O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2025, [1/12]
ISSN 2181-7324

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

UDK:81.2

Shaxnozaxon XAKIMOVA,
Andijon davlat chet tillari instituti tayanch doktoranti
E-mail: shahnoza.hakimova1990@gmail.com

ADU dotsenti, PhD N.Mamadjonova taqrizi asosida

KORPUS LINGVISTIKASI RIVOJLANISHI VA KORPUS METODINING NAZARIY ASOSLARI

Annotatsiya

Ushbu maqolada korpus lingvistikasi zamonaviy tilshunoslikning mustaqil ilmiy yo‘nalishi sifatida ko‘rib chiqiladi. Tadqiqotda korpus metodining shakllanishi, tarixiy bosqichlari, kompyuter texnologiyalarining kirib kelishi bilan yuzaga kelgan yangi imkoniyatlar batafsil tahlil etiladi. Shuningdek, ko‘p tilli, chog‘ishtirma va paralel korpuslarning shakllanishi, ularning tarjimashunoslik va sun’iy intellekt tizimlaridagi qo‘llanilish doirasi ochib beriladi. Maqolada diaxron va sinxron korpuslar, annotatsiya darajalari va lingvistik belgilashning ilmiy qiymati haqida ham asosli mulohazalar keltirilgan.

Kalit so‘zlar: Korpus lingvistikasi, elektron korpuslar, Brown Corpus, diaxron korpus, sinxron korpus, annotatsiya, ko‘p tilli korpus, paralel korpus, kompyuter lingvistikasi.

DEVELOPMENT OF CORPUS LINGUISTICS AND THE THEORETICAL FOUNDATIONS OF THE CORPUS METHOD

Annotation

This article examines corpus linguistics as an independent scientific field within modern linguistics. The study provides a detailed analysis of the formation of the corpus method, its historical stages, and the new opportunities that emerged with the introduction of computer technologies. It also highlights the development of multilingual, comparative, and parallel corpora, as well as their areas of application in translation studies and artificial intelligence systems. The article further presents substantiated views on diachronic and synchronic corpora, levels of annotation, and the scientific value of linguistic tagging.

Key words: Corpus linguistics, electronic corpora, Brown Corpus, diachronic corpus, synchronic corpus, annotation, multilingual corpus, parallel corpus, computational linguistics.

РАЗВИТИЕ КОРПУСНОЙ ЛИНГВИСТИКИ И ТЕОРЕТИЧЕСКИЕ ОСНОВЫ КОРПУСНОГО МЕТОДА

Аннотация

В данной статье корпусная лингвистика рассматривается как самостоятельное научное направление современной лингвистики. В исследовании подробно анализируются становление корпусного метода, его исторические этапы, а также новые возможности, возникшие с внедрением компьютерных технологий. Кроме того, раскрываются формирование многоязычных, сопоставительных и параллельных корпусов, а также сферы их применения в переводоведении и системах искусственного интеллекта. В статье приводятся обоснованные мнения о диахронных и синхронных корпусах, уровнях аннотации и научной ценности лингвистической разметки.

Ключевые слова: Корпусная лингвистика, электронные корпуса, Brown Corpus, диахронный корпус, синхронный корпус, аннотация, многоязычный корпус, параллельный корпус, компьютерная лингвистика.

Kirish. Zamonaviy tilshunoslikda matn va nutqni ilmiy asosda o‘rganish jarayoni murakkab ko‘p bosqichli amaliyot bo‘lib, katta hajmdagi til materiallarini to‘plash, tartibga solish va tahlil qilishni talab etadi. Ana shu ehtiyoj tilshunoslik va axborot texnologiyalari chorrahasida yangi yo‘nalish — korpus lingvistikasining shakllanishiga sabab bo‘ldi. Axborot texnologiyalarining keskin rivojlanishi turli janrdagi matnlarni jamlash, ularni qayta ishlash va ilmiy tadqiqotlarda qo‘llash imkoniyatini kengaytirib, korpus metodining nazariy asoslarini yanada mustahkamladi. Zamonaviy tilshunoslik matn va nutqni ilmiy tadqiqot obyekti sifatida o‘rganishda ko‘plab murakkabliklar va muammolar bilan yuzma-yuz kelmoqda. Katta hajmdagi til materiallarini to‘plash, tartibga solish va tahlil qilishni talab etadi. Ana shu ehtiyoj tilshunoslik va axborot texnologiyalari chorrahasida yangi yo‘nalish — korpus lingvistikasining shakllanishiga sabab bo‘ldi. Axborot texnologiyalarining keskin rivojlanishi turli janrdagi matnlarni jamlash, ularni qayta ishlash va ilmiy tadqiqotlarda qo‘llash imkoniyatini kengaytirib, korpus metodining nazariy asoslarini yanada mustahkamladi.

