

TIL KORPUSLARI

Salohiddinova Muslima

*ToshDO‘TAU Kompyuter lingvistikasi
mutaxassisligi 1-kurs magistranti*

Annotatsiya: Mazkur maqolada til korpuslari tushunchasi, ularning lingvistik tadqiqotlardagi o‘rni va ahamiyati yoritilgan. Maqolada korpus lingvistikasi zamonaviy tilshunoslikning empirik yo‘nalishi sifatida talqin qilinib, til korpuslarining tuzilishi, annotatsiya turlari hamda ularning tabiiy tilni qayta ishlash jarayonlaridagi roli tahlil etiladi. Shuningdek, jahon tilshunosligida yaratilgan yirik korpus loyihalari va o‘zbek tilshunosligida olib borilayotgan amaliy korpus tadqiqotlari misolida elektron korpuslarning ilmiy-amaliy ahamiyati ochib beriladi.

Kalit so‘zlar: til korpusi, korpus lingvistikasi, annotatsiya, lingvistik teglash, elektron korpus, tabiiy tilni qayta ishlash.

Abstract: This article discusses the concept of linguistic corpora and their role in modern linguistic research. Corpus linguistics is considered as an empirical approach in contemporary linguistics, and the structure of language corpora, types of annotation, and their significance in natural language processing are analyzed. The article also highlights the scientific and practical value of electronic corpora through examples of major international corpus projects and recent applied corpus studies in Uzbek linguistics.

Keywords: language corpus, corpus linguistics, annotation, linguistic tagging, electronic corpus, natural language processing.

So‘nggi yillarda tilshunoslik fanida empirik yondashuvning ustuvor ahamiyat kasb etishi tilni real nutq materiallari asosida o‘rganish zaruratini yuzaga chiqardi. An’anaviy tavsifiy va intuitiv metodlar bilan bir qatorda, statistik va kompyuterga asoslangan tadqiqotlar keng qo‘llanila boshladi. Aynan shu jarayonda korpus lingvistikasi shakllandi va zamonaviy tilshunoslikning muhim ilmiy yo‘nalishlaridan biriga aylandi. Korpus lingvistikasi kompyuter lingvistikasining tarkibiy qismi bo‘lib, til korpuslarini yaratish, ularni tahlil qilish va lingvistik tadqiqotlarda qo‘llashning nazariy hamda amaliy masalalarini o‘rganadi [Zaxarov, Mengliyev, Hamroyeva, 2023: 4].

Til korpusi tushunchasi zamonaviy tilshunoslikda markaziy o‘rin egallaydi. Til korpusi deganda muayyan tilga oid yozma yoki og‘zaki matnlarning elektron shaklda, maxsus dasturiy ta’minot asosida yig‘ilgan, saqlangan va qidiruv imkoniyatiga ega bo‘lgan majmuasi tushuniladi. Bunday matnlar real nutq namunalariga asoslanadi va til hodisalarini tabiiy holatda tahlil qilish imkonini beradi. Til korpusi oddiy elektron

matnlar to'plami emas, balki lingvistik annotatsiyaga ega bo'lgan, tadqiqot uchun mo'ljallangan maxsus axborot tizimidir [Zaxarov, Mengliyev, Hamroyeva, 2023: 5]. Til korpusi oddiy elektron matnlar to'plamidan tubdan farq qiladi, chunki u faqat matnlarni saqlashni emas, balki ularni ilmiy tahlil qilishni maqsad qilgan maxsus axborot tizimi hisoblanadi. Elektron matnlar to'plamida matnlar ko'pincha faqat o'qish yoki qidirish uchun mavjud bo'lsa, til korpusida har bir matn lingvistik annotatsiya bilan ta'minlanadi. Annotatsiya jarayonida so'zlar va gap birliklariga grammatik, leksik, sintaktik yoki semantik ma'lumotlar biriktiriladi. Bu esa til birliklarining real nutqdagi qo'llanilishini tizimli va avtomatlashtirilgan tarzda o'rganish imkonini beradi. Shu sababli til korpusi lingvistik tadqiqotlar uchun empirik ma'lumotlar bazasi vazifasini bajaradi hamda leksikologiya, grammatika, uslubiyat va kompyuter lingvistikasi sohalarida keng qo'llaniladi.

