ISSN 2181-3187

ILMIY MATNLAR BO‘YICHA PARALLEL KORPUS YARATISHNING NAZARIY ILDIZLARI

Jo‘raqulova Mushtariy Abdumalik qizi
Mirzo Ulug‘bek nomidagi O‘zbekiston Milliy universiteti

Annotatsiya: Ushbu maqola korpus va uning ahamiyati, korpus turlari, o‘zbek tili ilmiy matnlari parallel korpusi lingvistik manbalarning yangi turi hamda matnlarning parallel korpuslarini tuzish muammolari masalalariga bag‘ishlangan.

Kalit so‘zlar: kompyuter lingvistikasi**,** korpus tilshunoslik, Braun korpusi,parallel korpus,manba tili , tarjima tili, ilmiy matnlar parallel korpusi.

Аннотация: Данная статья посвящена корпусу и его значению, видам корпусу узбекских научных текстов в новому типу лингвистических источников и проблемам создания параллельного корпуса текстов.

Ключевые слова: компьютерная лингвистика**,** корпусная лингвистика**,** Браун корпус, параллельный корпус, исходный язык, язык перевода, параллельный корпус научных текстов.

Annotation: This article deals with the corpus and its significance, types of corpus, parallel corpus of Uzbek scientific texts, a new type of linguistic sources and the problems of creating parallel corpus of texsts.

Key words: computer linguistics, corpus linguistics, Brown Corpus, parallel corpus, source language, translation language,parallel corpus of scientific texts.

KIRISH

Ochig‘i korpusni insoniyat hali korpus lingvistikasi fani paydo bo‘lmasdan oldin, ya‘ni XVIII asrdan tadqiq eta boshlashgan. Misol qilib oladigan bo‘lsak, Bibliyani tadqiq etish, lug‘atlar yaratish (Johnson, Oxford English Dictionary, Webster Dictionary), tillarni o‘qitish (chastotali lug‘at Thorndike'a, 1921), deskriptiv grammatika (Fries, 1940, Quirk, 1968) va boshqalar. Kvirk korpusi bir million so‘z birikmalarini o‘zida jamlagan bo‘lib, har biri o‘n yetti qator matndan iborat million kartotekadan iborat. Bu korpus oxirgi elektron shaklda bo‘lmagan korpus edi. Mazkur korpusning yaratilishiga 25 yil vaqt sarflangan. Kvirk korpusi 1989-yilda oxirgi ishlarini yakunlagan. Bu paytda texnologiyalar yuqori sur’atda rivojlanib ketayotgan edi. Buning natijasi o‘laroq korpus tezlikda elektron shaklga o‘tkazildi. Hozirda bu korpus Londondagi University kollejida saqlanmoqda.

Kompyuter yordamida yaratilgan korpuslarning asosiy davrlari:

1. 1960-yil, Braun korpusi (AQSh), bir million so‘zni qamrab oladi;
2. 1970-yil, LOB korpusi (Buyuk Britaniya, Norvegiya), bu korpus ham bir million so‘zni qamrab oladi;

ISSN 2181-3187

3. 1980-yil, Rus tilining mashinalashgan fondi (Машинный Фонд русского языка);

4. Rus tilining Uppsal korpusi (Shvetsiya), bir million so‘zni qamrab oladi;

5. 1990-yil, Britaniya milliy korpusi (British National Corpus), milliy korpuslar (venger, italyan, xorvat, chex, yapon millatlari tillari) 100million so‘zni o‘z ichiga oladi; [1]

6. Ingliz tili majmui (The Bank of English), Birmingham (Collins Cobuild), 600 million so‘zdan iborat;

7. 2000-yil, Amerika milliy korpusi (American National Corpus), 100 million so‘zni o‘z ichiga oladi;

8.Amerikacha ingliz tilining zamonaviy korpusi (Corpus of Contemporary American English), 400 million so‘zdan iborat;

9. Rus tilining milliy korpusi (Национальный корпус русского языка), 140 million so‘zni o‘z ichiga oladi;

10.Gigaword corpora: ingliz, arab, xitoy tillarini qamrab olgan bo‘lib, 2 milliard so‘zdan tashkil topgan;

11.Oxford English corpus, mazkur korpus ham 2 milliard so‘zdan tashkil topgan;

