TURKIY TILLARDA KORPUSLARNING KENG KO‘LAMI VA ULARNING RIVOJLANISHI

Eshmamatova Mahliyo
O‘zbekiston Milliy universiteti
O‘zbek filologiyasi Kompyuter lingvistikasi
va amaliy tilshunoslik kafedrasi stajyor-o‘qituvchisi.
E-mail: meshmamatova21@gmail.com

Annotatsiya: Umumiy korpus yozma yoki og‘zaki matnlarni o‘z ichiga olishi mumkin yoki ikkala ommaviy axborot vositalaridan matnlarni saralab oladi. Yaqinda biz yozish texnologiyalari va avtomatlashtirilgan transkripsiya dasturiy ta’minotining takomillashuvi tufayli og‘zaki korpuslar sonining ko‘payishiga guvoh bo‘ldik. Tilshunoslar til o‘zgarishini va korpus ma’lumotlaridagi til o‘zgarishini kuzatish uchun, birinchi navbatda, ma’lum bir nuqtada yoki vaqt oralig‘ida tildan foydalanishning “oniy tasvirini” taqdim etish maqsadida sinxron korpusini yaratdilar. Monitor korpusi bu yerda taqdim etilgan boshqalardan yangi materiallar qo‘shilishi bilan doimiy ravishda o‘sib borishi (dinamik) ma’nosida farq qiladi. Tilning o‘z vaqtida lingvistik hisobi uchun tuzilgan korpuslar odatda tekshirilayotgan tilning turli davrlarida tildan foydalanishni ifodalovchi matnlarni o‘z ichiga oladi.
Kalit so‘z: korpus,metama’lumot,umumiy korpus, NLP, lug‘at.

Abstract: A general corpus may include written or spoken texts, or a selection of texts from both media. Recently, we have seen an increase in the number of spoken corpora due to improvements in writing technologies and automated transcription software. linguists have created synchronic corpora to track language change and language change in corpus data, primarily to provide a "snapshot" of language use at a specific point or time period. The monitor case differs from the others presented here in the sense that it is constantly growing (dynamic) with the addition of new materials. Corpora constructed for a linguistic account of a language in time usually contain texts representing language use in different periods of the language under investigation.
Keywords: corpus, metadata, general corpus, NLP, dictionary.

Аннотация: Общий корпус может включать письменные или устные тексты или подборку текстов с обоих носителей. В последнее время мы наблюдаем увеличение количества устных корпусов благодаря усовершенствованиям в технологиях письма и программном обеспечении для автоматической транскрипции. лингвисты создали синхронные корпуса для отслеживания изменений языка и языковых изменений в данных корпуса, в первую очередь для того, чтобы предоставить «моментальный снимок» использования языка в определенный момент или период времени. Корпус монитора отличается от других представленных здесь тем, что он постоянно растет (динамично) с добавлением новых материалов. Корпуса, созданные для лингвистического описания языка во времени, обычно содержат тексты, представляющие использование языка в различные периоды исследуемого языка.
Ключевые слова: корпус, метаданные, общий корпус, НЛП, словарь.

Turk tilshunosligi tadqiqotlarida korpusdan foydalanish, yaqinda paydo bo‘lgan sohalardan biridir. Ushbu kechikishning asosiy sabablaridan biri 2000 - yillarning boshlariga qadar hech qanday lingvistik korpus mavjud emasligidir. Jarayonda ishtirok etayotgan ko‘pchilikka ma’lumki, korpus qurish ancha mehnat va ko‘p vaqt talab qiladigan faoliyat bo‘lib, u qat’iy institutsional yordamni talab qiladi. Korpusni qurishda birinchi navbatda e’tiborga olinadigan ba’zilari umumiy dizaynga taalluqlidir: masalan, kiritilgan matn turlari, matnlar soni, alohida matnlarni tanlash, matnlar ichidan matn namunalarini tanlash va matn namunalari uzunligi. Ularning har biri ongli ravishda namuna olish qarorini o‘z ichiga oladi.

