

PAPER

JAHON TILSHUNOSLIGIDA KORPUS LINGVISTIKASINING SHAKLLANISHI TARIXI

Shermamatova Madina Obidjon qizi¹, *

¹Samarqand davlat chet tillar instituti O'zbek tili, jurnalistika, kompyuter lingvistikasi kafedrasida magistranti

* shermamatova@gmail.com

Abstract

Ushbu maqolada jahon tilshunosligida korpus lingvistikasining shakllanishi va rivojlanish jarayoni tarixiy, nazariy hamda metodologik jihatdan keng yoritilgan. Tadqiqotda korpus lingvistikasining XIX–XX asrlardagi empirik tilshunoslik, statistik yondashuvlar va kompyuter texnologiyalarining rivoji bilan bog'liq bo'lgan ildizlari tahlil qilingan. Brown korpusi, London–Lund korpusi va Britaniya Milliy Korpusi kabi yirik loyihalarning shakllanishi korpus lingvistikasining alohida ilmiy yo'nalish sifatida mustahkamlanishiga asos bo'lgani ilmiy dalillar asosida ko'rsatib berilgan. Shuningdek, J. Sinclair maktabi tomonidan ilgari surilgan metodologik tamoyillar, kollokatsiya nazariyasi, corpus-based va corpus-driven yondashuvlarning ilmiy mazmuni chuqur tahlil qilingan. XXI asrda big data, sun'iy intellekt va mega-korpuslar davrining kirib kelishi natijasida korpus lingvistikasining zamonaviy yo'nalishlari, xususan, NLP, mashina tarjimasi, semantik modellashtirish va real vaqtli til tahlili kabi sohalaridagi o'rni alohida ko'rib chiqilgan.

Key words:

korpus lingvistikasi, Brown korpusi, BNC, J. Sinclair, kollokatsiya, empirik tilshunoslik, big data, NLP, sun'iy intellekt.

KIRISH

Korpus lingvistikasi XX asr oxirida alohida ilmiy yo'nalish sifatida tan olingan bo'lsa-da, uning ildizlari ancha chuqur – XIX asr oxiri va XX asr boshlaridagi empirik tilshunoslik, statistik yondashuvlar, strukturalizm hamda kompyuter

texnologiyalarining shakllanishiga borib taqaladi. Mazkur bo'limda korpus lingvistikasining paydo bo'lishi uchun zamin yaratgan ilmiy, nazariy va texnologik omillar izchil yoritiladi.

XIX asr–XX asr boshlarida empirik yondashuvlarning kuchayishi 1.1. Tilni miqdoriy o'rganishga intilish. XIX asrning oxirida filologlar

Compiled on: December 4, 2025.

Copyright: ©2025 by the authors. Submitted to *Advances in Science and Humanities* for possible open access publication under the terms and conditions of the [Creative Commons Attribution \(CC BY\) 4.0 license](https://creativecommons.org/licenses/by/4.0/).

va matematiklar tomonidan til birliklarining chastotasini aniqlashga oid birinchi urinishlar paydo bo'ldi. Masalan:

- A.Xerdning ingliz matnlaridagi so'zlar chastotasi bo'yicha dastlabki kuzatuvlari,
- korpus bo'lmasa-da, katta matn to'plamlarini qo'lda statistik hisoblash tajribalari, keyinchalik kompyuterlashtirilgan chastota lug'atlarini yaratishga asos bo'ldi.

Bu davrning ahamiyati – tilni eksperimental yo'l bilan o'lchash mumkinligi haqidagi ilmiy fikrning shakllanishida.

1.2. Zipf qonuni. XX asr boshidagi eng muhim hodisalardan biri G. Zipfning so'zlar chastotasi bo'yicha tadqiqotlari bo'ldi. Zipf:

- til iqtisodiy qonuniyatlarga bo'ysunadi,
- yuqori chastotali so'zlar juda kam, past chastotali so'zlar esa nihoyatda ko'p,
- tilning matematik-modellashuvga mosligi mavjud, deb ta'kidadi.