Bugungi kunda har bir rivojlangan til o‘z korpusini yaratish ustida ish olib bormoqda. Eng mashhur va ko‘zga ko‘ringan korpuslarni yuqorida sanab o‘tdik. Kompyuter texnologiyasining tez rivoji bundanda katta hajmdagi korpuslarni ilm ahliga taqdim etadi. V.V.Rikovning yozishicha, korpus lingvistikasi atamasi munozarali masala hisoblanadi. Chunki korpus lingvistikasi matnlar massivini yaratish masalasimi yoki korpus ma’lumotlari asosidagi lingvistikami? Amaliyotda korpus lingvistikasi deganda:

- birinchidan, korpus mazkur fan uchun nutq materiali;
- ikkinchidan, faoliyat natijasidir.

Xulosa qilib aytganda, quyidagilarga imkoniyat yaratdi:

1. Tilga oid ilgari qilingan tadqiqotlar ko‘lami va natijalarini kompleks holda aniqlash va tahlil qilish;

2. Yangi va keng ko‘lamda lingvistik tadqiqotlar olib borish.[1]

Korpus lingvistikasi fani o‘z obyekti va ish materialini o‘zi yaratdi, bu esa uni mustaqil lingvistik fan sifatida tan olishga asos bo‘ladi. Biz tadqiq etayotgan fanning asosiy maqsadi – til tizimini lingvistik tasvirlashdir.

Korpus lingvistikasining asosiy yo‘nalishlari:

Zamonaviy korpus lingvistikasining asosiy yo‘nalishlari quyidagilar:

- birinchidan, bu lug‘atlar yaratish hamda leksikografik tadqiqotlar olib borishdir, zamonaviy ingliz tilining barcha lug‘atlari korpusga asoslangan(Collins, Webster, MacMillan va boshqalar);

ISSN 2181-3187

- ikkinchidan, korpuslarni o‘rganish orqali tillarning leksik tarkibi haqida aniq ma’lumotlar olish, so‘zlarning qo‘llanish chastotalarini tuzish.

Korpus materialining necha tilda berilishiga ko‘ra uning bir va ko‘p tilli turlari mavjud. Korpus mutaxassislarini (asosan, tarjimon) doim bir necha tilli korpus yaratish qiziqtirib kelgan. Korpus yaratishning ilk davridan boshlab ingliz, fin, fransuz, nemis, grek, norveg, ispan, shved va hokozo tillar uchun ikki tilli korpuslar paydo bo‘la boshlagan. Bunday korpus Bitexts deb ham ataladi. Korpusni ikki tilli emas, balki uch, to‘rt va undan ortiq tilli qilishga hech qanday to‘siq yo‘q.

Mutaxassislar parallellik nuqtayi nazaridan korpusni bir, ikki va ko‘p tilli kabi turlarga ham bo‘lishadi. Bir tilli korpusda til varianti va shevalar bir-biriga qarama-qarshi qo‘yilsa, ikki va ko‘p tilli korpus bir mavzu doirasida turli tilda yozilgan matnlar majmuidan iborat bo‘ladi. Masalan, ma’lum ilmiy muammo borasida turli davlatda turli tilda o‘tkazilgan konferensiya materiallarini qamrab olishi mumkin.

Ko‘p tilli korpuslar, odatda, tarjimonlar tomonidan foydalaniladi. Ko‘p tilli korpusning yana bir ko‘rinishi original matn va tarjima matndan iborat bo‘ladi.

Korpusning ushbu turi qiyosiy chog‘ishtirma tadqiqot olib borishda, tarjima nazariyasi hamda kompyuter tarjimasini o‘rganishda juda muhim manba bo‘lib xizmat qiladi. Ko‘p tilli korpusning 2 turi mavjud:

1) bir-birining tarjimasi bo‘lgan matnli korpus;

2) bir mavzuga oid ikki tildagi matnli korpus.