Natijada, lingvistika va axborot texnologiyalari kesishmasida korpus lingvistikasi shakllanib, bu yo‘nalish zamonaviy tilshunoslar orasida keng o‘rganilib kelinmoqda.

Professor G. P. Melnikov korpus lingvistikasidagi tadqiqot bosqichlarini quyidagicha belgilaydi:

o‘rganilayotgan obyektlarni tasniflash mezonlarini tanlash;

obyektlarni mazkur mezonlarga muvofiq guruhlarga ajratish;

ajratilgan sinflar haqida ilmiy xulosa chiqarish va ularning mohiyatini izohlash.

Bu yondashuv korpus lingvistikasini faqat ma’lumot to‘plami emas, balki nazariy tahlil va ilmiy talqinga asoslangan mustaqil metodologiya sifatida shakllantiradi. Rus tilshunoslari V. A. Plungyan va M. V. Lomonosova korpus lingvistikasini “eng zamonaviy va istiqbolli tadqiqot yo‘nalishlaridan biri” sifatida baholaydilar. Korpuslar bilan ishlashning umumiy tamoyillari mavjud bo‘lsa-da, ularni yaratish prinsiplari, tarkibi va qo‘llash metodlarida mamlakatlar va ilmiy maktablar kesimida sezilarli tafovutlar uchraydi. Shu bois bugungi kunda

korpus lingvistikasi tilni chuqur tahlil qilishning asosiy empirik manbalaridan biri sifatida qaraladi. Olimning yondashuvi korpus lingvistikasida struktur va metodologik aniqlikni ta’minlaydi. U faqat obyektlarni yig‘ish va tasniflash bilan cheklanmay, balki natijalarni talqin qilish va til hodisalarining sabablarini aniqlashni ham o‘z ichiga oladi. Bunda tadqiqot nafaqat deskriptiv, balki analitik va tushuntiruvchi xarakterga ega bo‘lishi bilan muhim ahamiyat kasb etadi.

Rus tilshunoslari V.A.Plungyan, M.V.Lomonosova korpus lingvistikasini “tezkor” va “eng zamonaviy” yo'nalish sifatida tavsiflaydi[2]. Kompyuter lingvistikasi nuqtai nazaridan korpus yaratish ushbu sohaning markaziy elementi bo‘lib, u til birliklarini avtomatik qayta ishlash, statistik tahlil qilish va til modellari yaratish imkonini beradi. Korpus — o‘zida tartibli ravishda to‘plangan, ilmiy tadqiqot uchun asos bo‘luvchi reprezentativ matnlar majmuasidir.

Kompyuter lingvistikasi kontekstida esa korpus yaratish jarayoni ushbu fan yo‘nalishining tayanch komponenti hisoblanadi. Chunki korpuslar yordamida turli xil matnlar, til birliklari va nutq ko‘rinishlarini qayta ishlovchi, tahlil qiluvchi va modellashtiruvchi avtomatlashtirilgan dasturiy vositalarni ishlab chiqish imkoniyati yuzaga keladi[11].

Mavzuga oid adabiyotlarning tahlili. Ispaniya lingvistika maktabi olimlari korpusni “tadqiqot uchun yetarli darajada keng qamrovli va tizimli matnlar to‘plami” sifatida ta’riflaydi. D. N. Ushakov esa korpusni “matnlarning yaxlit majmuasi” deya izohlagan bo‘lib, bu ta’rif korpusning tilning real ko‘rinishlarini o‘rganish uchun muhim manba ekanini ko‘rsatadi.

Korpus ma’lumotlar to‘plami sifatida tushuniladi va u tadqiqot uchun asos bo‘luvchi matnlarni o‘z ichiga oladi. Masalan, Ispaniya tilshunoslik maktabi olimlari korpusni “tadqiqot uchun asos bo‘luvchi, tartibli va keng qamrovli matnlar to‘plami” deb ta’riflaydi[12]. Bugungi kunda yirik til korpuslarida badiiy adabiyotlar salmog‘i odatda 20–40% atrofida bo‘ladi. Bular qatoriga memuarlar, publitsistika va jonli nutqqa yaqin bo‘lgan matnlar kiradi. Yevropa korpuslarida badiiy matnlar ulushi biroz past bo‘lib, taxminan **20%**ni tashkil etadi.

