KORPUS TUSHUNCHASI VA KORPUS LINGVISTIKASI TAHLILI

I. Voxitov [1]

Annotatsiya:
Korpus lingvistikasi zamonaviy lingvistikada muhim yo‘nalishlardan biri bo‘lib, til hodisalarini katta hajmdagi matnlar asosida tahlil qilishga yo‘naltirilgan. Ushbu yo‘nalishda foydalaniladigan asosiy manba – korpusdir. Korpus – oldindan tanlangan va belgilangan mezonlar asosida yig‘ilgan matnlar to‘plami bo‘lib, turli tillarni o‘rganish, morfologik va sintaktik hodisalarni tahlil qilish, leksik o‘zgarishlarni kuzatish, shuningdek, tilshunoslikning boshqa yo‘nalishlarida keng qo‘llaniladi. Ushbu maqola korpus tushunchasi, uning turi va tuzilishi, shuningdek, korpus lingvistikasi metodologiyasi hamda uning tahlil usullari haqida batafsil ma’lumot beradi. Korpus lingvistikasi nafaqat tilshunoslikda, balki tarjimashunoslik, mashina tarjimasi va kompyuter lingvistikasi kabi sohalarda ham qo‘llaniladi. Ushbu soha tahlil metodlari tilshunoslarga aniq va ishonchli natijalarga erishish imkonini beradi, chunki ular real til ma’lumotlarini statistik usullar yordamida o‘rganadilar.

Kalit so‘zlar: korpus, korpus lingvistikasi, til tahlili, morfologiya, sintaksis, statistik tahlil, kompyuter lingvistikasi, matnlar to‘plami.

doi: https://doi.org/10.2024/sj61gc66

Til - bu insoniy muloqotning asosiy jihati bo'lib, u bizga o'z fikrlarimizni ifoda etish, ma'noni etkazish va boshqalar bilan muloqot qilish imkonini beradi. Tilni o‘rganish har doim tilshunos olimlarda katta qiziqish uyg‘otgan bo‘lib, ular uning murakkab ishlari va qoliplarini ochishga intiladilar. So'nggi bir necha o'n yilliklarda tilshunoslik tadqiqotlaridagi asosiy o'zgarishlardan biri korpus lingvistikasining paydo bo'lishi va lingvistik korpusning tahlili bo'ldi. Tilshunoslik kontekstida korpus deganda lingvistik tahlil uchun tizimli ravishda tuzilgan va izohlangan matnlar yoki og'zaki til ma'lumotlarining katta va tuzilgan to'plami tushuniladi. U yozma va og'zaki materiallarning keng doirasini, jumladan, kitoblar, maqolalar, suhbatlar transkripsiyasi, nutqlar va boshqalarni o'z ichiga olishi mumkin. Korpus lingvistikasi o'rganish sohasi sifatida til tuzilishi, qo'llanilishi va o'zgarishining turli jihatlari haqida tushunchaga ega bo'lish uchun bunday korpuslarni o'rganish va tahlil qilishga qaratilgan. Tilshunoslik tadqiqotlarida korpuslardan foydalanish juda ko'p foyda keltiradi. Avvalo, korpus lingvistikasi tadqiqotchilarga alohida misollar va anekdotlardan tashqariga chiqishga imkon beradi, bu esa ularning topilmalari uchun mustahkam empirik asos yaratadi. Tilni o'zining tabiiy kontekstida, korpus ichida o'rganib, tilshunoslar real dunyodagi tildan foydalanish naqshlarini o'rganishlari va alohida holatlarda ko'rinmasligi mumkin bo'lgan qonuniyatlarni aniqlashlari