Birinchi tipdagi korpus —parallel korpus (parallel corpora) deb nomlanib, ma’lum bir tarjimaning turli aspektini o‘rganish uchun qo‘llaniladi. Masalan, Kanada parlamenti yig‘ini (ingliz/fransuz) matnlari korpusi mavjud. Parallel korpus o‘z navbatida yana 2 turga – moslashtirilgan (aligned) va moslashtirilmagan (not aligned) korpusga ajraladi. “Moslashtirilgan” atamasi korpusda tarjima birliklari orasida bir-birini taqozo etuvchi aniq aloqa mavjudligini bildiradi. Bunday korpusning afzalligi u yoki bu gapning qanday tarjima qilinganini topishda qulaylik mavjudligida. Bu turdagi korpus tarjimon uchun ahamiyatli, chunki unda noyob resurs–tarjima xotirasi (translation memory) mavjud. “Moslashtirilmagan” korpus (—qiyosiy korpus)ning vazifasi matnni uning tarjimasi bilan moslashtirish, bir birlikning tarjimada qaysi birlikka to‘g‘ri kelishini ko‘rsatib turishdan iborat. To‘g‘irlash avtomatik ravishda yoki qo‘lda bajarilishi mumkin. Birinchi usul oson, lekin xatolari ko‘p. Masalan, tarjima jarayonida sodda gap qo‘shma gap shaklida berilishi mumkin. Bunday paytda qaysi qurilish original ekanligini aniqlash qiyin bo‘ladi. Ko‘p tilli —moslashtirilgan korpus namunasi sifatida Yevropa Ittifoqining Acquis Communautaire ma’lumotlar bazasini keltirishimiz mumkin.

Ikkinchi xildagi korpus —tarjima korpusi (translation corpora) deb atalib, ayni bir fikrning turli tildagi ifodasini o‘rganish uchun muhim.

Parallel korpusning qimmati uning hajmi va tillarning miqdori bilan [2]

ISSN 2181-3187

belgilanadi. Acquis Communautaire dunyodagi eng katta parallel korpus bo‘lib, muhim jihati bu korpusdan foydalanishning bepulligi va malti-eston, sloven-fin kabi kam uchraydigan tillar juftligining mavjudligi bilan baholanadi. Ushbu korpuslardan quyidagi maqsadlarda foydalanish mumkin:

1) tipik tarjima usullari va transformatsiyani yuzaga keltirish;
2) avtomatik tarjima tizimi statistikasini o`rganish;
3) bir va ko`p tilli lug’atlar yaratish;
4) ma’lumotni saqlash va uzatish dasturlarini o‘rganish va baholash;
5) tarjima to‘g‘riligini avtomatik tarzda tekshirish;
6) ekvivalent tanlash imkoni kengligi orqali tarjimon mehnatini osonlashtirish.

Parallel matn- matn va uning tarjimasi yoki tarjimonlar tomonidan tuzilgan to‘plam tarjima va matnlardan tuzilgan bo‘ladi. Bitext- 2 ta parallel matning tarjima variantidir. Parallel matnni muqobil variantlari token va bloklar o‘rtasidagi yozishmalarni aniqlash vazifasini bajaradi. Hozirgi kunda parallel matnlarning muqobil variantlari keng ko‘lamli sohalarda qo‘llaniladi. Masalan bilimlarni qidirish, mashinali tarjima, tabiiy tilni qayta ishlash va boshqalarda ko‘rishimiz mumkin.

Dunyo tillarini qamrab olgan parallel korpusning ahamiyati turkiy tillar mushtarak korpusi tuzish masalasining nechog‘liq dolzarbligini ko‘rsatib turibdi. Turkiy tillar mushtarak korpusi matnshunoslik, qiyosiy tilshunoslik, tarjima nazariyasi, adabiyotshunoslik, qarindosh tillararo munosabatlar, til leksikasining boyish manbalarini o‘rganish vositasi sifatida xizmat qilishi bilan ahamiyatli.

Bunday korpusning yaratilishi turkiy tillar oilasiga mansub tillarning rivojlanishini ta’minlashi bilan birga, foydalanuvchilari kam sonli turkiy tillarni asrab qolish garovi hamdir. Turkiy tillarning mushtarak (parallel) korpusi turkiy tillarning mushtarak yodgorliklari – Avesto, O‘rxun-Enasoy obidalari, turkiy xalqlar eposlarini turkiyzabon xalq farzandlariga o‘rgatishning eng zamonaviy ta’lim vositasi sifatida xizmat qilishi tabiiy.