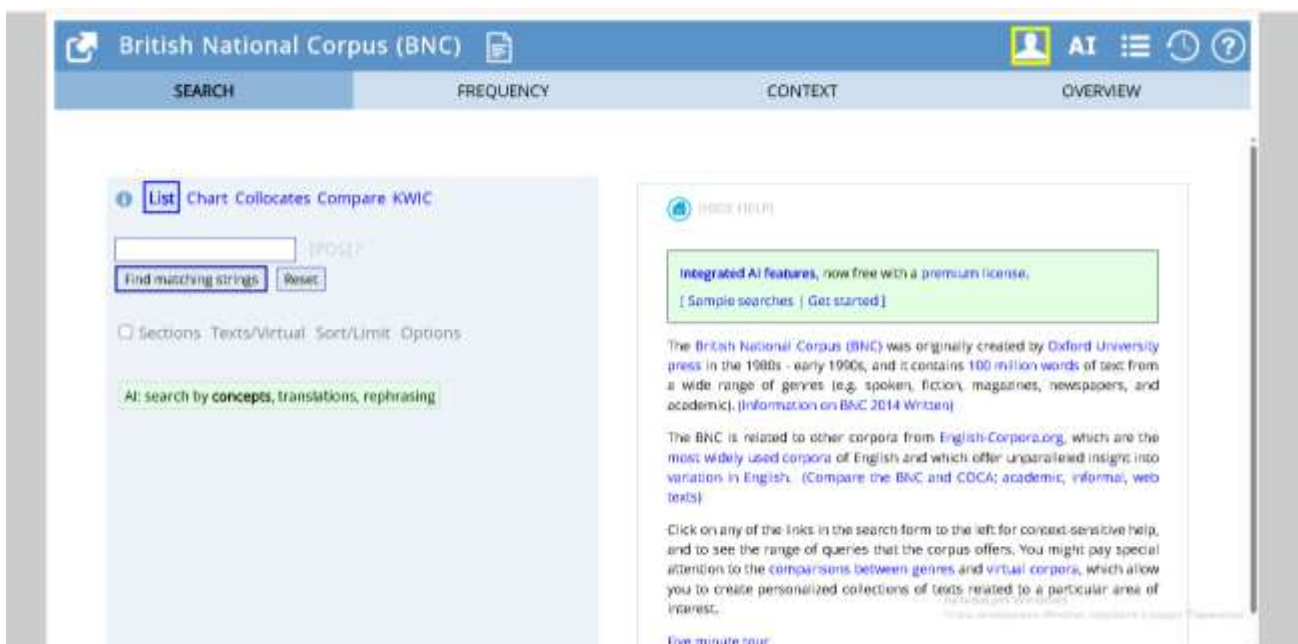
Til korpusi elektron kutubxonadan tubdan farq qiladi. Elektron kutubxona asosan badiiy va publitsistik matnlarni saqlashni maqsad qilsa, til korpusi lingvistik birliklarning shakli, ma'nosi va funksiyasini aniqlash uchun yaratiladi. Korpus matnlari morfologik, sintaktik, semantik va ekstralingvistik belgilar bilan teglanadi, bu esa murakkab lingvistik so'rovlarni amalga oshirish imkonini beradi. Shu jihatdan korpus lingvistik tadqiqotlar uchun ishonchli va samarali empirik baza hisoblanadi [Sinclair, 1991]. Korpus matnlarining morfologik, sintaktik, semantik va ekstralingvistik belgilar bilan teglanishi lingvistik tahlil imkoniyatlarini sezilarli darajada kengaytiradi. Morfologik teglash orqali so'z turkumlari, grammatik shakllar va so'z yasalishi kabi xususiyatlar aniqlanadi, sintaktik teglash esa gap tuzilishi va so'zlararo bog'lanishlarni tavsiflashga xizmat qiladi. Semantik teglash matn birliklarining ma'no jihatlarini aniqlash imkonini bersa, ekstralingvistik belgilar matnning muallifi, yaratilgan vaqti, janri va kommunikativ vaziyati kabi qo'shimcha ma'lumotlarni qamrab oladi. Bunday ko'p qatlamli annotatsiya tizimi murakkab lingvistik so'rovlarni, jumladan, muayyan grammatik shakl yoki ma'no toifasiga oid birliklarni avtomatik tarzda qidirish va tahlil qilish imkonini yaratadi. Natijada til korpusi lingvistik tadqiqotlar uchun ishonchli va samarali empirik baza sifatida xizmat qiladi.

