PEDAGOGIK TADQIQOTLAR JURNALI
ISSN: 3060-4923 Impact Factor –7,212

KORPUS LINGVISTIKIASINING ASOSIY TUSHUNCHALARI VA ULARNING ZAMONAVIY TILSHUNOSLIKDAGI O‘RNI

Alimova Mamuraxon Rahimberdiyevna
Andijon davlat chet tillar instituti
stajyor tadqiqotchisi

Annotatsiya

Ushbu maqolada zamonaviy tilshunoslikning eng jadal rivojlanayotgan yo‘nalishlaridan biri bolgan korpus lingvistikasining fundamental tushunchalari tahlil qilingan. Tadqiqotda korpus tushunchasi, uning turlari, muvozanatlashganlik (balancedness) mezonlari hamda lingvistik annotatsiya (belgilash) turlari batafsil yoritilishga harakat qilingan. Maqola IMRAD strukturasi asosida tuzilgan bo‘lib, korpus texnologiyalarining tilni o‘rganish va kompyuter leksikografiyasidagi amaliy ahamiyatini ko‘rsatib beradi.

Kalit so‘zlar: korpus lingvistikasi, korpus, annotatsiya, reprezentativlik, konkordans, token, lemma, kompyuter tilshunosligi.

Основные понятия корпусной лингвистики и их роль в современной лингвистике

Аннотация

В данной статье анализируются фундаментальные понятия корпусной лингвистики, одного из наиболее бурно развивающихся направлений современного языкознания. В исследовании подробно раскрыто понятие корпуса, его виды, критерии репрезентативности, сбалансированности, виды лингвистической аннотации. Статья построена на основе структуры IMRAD и демонстрирует практическую значимость корпусных технологий в изучении языка и компьютерной лексикографии

Basic concepts of corpus linguistics and their role in modern linguistics

Abstract

This article analyzes the fundamental concepts of corpus linguistics, one of the most rapidly developing areas of modern linguistics. The concept of corpus, its types, criteria of representativeness, balancedness, and types of linguistic annotation are detailed in the study. The article is structured on the basis of the IMRAD structure and demonstrates the practical importance of corpus technologies in language learning and computer lexicography

Kirish

Korpus lingvistikasi jahon kompyuter lingvistikasining juda tez rivojlanib kelayotgan sohalaridan biri bo‘lib, bu borada ancha yutuqlarga erishilgan. Dunyo oliy ta’lim muassasalarida

korpus lingvistikasi fan sifatida ham o‘qitiladi. Bu sohaning predmeti korpus yaratish nazariyasi va amaliyoti bo‘lsa, fan sifatida korpusning o‘ziga xosligi, dasturlash asoslari kabi jihatlari o‘qitiladi. Korpus lingvistikasi kompyuter lingvistikasining tarkibiy qismi, til korpusini yaratish, kompyuter texnologiyasi yordamida ulardan foydalanishning umumiy nazariyasi va amaliyoti bilan shug‘ullanadi.[1] Korpus lingvistikasining predmeti – til korpusi. U ingliz tilida linguistic corpus yoki text corpus kabi atamalar bilan qo‘llanilib, ilmiy adabiyotlarda turlicha ta’riflangan.[2]

Korpus – ma’lum maqsadda yig‘ilgan matnlar majmuasini tashkil etuvchi til birliklari yig‘indisi, tabiiy tildagi elektron shaklda saqlanadigan yozma yoki og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan online yoki offline tizimda ishlaydigan matnlar jamlanmasi.

Demak, til korpusi[3] ma’lum tilning belgilangan davrdagi, xilma-xil janr, rang-barang uslub, hududiy hamda ijtimoiy variantdagi matnlarining elektron shakldagi maxsus dasturiy ta’minot asosidagi yig‘indisidir. Korpus haqida yuqorida keltirilgan ta’riflar ham u haqida muayyan tilning o‘ziga xos xususiyati va variantlarini aks ettiruvchi bir necha belgi asosida tanlab olingan elektron shakldagi matn parchalari, lingvistik tadqiqot uchun asos vazifasini o‘taydigan tizim deya xulosa qilishga asos bo‘ladi

Lingvistik tadqiqot va amaliy topshiriqlar yechimi uchun til korpuslari zamonaviy tilshunoslikning inkor etib bo‘lmas ish quroli. U oddiy elektron kutubxonadan farqlanadi. Elektron kutubxona xalqning ijtimoiy-siyosiy, ma’naviy, iqtisodiy hayotini aks ettiruvchi badiiy va publisistik asarlarni nisbatan to‘liq qamrab olishni maqsad qilgan bo‘ladi. Elektron kutubxona matnlari tilshunoslik nuqtai nazaridan leksik, morfologik, grammatik, semantik belgilar asosida ishlov berilmaganligi sababli lingvistik tadqiqot uchun noqulaylik tug‘diradi. Chunki elektron kutubxona ilmiy tadqiqot materiali bazasini tayyorlash maqsadida tuzilmaydi, balki milliy-ma’naviy merosni to‘plashni maqsad qilgan bo‘ladi.