Jahon tillarining juda ko‘pi mukammallik darajasi va matnni ilmiy qayta ishlash imkoniyati bilan farq qiluvchi o‘z milliy korpusiga ega. Ingliz tilida yaratilgan Braun korpusi, Lankaster-Oslo/Bergen (LOB) korpusi, London-Lund korpusi, Leksikografik tadqiqotlar uchun Amerika meros korpusi, Lankaster ingliz tili so‘zlashuv korpusi, diaxronik korpus sanalgan. Ingliz matnlarining Xelsinki korpusi, lingvodidaktik tadqiqotlar uchun Ingliz tili o‘rganuvchilarining xalqaro korpusi, ingliz tilidagi korpuslarning eng so‘ngi avlodi sifatida Ingliz tili banki, Britaniya milliy korpusi, Xalqaro ingliz tili korpusi, Amerika milliy korpusi kabi mashhur korpuslarning mavjudligi milliy va davlat tili taraqqiyotida milliy korpusning ahamiyati va o‘rni nechog’lik muhimligini ko‘rsatadi.

Ilmiy matnlar ko‘pincha tarjima ob’ektiga aylanadi, lekin bu yerda bir qator tushuntirishlar berish kerak. Ko‘plab olimlar - ilmiy matnlarning asosiy qabul

ISSN 2181-3187

qiluvchilari - chet tillarida gaplashadilar. Ko‘pincha olimning o‘zi tinglovchilarining ko‘pchiligiga tanish bo‘lgan tilda yozadi (lotin – o‘rta asrlarda, fransuz yoki nemis - 19 -asrda, ingliz tili - hozirgi paytda). Shuning uchun faqat mumtoz ilmiy asarlar (Darvin, Marks, Karlisl, Sossyur va boshqalar) ko‘p tillarga tarjima qilingan. Shunday qilib, faqat “fan tillari” deb nomlangan matnlar parallel matnlar korpusini olish uchun yetarli miqdordagi matnli materialni ta’minlay oladi va ko‘plab til juftlari uchun parallel ilmiy matnlarni faqat uchinchi til orqali olish mumkin.

Xulosa:

Xulosa o‘rnida shuni aytish mumkinki, har qanday til jamiyatda saqlanib qolishi uchun va ayni navbatda dunyo ilm -fani rivojida o‘zining o‘rnini topishi lozim va bunda parallel korpusning o‘rni nihoyatda katta. Korpus tilshunoslikning istalgan sohasida tadqiqotlarrni olib borishning sifatli va samarali bo‘lishiga yordam beradi. Ilmiy matnlar parallel korpusini yaratishda madaniyatlararo munosabatlar omilini ham hisobga olish zarur.

Foydalanilgan adabiyotlar ro‘yxati:

1. Ruziev, K. B. (2020). Proverbs and corpus linguistics. Актуальные проблемы гуманитарных и естественных наук, (6), 64-67.
2. Ruziev, K. B The Approach of Paremias in Parallel Corpora. JournalNX, 6(05), 216-222.
3. Ruziyev, K. B. (2020). PARALLEL AND COMPARABLE CORPORA. Актуальные научные исследования в современном мире, (11-12), 43-46.
4. Khodjaeva, N. T. (2020). MODIFYING MATERIALS ON LISTENING COMPREHENSION. Актуальные научные исследования в современном мире, (11-12), 28-31.
5. Khodjaeva, N. (2021). TEACHING GRAMMAR AND UNDERSTANDING MEANING IN CONTEXT. InterConf. ACADEMIC RESEARCH IN EDUCATIONAL SCIENCES VOLUME 2 | ISSUE 9 | 2021ISSN: 2181-1385
6. Scientific Journal Impact Factor (SJIF) 2021: 5.723 Directory Indexing of International Research Journals-CiteFactor 2020-21: 0.89DOI: 10.24412/2181-1385-2021-9-515-522
7. Khodjaeva, N. T. (2019). Some Peculiarities And The Ways Of Giving Instructions On Reading Tests. International Journal of Research, 499-505.
8. Ходжаева, Н. Т., & Бахриддинова, М. Ш. (2020). Стилистические характеристики специальных текстов при информативном переводе. Актуальные проблемы гуманитарных и естественных наук, (6), 95-98.
9. https://shokiryuldash.blogspot.com/2020/06/korpus-nima-korpustilshunosligi.html
10. André Santos A survey on parallel corpora alignment. Proceedings of MI-Star, 2011