Til korpusining asosiy tarkibiy qismlari va ularning vazifalari keltirilgan. Jadval til korpusining qanday elementlardan tashkil topishini, shuningdek, matnlar bazasi, lingvistik annotatsiya qatlamlari va qidiruv tizimi o'rtasidagi o'zaro bog'liqlikni umumiy ko'rinishda aks ettiradi. Ushbu tarkibiy qismlar til korpusini oddiy elektron matnlar to'plamidan farqlovchi muhim belgilardan hisoblanadi va uning lingvistik tadqiqotlar uchun mo'ljallangan maxsus axborot tizimi ekanligini ko'rsatadi.

Tarkibiy qism	Tavsifi
---------------	---------

Matnlar bazasi	Yozma va og‘zaki nutq namunalaridan iborat bo‘lib, turli janr, uslub va davr matnlarini qamrab oladi
Morfologik annotatsiya	So‘z turkumlari, grammatik shakllar va lemmatizatsiya haqidagi ma’lumotlarni o‘z ichiga oladi
Sintaktik annotatsiya	Gap tuzilishi, so‘zlararo bog‘lanishlar va sintaktik munosabatlarni belgilaydi
Semantik annotatsiya	So‘z va birikmalarning ma’no jihatlarini, semantik toifalarini aks ettiradi
Ekstralingvistik ma’lumotlar	Matn muallifi, yaratilgan vaqti, janri, hududi va kommunikativ vaziyati haqida axborot beradi
Qidiruv va tahlil tizimi	Korpusdan murakkab lingvistik so‘rovlar asosida ma’lumot olish imkonini yaratadi

Korpus lingvistikasi rivoji jahon tilshunosligida muhim bosqichlarni bosib o‘tgan. Ingliz tilidagi Brown Corpus va British National Corpus kabi loyihalar tilni statistik va kontekstual jihatdan o‘rganishda yangi davrni boshlab berdi. McEnery va Wilson ta’kidlashicha, korpusga asoslangan tadqiqotlar tilning haqiqiy ishlatilish manzarasini ochib beradi va nazariy xulosalarning ishonchligini oshiradi [McEnery, Wilson, 2001].



1-rasm. British National Corpus (BNC)

Shuningdek, Biber korpus ma’lumotlari asosida yozma va og‘zaki nutq o‘rtasidagi farqlarni aniqlash mumkinligini ilmiy jihatdan isbotlagan [Biber, 1998: 32].

Til korpuslari tuzilishi va maqsadiga ko'ra turli turlarga bo'linadi. Nutq shakliga ko'ra yozma, og'zaki va aralash korpuslar ajratiladi. Zamonaviy milliy korpuslar, odatda, aralash bo'lib, yozma matnlar bilan bir qatorda og'zaki nutq namunalari hamni o'z ichiga oladi. Tiliga ko'ra korpuslar bir tilli va ko'p tilli bo'lishi mumkin, parallel korpuslar esa mashina tarjimai va qiyosiy tilshunoslikda muhim ahamiyatga ega [Zaxarov, Mengliyev, Hamroyeva, 2023: 25].