• **Umumiy korpus:** Umumiy korpusni qurishda harakatlantiruvchi kuch teng va nomoyon bo‘lishi mumkin bo‘lgan tildan foydalanishning mos yozuvlar korpusini yaratishdir. Umumiy korpus yozma yoki og‘zaki matnlarni o‘z ichiga olishi mumkin yoki ikkala ommaviy axborot vositalaridan matnlarni saralab oladi. Asosiy maqsad - turli janrlar, sohalar va turlardagi matnlarni teng tarzda taqdim etish, natijada xulosalar shuni ko‘rsatadiki korpus ma’lumotlarining miqdoriy va sifat tahlillaridan olingan ma’lumotlar tildan foydalanish uchun to‘g‘ri keladi. Britaniya Milliy Korpusi (BNC) 100 million so‘zni o‘z ichiga olgan va 4048 yozma matn va o‘n million so‘z transkripsiya qilingan og‘zaki ma’lumotlarni o‘z ichiga olgan zamonaviy ingliz tilining shunday umumiy ma’lumot korpusidir[Abdurakhmonova, N. 2021,2022]. Respublika va mintaqaviy gazeta va jurnallar, ommabop va akademik kitoblar, universitet insholari, elektron pochta namunalari, nashr etilmagan xatlar va turli yoshdagi, muassasalar va kitobxonlarning hisobotlari. BNC muvaffaqiyati vakili va muvozanatli umumiy korpus sifatida boshqalarni uning asosiy dizayn tamoyillarini, jumladan Amerika Milliy Korpusini, Koreya Milliy Korpusini, Polsha Milliy Korpusini va yaqinda Turkiya Milliy Korpusini qabul qilishga olib keldi.

• **Ixtisoslashgan korpus:** Nisbatan kichik o‘lchamli va janr yoki soha bo‘yicha ixtisoslashgan, bu turdagi korpuslar ko‘proq xilma - xil va ko‘proq mavjud. Ixtisoslashgan korpus yaratishdagi hozirgi uslublar asosan professional va akademik sohalarda kuzatiladi. Bunday ixtisoslashgan korpusning ba’zi turlari orasida Professional Spoken American English Corpus (CPSA) va Michigan Corpus of Academic Spoken English (MICASE) kiradi. Ixtisoslashgan korpus kattaroq umumiy korpusdan tegishli matn ma’lumotlarini olish orqali ham yaratilishi mumkin.

• **Yozma korpus:** Jigarrang korpus nafaqat birinchi korpus, balki ayni paytda ingliz tilining zamonaviy davrdagi birinchi yozma korpusidir. Korpus ma’lumotlarini tashkil etuvchi matnlar yozma ommaviy axborot vositalaridan to‘plangan.

• **Og‘zaki korpus:** Umumiy yoki yozma korpus bilan solishtirganda, tilning og‘zaki korpusini qurish va izohlash qiyinroq. Yaqinda biz yozish texnologiyalari va avtomatlashtirilgan transkripsiya dasturiy ta’minotining takomillashuvi tufayli og’zaki korpuslar sonining ko‘payishiga guvoh bo‘ldik. Og‘zaki ingliz tili uchun korpuslar 1960-yillarning oxirida yaratilgan masalan, London - Lund Corpus (LLC) (Greenbaum and Svartvik 1990), keyin esa boshqalar, jumladan Lancaster/IBM Spoken English Corpus (SEC), Kembrij va Nottingem. Ingliz tilidagi nutqlar korpusi (CANCODE) (Karter va MakKarti 2004). Turk tilining yagona mavjud va lingvistik jihatdan ishonchli yangi avlod so‘zlashuv korpusi Og‘zaki Turk korpusidir (Ruhi et al.2010). Turk Milliy Korpusida (TNC) (Aksan et al. 2012) ham mavjud bir million so‘zning og‘zaki komponenti BNC dizayn tamoyillariga rioya qilishning aksi sifatida.

• **Sinxron korpus:** tilshunoslar til o‘zgarishini va korpus ma’lumotlaridagi til o‘zgarishini kuzatish uchun, birinchi navbatda, ma’lum bir nuqtada yoki vaqt oralig‘ida tildan foydalanishning

“oniy tasvirini” taqdim etish maqsadida sinxron korpusini yaratdilar. Bunday korpuslarda barcha matnlargacha bo‘lgan vaqt oralig‘idan tanlanishi kerak, sinxron ravishda mavjud bo‘lgan til turlarini hisobga olish. Ingliz tilining xalqaro korpusi (ICE) Buyuk Britaniya, AQSh, Avstraliya, Kanada va ingliz tili birinchi til bo‘lgan boshqa mamlakatlarda so‘zlashuvchi ingliz tilini sinxron tahlil qilish uchun qurilgan (Greenbaum 1991). U bir millionlik yigirma korpusdan iborat.