Bu nafaqat korpus tilshunosligining statistik asosini yaratdi, balki keyingi kompyuter lingvistikasining ham poydevorini tayyorladi.

Strukturalizm va distribusional yondashuvlar

2.1. F.de Sossyur va til-nutq bo'linishi. Sossyur "til tizimi" (langue) va "nutq" (parole) o'rtasidagi farqni ajratdi. U bevosita korpus haqida gapirmagan bo'lsa-da, uning "nutq real faktlar majmuasidir" degan g'oyasi keyinchalik korpus lingvistikasiga metodologik zamin yaratdi.

2.2. Bloomfield va Amerika deskriptiv maktabi. L. Bloomfield boshchiligidagi Amerika strukturalistlari:

- tilni obyektiv, kuzatiladigan birliklar asosida tavsiflash,
- intuitiv mulohazalarni cheklash,
- kattik annotatsiya (tagging) amaliyotiga asos solish, kabi prinsiplarni ilgari surdi.

Distribusionalizm (Z. Xarris) esa "tilning tuzilishini uning birliklarining taqsimlanishi orqali bilish mumkin" degan g'oyani ilgari surdi – bu keyinchalik konkordans va kollokatsion tahlilning nazariy tagzaminiga aylandi. XX asr o'rtasi: texnologik burilish – kompyuterlarning paydo bo'lishi.

3.1. Elektron hisoblash qurilmalarining tilshunoslikdagi ilk qo'llanilishi. 1950–1960-yillarda kompyuterlar tilshunoslikka kirib kelishi

bilan birinchi marta:

- matnni mashinada kiritish,
- avtomatik statistik hisoblash,
- indekslash,
- matnlarni tez qayta ishlash imkoniyati paydo bo'ldi. Bu jarayon korpus lingvistikasi uchun texnik asos bo'lib xizmat qildi.

3.2. Machine Translation loyihalari (1950 – 1960-yillar). AQSh va Sovet Ittifoqidagi birinchi mashina tarjimasini loyihalari matnlarni elektron shaklda to'plash, kodlash va tahlil qilishga majbur edi. Aynan mana shu zaruriyat elektron matn korpuslarining paydo bo'lishiga sabab bo'ldi. Brown Corpus – zamonaviy korpus lingvistikasi boshlanishi (1964) Brown Corpus'ning fiziki yaratilishi korpus lingvistikasining ilmiy yo'nalish sifatida boshlanish nuqtasi hisoblanadi.

4.1. Brown korpusining yangiliklari. Bu korpus:

- 1 million so'zdan iborat ilk mashinada o'qiladigan korpus edi;
- turli janrlar bo'yicha tasniflangan;
- texnik jihatdan standartlashtirilgan matnlar bilan to'ldirilgan;
- chastota lug'atlarini yaratish imkonini berdi.

Uning metodologik ahamiyati shundan iboratki, u keyingi barcha korpuslar uchun janrlar balansi, meta-ma'lumotlar, teglash, korpus dizayni kabi me'yorlarni yaratdi.

4.2. Brown korpusining ta'siri. Brown korpusi:

- LOB (British counterpart),
- London-Lund og'zaki nutq korpusi,
- Kolmogorov statistika metodlarining lingvistik tadqiqotlarga kirib kelishi uchun asos bo'ldi.

Yevropa korpus maktabining shakllanishi (1970–1990-yillar)

5.1. London-Lund Corpus (1975). Tilshunoslik tarixida birinchi kattik annotatsiya qilingan fonetik-og'zaki nutq korpusi. Unda:

- intonatsiya,
- pauza,
- urg'u belgilarining teglanishi korpus lingvistikasini prosodiya va nutqshunoslik bilan bog'ladi.

5.2. COBUILD va J. Sinclair maktabi. Sinclair korpus lingvistikasiga tub metodologik burilish olib kirdi:

- "meaning is in the co-text" – semantika kontekst orqali aniqlanadi;

- “idiom principle” – til konstruksiyalarning tayyor birliklardan iborat;
- yangi lug‘atlar – empirik asosli, kontekstual ma’lumot bilan boyitilgan.