Korpusning xususiyatlari. Til korpusi hajm jihatdan imkon qadar katta bo‘lishi, 100 millionlab so‘z qo‘llanish holatini qamrab olishi lozim. Misol tariqasida rus tili milliy korpusi tarkibiga kirgan “Dostoyevskiy asarlari korpusi” uchun yig‘ilgan yozuvchining barcha asarlari matni 2 million so‘zdan iborat ekanligini keltirish o‘rinli. Shu sababli korpusni til modelining kichraytirilgan ko‘rinishi deb atash mumkin.[4]

Adabiyotlar sharhi

XXI asr boshlariga kelib, tilshunoslikda katta hajmdagi matnlarni elektron shaklda qayta ishlash va tahlil qilish ehtiyoji tug‘ildi. Bu ehtiyojdan kelib chiqib, korpus lingvistikasi mustaqil yo‘nalish sifatida shakllandi va taraqqiy eta boshladi.[5] Biz bilamizki, avvallari tilshunoslar ko‘proq intuitiv, ya’ni ichki hissiyotga asoslangan usullardan foydalanganlar,[6] endi vaqt o’tib, bugungi kunda empirik

[1] Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.
[2] Кутузов А.Б. Корпусная лингвистика. − (Электрон ресурс): Лицензия Creative commons Attribution Share-Alike 3.0 Unported (Электрон ресурс)
[3] Национальный корпус русского языка // Вопросы языкознания. – Москва: Наука, 2006. – С. 149-155.
[4] Захаров В.П. Корпусная лингвистика. Учебно-методическое пособие. – Санкт-Петербург, 2005. – 48 с. – С.
[5] McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
[6] Yo‘ldoshev, B. (2021). O‘zbek korpus lingvistikasi muammolari va istiqbollari. Toshkent: Fan.

(dalillarga asoslangan) yondashuv ustuvorlik qilmoqda. Mazkur tadqiqotning maqsadi korpus lingvistikasining tayanch tushunchalarini tizimlashtirish va ularning amaliy tilshunoslikdagi metodologik ahamiyatini ochib berishdan iborat.[7]

Ushbu tadqiqotni olib borishda tizimli tahlil, qiyosiy-tiplogik metod hamda deskriptiv (tavsifiy) usullardan foydalanildi. Korpus lingvistikasining shakllanish tarixi va bugungi holatini o‘rganish uchun xalqaro tajriba (BNC - Britaniya milliy korpusi, COCA - Zamonaviy Amerika ingliz tili korpusi)[8] va o‘zbek tili milliy korpusining (uzcorpora.uz) tuzilish prinsiplari solishtirma tahlil qilindi[9]. Tadqiqot doirasida korpus yaratishning quyidagi asosiy bosqichlari metodologik baza sifatida olindi:

- Matnlarni tanlash va raqamlashtirish (Data collection).
- Matnlarni tozalash va standart shaklga keltirish (Preprocessing).
- Lingvistik annotatsiya (Metadasturlash).

Jahon ilm-fanida korpus lingvistikasining nazariy poydevori aynan g‘arb tilshunos olimlarining asarlarida o‘z aksini topgan. Xususan, soha asoschilaridan biri Jon Sinkler (John Sinclair) o‘zining fundamental tadqiqotida korpus tahlilining eng asosiy tushunchalari - matnlarni muvozanatlash va so‘zlarning birikuvchanlik ehtimoli (kollokatsiya) qonuniyatlarini mukammal ochib bergan.[10]

Korpuslar yordamida nutq uslublarini (badiiy, so‘zlashuv, ilmiy) statistik tahlil qilish metodologiyasi esa Duglas Biber (Douglas Biber) nomi bilan bog‘liq Olimning tadqiqotlari an’anaviy grammatik tushunchalarni empirik ma’lumotlar asosida qayta shakllantirishga xizmat qildi.[11]