• **Diaxronik korpus:** Tilning o‘z vaqtida lingvistik hisobi uchun tuzilgan korpuslar odatda tekshirilayotgan tilning turli davrlarida tildan foydalanishni ifodalovchi matnlarni o‘z ichiga oladi. Ovoz yozish texnologiyalarining so‘nggi tarixini hisobga olsak, diaxronik korpus vaqt o‘tishi bilan yozma tilni ifodalaydi, masalan, Xelsinki diaxronik korpusi inglizcha matnlar (Rissanen va boshq. 1991) Learner Corpora: Til sinfida korpusdan foydalanish, o‘qitish va o‘rganish kontekstlarida o‘z o‘rnini topdi. Masalan, International Corpus of Learner English (ICLE) (Granger 2003) va uning subkorpusi sifatida Turk Xalqaro O‘rganuvchi Corpus of English (TICLE) (Kilimci and Can 2009) so‘nggi yillarda o‘qitish kontekstida tadqiqot manbai bo‘lib kelgan.

• **Monitor korpusi:** Monitor korpusi bu yerda taqdim etilgan boshqalardan yangi materiallar qo‘shilishi bilan doimiy ravishda o‘sib borishi (dinamik) ma’nosida farq qiladi. Bank of English (BoE) va Corpus of Contemporary American English (COCA) (Devis 2008) ingliz tili uchun ushbu turdagi taniqli korpuslardir. Korpus, birinchi navbatda, tilni haqiqiy ma’lumotlar asosida empirik tarzda o‘rganish uchun tildan foydalanishni muvozanatli tarzda ifodalash uchun tuzilgan. Tilshunoslik tahlillarida korpusning roli va funksiyasiga qarab, turli nuqtai nazardan qarash mumkin, tadqiqot savollari bo‘yicha. Lüdeling and Kytö lingvistik tahlillarda korpusdan foydalanishni uchta asosiy maqsadda umumlashtiradi:

1) **empirik yordam**
2) **chastotali ma’lumot**
3) **metama’lumot**

Korpus qidiruv tizimi tadqiqotchilarga yordam beradi. Ularning savollariga mos bo‘lgan haqiqiy tildan foydalanish uchun ma’lumotlar topish, ya’ni ular hozirda muqobil emas, balki haqiqiy tildan foydalanish uchun ma’lumotlar keltirish mumkin[Abdurakhmonova, N. 2021,2022]. Tilning tuzilishi va ishlatilishiga oid dalillarni korpusdan topish, lingvistik tahlilning ma’lum darajasi bilan chegaralanib qolmaydi, balki tovushdan shaklgacha va funktsiyaga qadar barcha darajalarda ishlaydi. Korpusdagi ma’lumotlar saralanadi va izohlanadi, shuning uchun namuna olishning aniq turini ta’minlaydi. Bu gipotezalarni empirik tarzda tasdiqlaydi. Haqiqiy til namunalari jamlanmasi sifatida korpus so‘rovi tildan foydalanishning ilgari ko‘zda tutilmagan iqtiboslarini qaytaradi. Korpusdan olingan iqtiboslar shunchaki ma’lum bir lingvistik ko‘rinishni ifodalamaydi, balki miqdoriy ma’lumotni ham beradi.

Bugungi kunda turkiy tilda kamida uch xil turdagi korpus mavjudligini aytib o‘tishimiz mumkin:

1) katta o‘lchamli umumiy tilshunoslik korpusi

2) aniq tadqiqot savollarini o‘rganish uchun tuzilgan va faqat quruvchilar uchun mo‘ljallangan kichik o‘lchamli ixtisoslashtirilgan korpuslar

3) hech qanday til mezonlarisiz qurilgan NLP korpuslari yaratilgan va foydalanuvchilar uchun taqdim etilgan, ong balki turli ilovalar uchun ishlab chiqilgan algoritmlarni sinab ko‘rish vositalari sifatida

Turk tilini boshqa tillarga nisbatan yaxshi o‘rganilgan til deb ayta olmaymiz, uning tarixi va grammatikasi yaxshi hujjatlashtirilgan. Hozirgi turkshunoslikdagi aksariyat tilshunoslik asarlari

