar natijasida korpusning konkordans funksiyasi tilning leksik, semantik, sintaktik, stilistik xususiyatlarini avtomatik aniqlab berishi isbotlandi.

Korpusda konkordans funksiyasi bu - yuqori tezlikda so‘zlarni izlashga mo‘ljallangan funksional qurilma hisoblanib, u korpus tarkibidagi so‘zlar, frazalar, teglar, hujjatlar va matn turlarini aniqlaydi va korpus oynasida umumiy sanog‘ini namoyon qiladi.

ADABIYOTLAR

1. Anorxon Axmedova. Parallel korpusda o‘xshatishlarning leksik-semantik munosabatlari tadqiqi. Monografiya. Toshkent.2022. B 141.
2. Awil Hashi. Developing a Model Corpus for Endangered Languages. CALGARY, ALBERTA JULY, 2014. P 246.
3. Amano T, Sandel B, Eager H, Bulteau E, Svenning J-C, Dalsgaard B, Rahbek C, Davies RG, Sutherland WJ. 2014 Global distribution and drivers of language extinction risk. P 281: 20141574. http://dx.doi.org/10.1098/rspb.2014.157
4. Ana Frankenberg Garcia. Compiling and using a parallel corpus for research in translation. 2009.
5. Claudio Fantinuoli and Federico Zanettin (ed.). 2015. New directions in corpus-based translation studies (Translation and Multilingual Natural Language Processing 1). Berlin: Language Science Press. P 103.
6. Elke St.John. A Case for Using a Parallel Corpus. Language Learning & Technology http://llt.msu.edu/vol5num3/stjohn September 2001, Vol. 5, Num. 3 pp. 185-203.
7. Garvin, Paul L.. "The Georgetown-IBM Experiment of 1954: An Evaluation in Retrospect". Papers in linguistics in honor of Léon Dostert, edited by William Mandeville Austin, Berlin, Boston: De Gruyter Mouton, 1967, pp. 46-56. https://doi.org/10.1515/9783111675886-006
8. M.Baker “Corpus linguistics and translation studies: implications and applications” 1993 p 243. https://doi.org/10.1075/z.64.15bak
9. Philip Resnik, Jons Hopkins. Jons Hopkins. Computational Linguistic. Volume 29, Number 3.349-380.
10. Zhaorong Zong1, Changchun Hong. Research on Alignment in the Construction of Parallel Corpus IOP Conf. Series: Journal of Physics: Conf. Series 1213 (2019) 042003 IOP Publishing doi:10.1088/1742-6596/1213/4/042003.p 1-5.
11. Zhaorong Zong1, Changchun Hong. Research on Alignment in the Construction of Parallel Corpus IOP Conf. Series: Journal of Physics: Conf. Series 1213 (2019) 042003 IOP Publishing doi:10.1088/1742-6596/1213/4/042003.p 1-5.
12. Захаров В.П., Богданова С.Ю. КОРПУСНАЯ ЛИНГВИСТИКА. (Учебник) - Санкт-Петербург, 2013, - С.62
13. Захаров В.П., Богданова С.Ю. КОРПУСНАЯ ЛИНГВИСТИКА. (Учебник) - Санкт-Петербург, 2013, - С.64
14. About the World Atlas of Languages | UNESCO WAL.30.05.2024.
15. https://www.clarin.ac.uk/hansard-corpus.10.06.2024.
16. http://www.perseus.tufts.edu/hopper/.2.06.2024.