Bu tamoyillar korpus lingvistikasi nazariyasining shakllanishiga sabab bo‘ldi.

Milliy korpuslar davri (1990–yillar)

6.1. BNC (British National Corpus). 100 million so‘zdan iborat yirik milliy korpus bo‘lib:

- og‘zaki + yozma nutqni birlashtirdi,
- ijtimoiy–lingvistik balansni ta’minladi,
- zamonaviy annotatsiya standartlarini joriy qildi.

Shundan keyin dunyoning ko‘plab mamlakatlarida milliy korpuslar yaratila boshlandi:

- Germaniya (IDS Mannheim),
- Chexiya (Prague Dependency Treebank),
- Rossiya (Milliy Korpus),
- Xitoy, Yaponiya, Koreya.

XXI asr: internet, big data va NLP davri

7.1. Internet korpuslarining paydo bo‘lishi. Web-as-corpus konsepsiyasi tilni:

- real vaqtli,
- cheksiz miqdorda,
- dinamik o‘zgaruvchan ko‘lamda o‘rganish imkonini berdi.

7.2. AI va yirik til modellari. Sun’iy intellekt modellarining (GPT, BERT, DeepL) paydo bo‘lishi korpuslarni yangi ilmiy bosqichga olib chiqdi:

- milliardlab tokenlardan iborat mega-korpuslar,
- vektor semantika,
- o‘z-o‘zidan o‘rganadigan modellar.

Bu jarayon korpus lingvistikasining ta’sirini lingvistikaning barcha bo‘limlariga tarqatdi.

METODLAR

Korpus lingvistikasi murakkab, ko‘p qatlamli tarixiy, texnologik va nazariy jarayonlar bilan bog‘liq bo‘lgani sababli ushbu tadqiqotda bir necha ilmiy metodlar kompleks tarzda qo‘llanildi. Har bir metod korpus lingvistikasi shakllanishining ma’lum bir jihatini ochib berishga xizmat qildi.

Tarixiy-komparativ metod. Tarixiy-komparativ metod mazkur tadqiqotning asosiy metodlaridan biri bo‘lib, korpus lingvistikasining shakllanishi bosqichlarini ketma-ketlikda tahlil qilish, turli davrlar lingvistik yondashuvlarini qiyosiy asosda o‘rganish imkonini berdi. Bu metod

korpus lingvistikasi uzoq tarixiy jarayon natijasi ekanini, u birdan paydo bo‘lgan emas, balki tilshunoslikning turli ilmiy yo‘nalishlari, ijtimoiy ehtiyojlar va texnologik rivojlanish bilan bog‘liq evolyutsion hodisa ekanini aniqlashga yordam berdi.

Deskriptiv tahlil. Deskriptiv tahlil – korpus lingvistikasining amaliy tomonini o‘rganishga yo‘naltirilgan metod bo‘lib, korpuslarning tuzilishi, mazmuni, ijtimoiy–lingvistik mezonlari va funksional imkoniyatlarini tavsiflash imkonini berdi. Deskriptiv tahlil korpus lingvistikasi amaliyotining real holatini aniq ko‘rsatadi, korpuslarning texnik va strukturaviy xususiyatlari haqida to‘liq tasavvur beradi va ularning lingvistik tadqiqotlarda tutgan o‘rnini faktlarga asoslangan holda yoritish imkonini yaratadi.

Metodologik tahlil. Metodologik tahlil – korpus lingvistikasining nazariy asoslarini o‘rganuvchi metod bo‘lib, korpuslar bilan ishlashning ilmiy tamoyillari, teglash standartlari, statistik modellari va konseptual yondashuvlarini chuqur tahlil qilish imkonini beradi. Metodologik tahlil korpus lingvistikasining ilmiy apparatini aniqlashtiradi, uning tadqiqot shakli, ilmiy yo‘nalishi va lingvistik nazariyalar bilan bog‘liqligini tushuntiradi. Bu qism korpus lingvistikasining faqat amaliy emas, balki nazariy jihatdan ham mustahkam yo‘nalish ekanini isbotlaydi.

