PARALLEL KORPUS MASHINA TARJIMASINING O‘RGANISH OBYEKTI SIFATIDA

Jo‘raqulova Mushtariy Abdumalikovna
O‘zbekiston Milliy universiteti 2-kurs magistranti.
E-mail: mushtariy00712@gmail.com

Annotatsiya: Ushbu maqola parallel korpuslarning turlari uning ahamiyati, va uni tuzishdagi muammolari hamda mashina tarjimasini rivojlantirish masalalariga bag‘ishlangan.

Kalit so’zlar: korpus tilshunosligi, parallel korpus matnlari (PKM), mashina tarjimasi, Ingliz va Norvegiya Parallel Korpusi, CzEng parallel korpusi, Scielo parallel korpusi, FAPESP parallel korpusi.

Abstract: This article is devoted to the types of parallel corpora, their importance, and the problems of their creation, as well as the development of machine translation.

Keywords: corpus linguistics, parallel corpus texts (PKM), machine translation, English and Norwegian Parallel Corpus, CzEng parallel corpus, Scielo parallel corpus, FAPESP parallel corpus.

Аннотация: Данная статья посвящена важности типов параллельных корпусов, их проблемам и совершенствованию машинного перевода.

Ключевые слова: корпусная лингвистика, параллельный корпус текстов (PKM), машинный перевод, параллельный корпус английского и норвежского языков, параллельный корпус CzEng, параллельный корпус Scielo, параллельный корпус FAPESP.

Zamonaviylashib borayotgan jamiyatimizda avtomatik informatsion texnologiyalar va kompyuter muhim ahamiyat kasb etmoqda. Kompyuter hozirgi kunda barcha sohalariga kirib borayotgani hammamizga birday ma’lumdir. Buning natijasida ko‘plab yangi sohalar rivojiga o‘z hissasini qo‘shmoqda.Shu jumladan, kompyuter texnologiyalarining tilshunoslik fani bilan hamkorligini boshlash natijasida kompyuter lingvistikasi fani vujudga keldi.

Korpus lingvistikasining ilk bo‘g‘ini Naom Chomskiy nomi nilan bog‘liq bo‘lib, u generativ tilshunoslikka oid qarashlarni ilgari surgan bo‘lsa ham, uning matnlarni amaliy jihatdan o‘qiy oladigan shaklda to‘plab, uni kompyuter tahlil qiladigan darajaga ko‘targan bosqich korpus lingvistikasi uchun dastlabki qadam bo‘ldi.

Kompyuterda yaratilgan birinchi matnlar korpusi Braun korpusi (БК , inglizcha Brown Corpus, BC) hisoblanadi, u 1961 -yilda Braun universitetida yaratilgan, har biri 2000 so‘zli 500 ta matn fragmentini o‘z ichiga oladi. 1970-yillarda 1 mln so‘zni o‘z ichiga olgan matnlar korpusi asosida rus tilining chastotali lug‘ati yaratildi. 1980-yillarda Shvetsiyaning Upsala universitetida ham rus tilida matnlar korpusi yaratildi[Abdurakhmonova, N.2021,2022,2023].

Parallellik mezoniga ko‘ra korpuslar bir tili, ikki tili va ko‘p tiliga bo‘linadi. Bir tili korpuslarda dialektlar va til variantlariga qarama –qarshi bo‘ladi. Masalan, ingliz tili ona tili va ingliz tili chet tili kabi yangi texnologiyalar paydo bo‘lgunga qadar ilmiy qiziqish doirasidan tashqarida qoldi, bu esa qarama-qarshi tildagi nutq asarlarini ancha ko‘p jalb qilishga imkon berdi.

Ingliz va Norvegiya Parallel Korpusi (ENPC) asl badiiy va badiiy bo'lmagan ingliz va norvegcha matnlar va ularning tarjimalaridan olingan ikki million tokendan iborat. Asl matnlar ularning tarjimalari bilan jumla darajasida moslashtirildi. Ushbu korpus haqiqiy matnlar va ularning tarjimalari, har ikki tildagi asl nusxalar, ikkala tildagi tarjimalar va asl nusxalar hamda bir tildagi tarjimalarni taqqoslashga qaratilgan edi. Keyinchalik ENPCga nemis-norvegiya, fransuz-norvegiya, rus va norvegiya parallel korpuslari qo‘shildi va bu Oslo ko‘p tilli korpusini tashkil qildi[H. Hasselgard, 2010; 98-101].

ENPC dan keyin ingliz tilini o‘z ichiga olgan yana bir qancha parallel korpuslar yaratildi. Ular quyidagilarni tashkil qiladi: ingliz-shved parallel korpusi, ingliz-fransuz korpusi, ingliz-german korpusi va ingliz-ispan korpuslaridir. Bunga misol qilib, JRC-ACQUIS Multilingual Parallel Corpusini ham ko‘rsatib o‘ta olamiz. U 22 tildan bir milliarddan ortiq tokendan iborat[G. R. Yepes,2011;65-80]. Qo‘shimcha qilib,Ochiq Parallel Korpus (OPUS) Yevropa va Osiyo tillarining parallel korpusi, shu jumladan arab tillaridagi 352 millionga yaqin tokenlardan iborat.