Zamonaviy korpus lingvistikasining darslik darajasiga ko‘tarilgan eng mashhur qo‘llanmasi mualliflari Toni MakEneri (Tony McEnery) va Endryu Hardi (Andrew Hardie) hisoblanadi. Ular korpus tushunchasining empirik tabiatini va annotatsiya (belgilash) turlarining metodologik ahamiyatini to‘laqonli tizimlashtirgan.[12] Sohaning amaliy samaradorligi, ayniqsa, korpus texnologiyalari va leksikografiya (lug‘atchilik) bog‘liqligini yangi bosqichga olib chiqqan Adam Kilgarriff ishlarida ko‘rinadi. Olim zamonaviy korpus dasturlarining asosi bo‘lgan "Sketch Engine" tizimini yaratdi va asarlarida so‘zlar korpusda qanday yashashi (Word Sketches) konsepsiyasini ilmiy asoslab berdi.[13]

Natijalar

Olingan natijalar shuni bilish mumkinki, korpus lingvistikasi an’anaviy tilshunoslikdagi ko‘plab bahsli masalalarni, aniq raqamlar va statistik ma’lumotlar yordamida hal qilish imkonini beradi.

Masalan, o‘zbek tilida "yurak" va "qalb" so‘zlarining qaysi biri ko‘proq publisistikada, hamda qaysi biri badiiy uslubda qo‘llanishini intuitiv aniqlash qiyin. Korpus dasturlari (masalan, Sketch

---

[7] Garside, R., Leech, G., & McEnery, T. (1997). *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Routledge.
[8] Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.
[9] Jurafsky, D., & Martin, J. H. (2020). *Speech and Language Processing* (3rd ed. draft). Stanford University.
[10] Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.
[11] Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge University Press.
[12] Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.
[13] McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.

Engine yoki NoSketch Engine) yordamida bir necha soniya ichida ularning chastotasini va qaysi so‘zlar bilan ko‘proq kollokatsiya hosil qilishini aniqlash mumkin.

Biroq, korpus lingvistikasida ham koplab muammolarga duch kelishimiz mumkin . Eng katta muammo -reprezentativlik muammosidir. Hech bir korpus, hatto u milliardlab so‘zdan iborat bo‘lsa ham, butun bir millat tilini 100 foiz qamrab ololmaydi. Shuning uchun korpus doimiy ravishda yangilanib va boyitilib borishi zarur. O‘zbek tili milliy korpusining rivojlanishi ham ayni shu bosqichda to‘g‘ri annotatsiya tizimini shakllantirishga muhtoj.

Xulosa

Korpus lingvistikasi shunchaki tilni o‘rganish usuli emas, balki zamonaviy tilshunoslikning yangi metodologik poydevoridir. Uning asosiy tushunchalari (korpus, token, lemma, annotatsiya, konkordans) tilni kompyuter tahlili orqali xolis o‘rganish imkonini beradi. Korpus texnologiyalarini rivojlantirish kelajakda mukammal elektron lug‘atlar yaratish, mashina tarjimasi (Machine Translation) sifatini oshirish va o‘zbek tilining sun’iy intellekt tizimlariga integratsiyalashuvini ta’minlashda hal qiluvchi rol o‘ynaydi.

Adabiyotlar ro‘yxati:

1. McEnery, T., & Hardie, A. (2011). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.
2. Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.
3. Yo‘ldoshev, B. (2021). O‘zbek korpus lingvistikasi muammolari va istiqbollari. Toshkent.
4. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.
5. Davies, M. (2009). The CoCA corpus: Design and applications. International Journal of Corpus Linguistics, 14(2), 159-190.
6. Yo‘ldoshev, B. (2021). O‘zbek korpus lingvistikasi muammolari va istiqbollari. Toshkent: Fan.
7. Garside, R., Leech, G., & McEnery, T. (1997). Corpus Annotation: Linguistic Information from Computer Text Corpora. Routledge.
8. Lüdeling, A., & Kytö, M. (Eds.). (2008). Corpus Linguistics: An International Handbook (Vol. 1). Walter de Gruyter.
9. Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.
10. Biber, D., Conrad, S., & Reppen, R. (1998). Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press.
11. Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). Longman Grammar of Spoken and Written English. Longman.
12. McEnery, T., & Hardie, A. (2011). Corpus Linguistics: Method, Theory and Practice. Cambridge University Press.
13. Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004). Itri: The Sketch Engine. In Proceedings of EURALEX (pp. 105-116).