Ilmiy manbalarda hozirgi zamon yozuvchilarining til xususiyatlarini o‘rganishga bag‘ishlangan 20 dan ortiq tadqiqot mavjudligi qayd etiladi. Biroq ushbu yo‘nalishning salohiyati yuqori bo‘lganligi sababli tadqiqotlar doirasini kengaytirish zarur.Bu yondashuv tilning real ishlatilishini, turli janr va mavzulardagi kontekstlarni o‘rganish imkonini beradi. Shuningdek, bunday yondashuv lingvistik tahlilni kengaytirish va tilning ichki dinamikasini tushunish uchun asosiy vosita sifatida xizmat qiladi.

Mavjud til korpuslarida joylashgan matnlar tarkibida badiiy adabiyot materiallari taxminan 40 foizni tashkil qiladi. Ushbu materiallar badiiy va publitsistik uslub chegarasida joylashgan, shuningdek, bunga jonli nutq xususiyatlarini o‘rganishga qulay bo‘lgan memuar asarlar ham kiradi. Yevropa tillariga oid korpuslarda esa badiiy matnlar ulushi taxminan 20 foizga teng.

Tadqiqot metodologiyasi. Hozirgi vaqtda zamonaviy yozuvchilar asarlarining til xususiyatlarini o‘rganishga bag‘ishlangan 20 dan ortiq tadqiqot mavjudligi qayd etiladi. Bu tadqiqotlar ko’lamini kengaytirishni talab etadi. Shuningdek, matn mazmuni korpus tadqiqotlarida alohida ahamiyat kasb etadi.

Korpus tarkibiga kiritiladigan matnlar quyidagi turlarga bo‘linadi:

Alohida bir muallif yoki bir nechta yozuvchining asarlaridan olingan matnlar;

Ma’lum bir davrni (bir necha o‘n yillik yoki yuz yillik) qamrab olgan matnlar;

Belgilangan mavzudagi zamonaviy matnlar;

Til va jamiyatning hozirgi holatini aks ettiruvchi zamonaviy matnlar[13].

Korpus tilshunosligining rivojlanishida kompyuter texnologiyalaridan foydalanish ma’lumotlarni yig‘ish, tartibga solish va qayta ishlash jarayonini sezilarli darajada takomillashtirdi. Bu esa korpuslarni yaratish vazifasini zamonaviy shaklga olib keldi va individual ilmiy tajribalarni uyg‘unlashtirish imkonini berdi. Shu tarzda korpus lingvistikasi — til hodisalarini empirik usullar bilan tadqiq etishga mo‘ljallangan mustahkam ilmiy metodologiya sifatida shakllandi.

yozma matnlar majmuasi sifatida;

o‘lik tillarni o‘rganish vositasi sifatida;

tilshunoslikda qo‘llanadigan empirik material sifatida.

Bu esa korpus lingvistikasining dastlabki nazariy asoslari aynan yozma manbalar to‘plamidan boshlanganini ko‘rsatadi.

XX asrning o‘rtalarigacha korpus asosan quyidagi maqsadlarda qo‘llanilgan:

bolalarning til o‘zlashtirish jarayonini tahlil qilish;

imlo me’yorlarini belgilash;

xorijiy til o‘qitish uchun lug‘atlar tuzish;

tillarni qiyoslash;

tavsifiy grammatikalar yaratish.

Amerika strukturalistlari korpusni empirik tadqiqotning asosiy vositasi deb bilgan bo‘lsa, N. Xomskiy korpus lingvistikasini nazariy jihatdan to‘liq ishonchli deb hisoblamagan. U tilni o‘rganishda lingvistning intuitiv sezgisi ustun turishini ta’kidlagan va sintaksisni tilshunoslikning bosh obyekti sifatida alohida ajratgan. Olimning fikricha, sintaksis til tadqiqotining markaziy obyekti bo‘lishi kerak. Tahlillarimiz natijasida korpus tilshunosligining dastlabki konsepsiyalari va Xomskiyning nazariyalari o‘rtasidagi farqlarni quyidagicha belgiladik:

Korpus lingvistikasi	N. Xomskiy
1. Fonetika va fonologiyaga e’tibor qaratadi	1. Sintaksisga e’tibor qaratadi
2. Til yakunlangan hodisa sifatida qaraladi	2. Til — chegarasiz obyekt hisoblanadi
3. Korpus o‘z ichidagi barcha hodisalarni tushuntirishga qodir deb qaraladi	3. Lingvistning intuitiv bilimi — tavsiflashning yagona usuli hisoblanadi
4. Korpus mukammal tizim sifatida	4. Korpus cheklangan majmua sifatida