nutq tahlili, pragmatika yoki sintaksis kabi oz sonli sohalarga qaratilgan. Biz kamdan - kam hollarda semantika yoki leksikologiya yoki boshqa sohalardagi ishlarni topamiz, chunki ular boyitilgan ma'lumotlar to'plamini talab qiladi. Jumladan, yangi tashkil etilgan Respublikaning dastlabki yillarida leksikaning turkiylashuvidan kelib chiqqan holda, o'sha davrdagi mavjud shevalardan lug'at tuzishga qaratilgan. Lug'at XIII asrdan boshlab 160 ga yaqin turli tarixiy matnlardan turkiy tilga oid lug'aviy birikmalarni jamlagan. Juda kam nazariy tadqiqotlardan tashqari, deyarli barcha lingvistik tahlillar empirik va ma'lumotlarga asoslangan. Turk tilshunosligidagi tipik tadqiqot savolni tahlil qilishda " ma'lumotlar bazasi " yoki "ma'lumotlar to'plami" ni jamlaydi. Aytishimiz mumkinki, deyarli barcha foydalanishga asoslangan empirik tadqiqotlarda qo'llaniladigan juda kichik o'lchamli maxsus korpuslar mavjud. Biroq, ular o'zlarining shakli va hajmi bilan qattiq cheklangan, ular boshqa tadqiqotchilar uchun mavjud emas va ma'lumotlar muayyan muammo bilan to'plangan. Turk tilidagi hisoblash tilshunosligi bo'yicha ish turkiy korpus lingvistikasiga qaraganda uzoqroq tarixga ega. Turk tilida korpus tadqiqotining dastlabki boshlanishi NLP tadqiqotlari va hisoblash tilshunosligi tahlillari bilan bog'liq edi. Hisoblash tilshunosligida va NLPda keng ko'lamli korpuslar "amaliy" maqsadlar uchun qurilgan. Turkiyada hisoblash tilshunosligining keng qamrovli tarixi hali yozilmagan; ammo, bu sohada oldingi ishlar haqida ba'zi – ba'zi joylarda havolalar bor. Lingvistik tahlil uchun ma'lum bo'lgan birinchi elektron korpus Koksal (1976) tomonidan "avtomatik morfologik tahlil" uchun yaratilgan. Koksal o'z algoritmini kundalik gazetalardan tasodifiy tanlangan 1534 so'zli matn namunasi korpusida sinab ko'rdi va baholadi. Koksalning ishi turkiy tillarning boy morfologiyasi va mumkin bo'lgan morfema birikmalarini tan oladi, shuningdek, kelajakdagi asosiy vazifalarga ishora qiladi, qo'llashning potentsial sohalariga e'tibor qaratadi, avtomatlashtirilgan til tahlillari uchun kattaroq korpus yaratishga undaydi. Zamonaviy turk tilini ifodalash uchun ishlab chiqilgan va tuzilgan birinchi elektron lingvistik korpus.

Foydalanilgan adabiyotlar:

1. Leech, G. 100 million words of English: The British National Corpus (BNC). Language research, 1992.
2. Xiao, Z., McEnery, A. Two approaches to genre analysis: Three genres in modern American English. Journal of English Linguistics, 33(1), 2005. 62-82.
3. Gomez, I. F. Interaction in academic spoken English: the use of 'I'and 'you'in the MICASE. Information technology in languages for specific purposes: Issues and prospects, 2006. 35-51.
4. Rissanen, M., Ihalainen, O., Kytö, M. The Helsinki Corpus of English Texts. Diachronic and Dialectal. Helsinki:1991.
5. Granger, S. The international corpus of learner English: a new resource for foreign language learning and teaching and second language acquisition research. Tesol Quarterly, *37*(3), 2003538-546.
6. Kilimci, A. Negotiation of meaning in L2 academic writing. In Proceedings of 1st international conference on literature, languages and linguistics. Athens:2009.
7. Davis, M. The Corpus of Contemporary American English (COCA): One Billion Words, 1990–2019.
8. Абдурахмонова, Н., & Абдувахобов, Г. (2021). O 'quv lug 'atini tuzishning nazariy metodologik asoslari. *Международный журнал искусство слова*, *4*(6).

9. Abdurakhmonova, N. (2021). Formal-Functional Models of The Uzbek Electron Corpus. *ANGLISTICUM. Journal of the Association-Institute for English Language and American Studies*, *10*(8), 59-66.
10. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language Processing. In *2022 7th International Conference on Computer Science and Engineering (UBMK)* (pp. 73-75). IEEE.
11. 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.
12. Kytö, M., & Lüdeling, A. (Eds.). Corpus linguistics: an international handbook. De Gruyter Mouton.:2009.
13. Köksal, F. J coupling measurements using Carr—Purcell spin-echo techniques. Chemical Physics Letters, *44*(1), 1976187-189.
14. Aksan, M., & Aksan, Y. Лингвистические корпуса: взгляд с турецкого языка. В обработке турецкого естественного языка.Шпрингер, Чам: 2018. (стр. 291-315)
15. Мирзахалов Дж., Бабу А., Атаман Д., Кариев С., Тайерс Ф., Абдурауфов О.,Челлаппан С. Масштабное исследование машинного перевода на тюркских языках, 2021.