Zamonaviy texnologiyalarni baholash. Mazkur metod korpus lingvistikasi XXI asr ilmiy jarayonlari – sun’iy intellekt, katta ma’lumotlar (Big Data), neyron tarmoqlar, til modellarining rivoji – bilan qanday uzviy bog‘langanini aniqlashga qaratilgan. Zamonaviy texnologik yondashuvlar korpus lingvistikasini klassik metodlardan katta ma’lumotlar va sun’iy intellekt bilan integratsiyalangan yangi bosqichga olib chiqqanini ko‘rsatadi. Korpus endilikda nafaqat tadqiqot vositasi, balki AI tizimlarining asosiy o‘quv materiali hisoblanadi.

NATIJALAR

Ushbu tadqiqot natijalari korpus lingvistikasining tarixiy shakllanishi, metodologik rivoji, korpus turlarining ilmiy ahamiyati va zamonaviy sun’iy intellekt texnologiyalari bilan integratsiyasini kompleks asosda aniqlashga imkon berdi. Tahlil

jarayoni korpus lingvistikasining jamiyatdagi o'rnini, lingvistik tadqiqotlar uchun yaratgan yangi imkoniyatlari va XXI asr talablari bilan uyg'unlashuvini keng qamrovli tarzda yoritdi. Tadqiqot natijasida korpus lingvistikasining taraqqiyoti uch asosiy bosqichdan iborat ekanini ilmiy asosda tasdiqlandi. Bu bosqichlar nafaqat vaqt oralig'ini, balki metodologiya, texnologiya va ilmiy qarashlarning evolyutsiyasini aks ettiradi.

1) Dastlabki empirik-statistik davr (XIX–XX asr boshlarida)

2) Asos solish davri (1960–1990)

3) Global kengayish va metodologik takomillashuv davri (1990–hozir)

Tadqiqot korpus lingvistikasining shakllanishi bosqichma-bosqich kechgan murakkab tarixiy-nazariy jarayon ekanini, uning zamonaviy texnologiyalar bilan integratsiyasi esa fan rivojida yangi paradigmaga asos bo'layotganini isbotladi. Korpus lingvistikasi bugungi kunda tilshunoslikning eng ishonchli empirik poydevori, AI modellarining asosiy o'quv materiali va lingvistik tadqiqotlarning universal vositasiga aylangan.

XULOSA

Sun'iy intellekt va Big Data texnologiyalarining jadal rivojlanishi korpus lingvistikasiga yangi bosqich olib kirdi: gigant korpuslar, neyron til modellari, vektor semantika, avtomatik teglash, multimodal korpuslar kabi jarayonlar korpus lingvistikasi va AI integratsiyasini chuqurlashtirdi. Bugungi kunda korpuslar nafaqat tilshunoslikning empirik asosi, balki mashina tarjimasi, matnni qayta ishlash, nutq texnologiyalari va kognitiv modellashtirishning ham markaziy omiliga aylandi. Umuman olganda, korpus lingvistikasi hozirgi davr tilshunosligining nazariy, amaliy va texnologik yo'nalishlarini birlashtiruvchi strategik fan sifatida shakllanmoqda. Kelgusida real vaqtli, multimodal, o'zini avtomatik yangilovchi korpuslar va neyrolingvistik modellar bilan uyg'unlashgan korpuslar ushbu sohaning yangi avlodini yaratishi kutilmoqda.

References

1. Biber, D., Conrad, S., Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
2. Brown, K., Yule, G. (1983). *Discourse analysis*. Cambridge: Cambridge University Press.
3. McEnery, T., Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
4. Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.
5. Johansson, S., Hofland, K. (1994). *The International Corpus of English (ICE): A Handbook*. Oslo: Norwegian Computing Center for the Humanities.
6. Leech, G., Rayson, P., Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. London: Longman.