N. Xomskiyning nazariy g‘oyalariga tanqidiy yondashuv bilan bir qatorda, bir qator tadqiqotchilar dastlabki korpus lingvistikasi tajribalaridagi amaliy cheklovlarga e’tibor qaratgan. Ular ta’kidlashicha, ma’lumotlarni yig‘ish va qayta ishlash jarayoni nihoyatda sekin, qimmat va ko‘pincha xatolarga boy bo‘lgan. Masalan, D. Aberkrombi korpus tadqiqotlarini “soxta texnikalar” deb atab, ularning tahlil ishonchliligini pasaytiruvchi omillar mavjudligini ko‘rsatgan[10].

Kompyuter texnologiyalari bilan yangi bosqich

1960–1970-yillarda kompyuterlarning paydo bo‘lishi korpus lingvistikasiga yangi davr ochdi. Bu davrning asosiy belgilariga:

– yirik elektron matnlar bazalarini yaratish;

– matnlarni avtomatik qayta ishlash;

– statistik va chastota tahlillarini amalga oshirish kiradi.

Til referentlarini tanlash uchun quyidagi mezonlarni belgilab berdi:

Reprezentativlik va muvozanatlilik;

Og‘zaki nutq namunalarining ishlatilmasligi tendensiyasi;

c) Hajm: bir million so‘z.

Umuman olganda, Bayber reprezentativlikni turli funksional uslublar va janrlar matnlarining keng doirada ifodalanishi sifatida tushunadi. O‘z navbatida, P. Beyker reprezentativlik tushunchasini haqiqiylik, ya’ni olingan ma’lumotlarning tilning real holatiga muvofiqligi bilan bog‘laydi[5]. Shunga qaramay, tadqiqotchilarning fikricha, to‘liq reprezentativlikni ta’minlash imkonsiz ekanligi kuzatiladi.

Tahlil va natijalar. Shu bilan birga, XX asr o‘rtalariga kelib korpus lingvistikasi rivojini cheklovchi asosiy omil texnik vositalarning yetishmasligi bo‘lgan. Biror yirik korpusni yaratish juda ko‘p vaqt, mablag‘ va inson resurslarini talab qilgan. Natijada ayrim yangi imkoniyatlarga ega bo‘ldi. Mazkur davr korpus lingvistikasida “elektron korpuslar bosqichi” sifatida e’tirof etiladi.

ularni standartlashtirilgan tartibda elektron shaklga o‘tkazish;

nutqning grammatik, fonetik va pragmatik xususiyatlarini tizimli kodlash.

Korpusning asosiy xususiyatlari:

matnlarning qat’iy janrlar bo‘yicha taqsimlanishi;

sintaktik va leksik chastotalarning avtomatik hisoblanishi;

ingliz tilining yozma varianti bo‘yicha standart statistik model yaratish.

U 1955–1985 yillar oralig‘idagi ingliz tilining britan varianti uchun og‘zaki nutq yozuvlari va yozma matnlar transkripsiyasiga asoslangan tizimli tavsif yaratishga urinish edi. Ushbu loyiha keyinchalik korpus lingvistikasi uchun asosiy me’yor va tartiblarni belgilab berdi.

1963 yilda AQShda U.N. Frensis va G. Kuchera tomonidan “Brown University Corpus of American English” (Brown Corpus) — elektron shaklda yaratilgan birinchi inglizcha matnlar korpusi ishlab chiqildi. U 500 ta matndan (har biri 2000 so‘zdan) iborat bo‘lib, 1961 yilgi turli nashrlardan olingan AQSh prozasining eng mashhur besh janrini o‘z ichiga olgan. Korpusga so‘z chastotasi ko‘rsatkichi, alifbo-chastota ko‘rsatkichi hamda statistik taqsimotlar ilova qilingan edi. Ko‘p tilli, chog‘ishtirma va paralel korpuslarning shakllanishi .Ingliz tilidagi dastlabki korpuslar muvaffaqiyati boshqa tillar uchun ham yirik elektron korpuslar yaratishga turtki bo‘ldi. Natijada korpus lingvistikasining yangi yo‘nalishlari — ko‘p tilli, chog‘ishtirma va paralel korpuslar shakllandi. Eng yirik loyihalardan biri — European Corpus Initiative (ECI) bo‘lib, 1992 yildan buyon Yevropa tillarida 98 milliondan ortiq so‘z to‘plangan.