Arabcha parallel korpus 1999 yild Job Xobkins universitetida mashina tarjimasini osonlashtirish maqsadida ingliz-arab parallel Misr korpusi ishlab chiqarilgandan so‘ng vujudga keldi. Bu korpus ingliz va arab tilidagi Qur’ondan iborat edi. Qur’on arabcha Corpus [http://corpus.quran.com/] Scielo korpusidir morfologik izohli klassik arab tilini tasvirlash uchun juda yaxshi manba hisoblandi.

Keyinchalik esa Al-Ajmiy tomonidan 2004-yilda ingliz-arab parallel korpusi yaratildi va unda tarix, iqtisod san’at, fan va adabiyotga oid 3 million so‘zdan iborat tarjima kitoblar turlaridan bor edi. Birlashgan Millatlar Tashkilotining Ingliz-Arab Parallel Korpusi (EAPCOUNT) 2013 yilda Hammuda Salhi tomonidan tuzilgan.

Keyingi CzEng korpusi bo‘lib, chex-ingliz parallel korpusi notijorat tadqiqot va ta’lim maqsadlarida erkin foydalanishlari mumkin. Bu korpusning aksariyat qismi kitoblar,film subtitrlari hamda Yevropa Ittifoqi qonunchilik manbalaridan iborat. Bu korpus tokeni har bir til uchun taxminan 200 millionni tashkil qiladi. Bu korpusning oxirgi nashriga ko‘ra morfologik teglar, sirt sintaktik va avtomatik havola xususiyatlari kiritilgan[O. Bojar ., 2016; 231-238].

Yana bir parallel korpus uchun na’munasi Scielo korpusidir. Biotibbiyot sohasi (biologiya fanlari va sog'liqni saqlash fanlari) uchun bepul mavjud bo'lgan parallel ilmiy nashrlar korpusidir. Bu korpus Scielo ma’lumotlar bazasidan olingan. Korpus uchta til juftligi uchun tuzilgan: portugal-ingliz (jami 86 000 ga yaqin hujjat), ispan-ingliz (taxminan 95 000 ta hujjat) va fransuz-ingliz (2000 ga yaqin hujjat). [M.Neves., 2016].

FAPESP korpusi ilmiy yangiliklar matnlari asosida ikki til juftlikda, portugal-ingliz va portugal-ispan tillarida tuzilgan. Ushbu matnlar onlayn ko‘p tilli Braziliya jurnalidan (Pesquisa FAPESP) avtomatik ravishda o‘rganildi. Korpus hujjat va jumla qilib tizimlashtiriladi. Unda 2700 ga yaqin parallel hujjatlar mavjud bo‘lib, ular har bir tilda 150 000 dan ortiq tizimlangan jumlalarni o‘z ichiga oladi.[28]

Bu ikki va ko‘p tilli parallel korpuslar qurilishi jihatdan qiyin bo‘lganligi sababli yuqoridagilargina bepuldir.

Parallel korpuslar mashina tarjimasi (MT) tizimlarini rivojlantirishda muhim rol o'ynaydi. Mashina tarjimasida uchta yondashuv mavjud bo‘lib, ular quyidagilarni tashkil etadi: lingvistik,

statistik va kompyuter yordamida amalga oshiriladigan bilimlardir[Abdurakhmonova, N.2021,2022,2023].

Birinchi yondashuv qoidaga asoslangan bo‘lib, manba va maqsad tillarning morfologik, sintaktik, semantik va idiomatik bilimlari kabi lingvistik bilimlarga bog‘liq. Bunday yondashuvda bosh gap tahlilchi yordamida maqsadli gapga tarjima qilinadi. Ushbu tahlilchi manba jumlani NP, VP, AdvP va PP kabi tarkibiy qismlariga tahlil qiladi va keyin ularni TL ekvivalentlari bilan almashtiradi.

Ikkinchi yondashuv - statistik mashina tarjimasi. U parallel korpusni tahlil qiladi, eng tez-tez mos keladigan SL-TL naqshlarini tanlaydi va ularni tarjimada ishlatadi. Masalan, Google Translate kabi statistik tizimlarda parallel korpus, monolingual korpus va ularning ma'lumotlarining statistik modellari matnlarni bir tildan boshqa tilga avtomatik ravishda ko‘rsatish uchun keng qo‘llaniladi[F.J.Och, 2005; Thailand].