Til korpusining eng muhim xususiyatlaridan biri uning annotatsiyalanganligidir. Annotatsiya matnga qo'shimcha lingvistik ma'lumot berish jarayoni bo'lib, u orqali so'z turkumi, grammatik shakl, sintaktik bog'lanish yoki semantik rol aniqlanadi. Annotatsiyalangan korpuslar yordamida nafaqat alohida so'zlar, balki grammatik konstruksiyalar va semantik tuzilmalarni ham avtomatik tahlil qilish mumkin. Bu esa tabiiy tilni qayta ishlash, mashina tarjimai va sentiment tahlili kabi yo'nalishlar uchun muhim hisoblanadi [Jurafsky, Martin, 2023: 67]. Shu bois annotatsiya korpusning funksional imkoniyatlarini belgilovchi asosiy omillardan biri hisoblanadi. Annotatsiya darajasi va turi qanchalik mukammal bo'lsa, korpus asosida bajariladigan lingvistik va hisoblashga asoslangan tahlillarning aniqligi ham shunchalik yuqori bo'ladi.

Til korpuslarining amaliy ahamiyati juda kengdir. Ular leksikografiyada zamonaviy lug'atlar tuzishda asosiy manba bo'lib xizmat qiladi, chunki korpusdagi misollar real nutqdan olingan va ishonchli hisoblanadi. Bundan tashqari, til korpuslari til o'qitishda, ayniqsa, chet tillarini o'rgatishda real kontekstga asoslangan misollar taqdim etish imkonini beradi. Korpuslar yordamida o'quvchi til birliklarining haqiqiy qo'llanilishini ko'rishi va til kompetensiyasini oshirishi mumkin [Gries, 2009: 25].

O'zbek tilshunosligida korpus lingvistikasi nisbatan yangi yo'nalish bo'lsa-da, so'nggi yillarda bu borada muhim ilmiy ishlar amalga oshirilmoqda. O'zbek tilshunosligida korpus lingvistikasi nisbatan yangi yo'nalish bo'lsa-da, so'nggi yillarda bu sohada amaliy yo'naltirilgan ilmiy ishlar amalga oshirilmoqda. Xususan, o'zbek tilidagi badiiy, publitsistik va ilmiy matnlar asosida mualliflik korpuslari shakllantirilgan hamda ushbu korpuslarda matnlarni elektron formatda yig'ish, belgilash va qidirish tizimlari joriy etilgan. Ayrim tadqiqotlarda o'zbek tilidagi matnlar morfologik teglashdan o'tkazilib, so'z turkumlari va grammatik shakllarni avtomatik aniqlash tajribalari amalga oshirilgan. Shuningdek, korpus materiallari asosida chastotali lug'atlar tuzish, so'zlarning real nutqdagi qo'llanilishini statistik tahlil qilish va lingvistik birliklarning kontekstual xususiyatlarini aniqlash bo'yicha amaliy natijalar olingan. Bu kabi ishlar o'zbek tilida elektron korpuslardan ilmiy tadqiqotlarda foydalanish imkoniyatlarini kengaytirib, ularni lingvistik va kompyuter texnologiyalari bilan bog'lashga xizmat qilmoqda.

Xulosa qilib aytganda, til korpusi zamonaviy tilshunoslikning ajralmas qismi bo'lib, tilni empirik asosda o'rganish imkonini beradi. Korpus lingvistikasi til birliklarining real qo'llanilishini ochib berib, nazariy va amaliy tadqiqotlar

samaradorligini oshiradi. O'zbek tilining milliy korpuslarini yaratish va ularni rivojlantirish esa nafaqat filologiya, balki sun'iy intellekt va raqamli til texnologiyalari rivoji uchun ham muhim ahamiyat kasb etadi.

Foydalanilgan adabiyotlar:

1. Zaxarov V.P., Mengliyev B.R., Xamroyeva Sh.M. Korpus lingvistikasi. – Toshkent, 2023.
2. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991.
3. McEnery T., Wilson A. Corpus Linguistics: An Introduction. – Edinburgh: Edinburgh University Press, 2001.
4. Biber D. Variation across Speech and Writing. – Cambridge: Cambridge University Press, 1998.
5. Jurafsky D., Martin J.H. Speech and Language Processing. – Stanford University, 2023.
6. Gries S.Th. Quantitative Corpus Linguistics with R. – New York: Routledge, 2009.