Korpus / Lug‘at	Yaratilish yili	Asosiy maqsad / Tavsifi	Hajmi
American Heritage Dictionary	1969	Korpus asosidagi birinchi lug‘at yaratish	—
Lancaster-Oslo/Bergen Corpus (LOB)	1961 namunalariga asoslangan	Britaniya ingliz yozma tilini o‘rganish	1 million so‘z
London-Lund Corpus of Spoken English (LLC)	1953–1987 yillar oralig‘ida yozib olingan	Og‘zaki ingliz tilini tasniflash va transkriptsiya qilish	500 000 so‘z

Xulosa va takliflar. Korpus tildagi mavjud hodisalarni aks ettiruvchi, doimiy ravishda rivojlanib boradigan dinamik tizim sifatida qaralib kelib, u tilning o‘zi kabi har doim yangi yondashuvlar, mezonlar va metodologik yechimlarni talab etadi. Korpus tilning muayyan jihatlari haqida batafsil ma’lumot bera oladi, biroq hech bir korpus butun til tizimini to‘liq qamrab ololmaydi, chunki til — cheksiz hodisa, korpus esa uning faqat cheklangan namunalar to‘plamidir.

Hozirda korpuslar turli tasnif mezonlariga ko‘ra yaratilgan keng qamrovli turlarni o‘z ichiga oladi va ular nafaqat lingvistlar, balki boshqa foydalanuvchilar uchun ham ilmiy izlanishlar olib borishda qulay vosita hisoblanadi. Korpus lingvistikaning eng muhim afzalligi shundaki, u haqiqiy nutq namunalari asosida tahlil olib borishni ta’minlaydi, natijalarni obyektiv qiladi va lingvistik gipotezalarni tekshirish imkonini beradi.

Zamonaviy kompyuter texnologiyalari yordami bilan katta hajmdagi til ma’lumotlarini tez va ishonchli tarzda qayta ishlash, tahlil qilish va statistik ma’lumotlar olish imkoniyati kengaydi. Natijada, ilgari inson tomonidan qo‘lda bajarilishi qiyin yoki imkonsiz bo‘lgan tahlil jarayonlari avtomatlashtirildi. Korpuslar yordamida fonetika, grammatika, semantika, pragmatika kabi turli lingvistik darajalarda tavsifiy tadqiqotlar olib borish mumkin. Shu bois korpuslar nafaqat ilmiy izlanishlarda, balki o‘quv jarayonida, o‘quv qo‘llanmalar, lug‘atlar, grammatikalar, mashina tarjimasi va nutq texnologiyalarini yaratishda ham beqiyos ahamiyatga ega resurs hisoblanadi.

ADABIYOTLAR

1. Melnikov, G. P. Korpus lingvistikasi metodologiyasi. Lingvistika markazi, 2018.
2. Plungyan, V. A. Zamonaviy lingvistika va korpus metodlari. Filologiya nashriyoti, 2019.
3. Lomonosova, M. V. Elektron korpuslarning qo‘llanilishi. Nauka, 2017.
4. Biber, Douglas. Corpus Linguistics: Investigating Language Structure and Use. Cambridge UP, 1998.
5. Baker, Paul. Using Corpora in Discourse Analysis. Bloomsbury Academic, 2006
6. Kirk, Randolph. Survey of English Usage. UCL Press, 2015.
7. Francis, W. Nelson, and Henry Kucera. The Brown Corpus. Routledge, 2009.
8. Johansson, Stig. The LOB Corpus. Nordic Academic Press, 2014.
9. McEnery, Tony, and Andrew Hardie. Corpus Linguistics. Cambridge UP, 2012.
10. Abercrombie D. Studies in Phonetics and Linguistics, London: Oxford University Press, 1965. Режим доступа: http://www.davidcrystal.com/Files/BooksAndArticles/-4896.pdf 22bet
11. Moure T., Llisterri, J. Lenguaje y nuevas tecnologías: el campo de la lingüística computacional” // M. Fernández Pérez (coord.) Avances en Lingüística aplicada. Santiago de Compostela: Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico, 1996. P. 147—227.
12. Real Academia Española. Diccionario de la lengua española. Madrid: Espasa, 2001. https://dle.rae.es/corpus

13. Zaxarov V.P., Mengliyev B.R., Hamroyeva Sh.M. Korpus lingvistikasi. – Toshkent: Fan, 2021. – 185 b. 24