Uchinchi yondashuv - kompyuter yordamida tarjima bo‘lib, u mashina va tarjimon o‘rtasidagi interaktiv jarayonni o‘z ichiga oladi. Tarjima jarayonini osonlashtirish va avtomatlashtirish uchun kompyuter yordamida tarjima dasturlarining ko‘p turlari (masalan, elektron lug‘atlar va tarjima xotiralari) ishlab chiqilgan. Ikki tilli va ko‘p tilli elektron lug'atlar nutqning bir qismi, talaffuz va birikmalar kabi SL va TL so'zlari haqidagi ma'lumotlarni o‘z ichiga oladi. Ushbu lug‘atlarni turli shakllarda topish mumkin: maxsus qurilmalar (masalan, Atlas zamonaviy lug'ati inglizcha-arabcha), kompyuter dasturlari (masalan, Parallel korpusni ishlab chiqish oddiy ish emas. Bu qurilishning ko'p bosqichlarida duch keladigan texnik va lingvistik qiyinchiliklar bilan bog'liq: matnni tanlash, o'zgartirish, segmentatsiya qilish, stamplash, tekislash va izohlash[Abdurakhmonova, N.2021,2022,2023].

Parallel korpusni yaratishda ko‘plab qiyinchililklar mavjud. Ular texnik va lingvistik qiyinchiliklar bilan bog‘liq: matnni tanlash, o'zgartirish, segmentatsiya qilish, shtamplash, tekislash va izohlash kabi bir qancha muammolari mavjud. Bunga sabab qilib esa, tilning avtomatik moslashtirilishiga yetarlicha parallel matnlarga manba yo‘qligi va texnalogik jihatdan kam rivojlanganlik hisoblanadi. Parallel korpusni loyihalash uchun universal retsept yo‘q. Parallel korpus uchun ma’lumotlarni tanlash jarayoni tarjima qilingan ma’lumotlarning mavjudligi, tarjima qilingan materialning sifati, maqsadi tillar va tarjima yo‘nalishlari uchun tarjima qilingan materialning proportsional mavjudligi kabi bir qator cheklovlar bilan tavsiflanadi.

Foydalanilgan adabiyotlar ro‘yxati:

1. N. Abduraxmonova. O‘zbek tili elektron korpusining kompyuter modellari (Monografiya). GlobeEdit, 2021.
2. H. Hasselgard. Contrastive Analysis / Contrastive Linguistics. The Routledge Linguistics Encyclopedia, K. Malmkjær, Ed., Third Edition, London, NY: Routledge, pp. 98-101, 2010.
3. G. R. Yepes. Parallel Corpora in Translator Education. Redit: Revista Electrónica de Didáctica de la Traducción y la Interpretación, pp. 65-80, 2011.
4. H. Al-Ajmi. A New English–Arabic Parallel Text Corpus for Lexicographic Applications. Lexikos, vol. 14, pp. 326-330, 2004
5. Абдурахмонова, Н., & Абдувахобов, Г. (2021). О ‘QUV LUG ‘ATINI TUZISHNING NAZARIY METODOLOGIK ASOSLARI. *МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА*, *4*(6).

6. Abdurakhmonova, N., Shakirovich, I. A., & O‘G‘Li, K. N. S. (2022). Morphological analyzer (morfoAnalyse) Python package for Turkic language. *Science and Education*, *3*(9), 146-156.
7. Ismailov, A. S., Shamsiyeva, G., & Abdurakhmonova, N. (2021). Statistical machine translation proposal for Uzbek to English. *Science and Education*, *2*(12), 212-219.
8. Абдурахмонова, Н., & Бойсариева, С. (2023). TABIIY TILNI QAYTA ISHLASHDA (NLP) OKKAZIONALIZMLARNING MORFEM TAHLILI. *МЕЖДУНАРОДНЫЙ ЖУРНАЛ ИСКУССТВО СЛОВА*, *6*(3).
9. Abdurakhmonova, N., Tuliyev, U., Ismailov, A., & Abduvahobo, G. (2022). UZBEK ELECTRONIC CORPUS AS A TOOL FOR LINGUISTIC ANALYSIS. In *Компьютерная обработка тюркских языков. TURKLANG 2022* (pp. 231-240).
10. O. Bojar, O. Dusek and others. Enlarged czech-english parallel corpus with processing tools dockered, in: Int. Conf. Text Speech Dialogue, Springer, 2016, pp. 231–238.
11. M. Neves, A.J. Yepes, A. Neveol. The scielo corpus: a parallel corpus of scientific publications for biomedicine, Proc. Tenth Int. Conf. Lang. Resour. Eval. LREC 2016 Paris Fr. Eur. Lang. Resour. Assoc., ELRA, 2016.
12. F. J. Och. Statistical Machine Translation: Foundations and Recent Advances, presented at the Tutorial at MT Summit, Phuket, Thailand, 2005.
13. L. Bowker and J. Pearson. Working with Specialized Language: A Practical Guide to Using Corpora, London, NY: Routledge, 2002.
14. Abduraxmonova N. Mashina tarjimasining lingvistik asoslari. GlobeEdit, 2012.