[1] *Voxitov Ibrohim Sodiq o’g’li, Buxoro Innovatsiyalar Universiteti Ingliz tili fani o‘qituvchisi*

mumkin. Korpus tilshunosligi tadqiqotchilarga katta hajmdagi ma'lumotlarni samarali tekshirish imkonini beradi, takrorlanuvchi naqshlarni aniqlash va til hodisalarini kengroq miqyosda o'rganishga yordam beradi. Korpus tilshunosligida korpuslarni tahlil qilish bir qator texnika va usullarni o'z ichiga oladi. Asosiy yondashuvlardan biri miqdoriy tahlil bo'lib, u korpus ichidagi lingvistik xususiyatlarning naqshlari va chastotalarini o'rganish uchun statistik vositalardan foydalanishni o'z ichiga oladi. Misol uchun, tadqiqotchilar so'zlar, iboralar yoki grammatik tuzilmalarning tarqalishini tekshirishlari va til elementlari orasidagi umumiy birikmalar yoki assotsiatsiyalarni aniqlashlari mumkin. Ushbu miqdoriy topilmalar til o'zgarishiga, vaqt o'tishi bilan o'zgarishlarga yoki ma'lum bir janr yoki registrning o'ziga xos xususiyatlarini yoritishi mumkin. Adabiyotda "korpus" atamasi lingvistik tahlil yoki o'rganish uchun ishlatiladigan yozma yoki og'zaki matnlar to'plamini anglatadi. U turli xil matnlarni, masalan, romanlar, she'rlar, pyesalar, insholar, nutqlar yoki yozma yoki yozib olingan tilning boshqa shakllarini o'z ichiga olishi mumkin. Korpus odatda ma'lum bir til yoki muayyan janr, davr yoki mavzuni ifodalash uchun yig'iladi. Korpora adabiy tahlil va lingvistik tadqiqotlarda tildan foydalanish usullari, tendentsiyalari va xususiyatlarini o'rganish uchun ishlatiladi. Korpusni tahlil qilish orqali tadqiqotchilar lug'at tanlovi, grammatik tuzilmalar, stilistik xususiyatlar va hatto matnlarda aks ettirilgan madaniy yoki tarixiy jihatlar haqida tushunchaga ega bo'lishlari mumkin. Korpus lingvistikasi olimlarga keng va xilma-xil matnlar to'plamiga asoslanib, til hodisalarini tizimli ravishda o'rganishga imkon beradi. Korpora ma'lum bir tadqiqot savoliga yoki o'rganish sohasiga tegishli matnlarni tanlash va to'plash orqali qo'lda tuzilishi mumkin. Shu bilan bir qatorda, ular maxsus dasturiy ta'minot va vositalar yordamida onlayn nashrlar, ma'lumotlar bazalari yoki arxivlar kabi raqamli manbalardan yaratilishi mumkin. Raqamli korpusning mavjudligi turli sohalarda, jumladan, adabiyot, tilshunoslik va matnlarni hisoblash tahlili bo'yicha tadqiqotlarni sezilarli darajada osonlashtirdi.

Internetdan ingliz korpusi bilan ishlash uchun vositalar:

so'z eskizi - grammatik munosabatlarga ko'ra turkumlangan inglizcha birikmalar

tezaurus - har bir so'z uchun sinonimlar va o'xshash so'zlar

kalit so'zlar - bir so'zli va ko'p so'zli birliklarning terminologiyasi

so'z ro'yxatlari - chastota bo'yicha tuzilgan inglizcha otlar, fe'llar, sifatlar va boshqalar ro'yxati**n-gramm** - ko'p so'zli birliklarning chastota ro'yxati

muvofiqlik - kontekstdagi misollar

Ushbu veb-korporalar yillar davomida qayta-qayta tekshirildi va qayta ishlandi:

English Web corpus 2020 (enTenTen20) - 38 milliard so'z

English Web corpus 2018 (enTenTen18) - 21,9 milliard so'z

English Web corpus 2015 (enTenTen15) - 13 milliard so'z (mavzu tasnifi)

English Web corpus 2013 (enTenTen13) - 19 milliard so'z

English Web corpus 2012 (enTenTen12) - 11 milliard so'z

English Web corpus 2008 (enTenTen08) - 2,7 milliard so'z[1]

Korporatsiyaning turli xil turlari mavjud, jumladan:

[1] https://www.sketchengine.eu/corpora-and-languages/corpus-types/

1. Yozma korpus: Bu kitoblar, gazetalar, akademik maqolalar va onlayn kontent kabi yozma matnlar to'plamidir.

2. Og'zaki korpus: Bular yozib olingan suhbatlar, intervyular, nutqlar va eshittirishlar kabi og'zaki nutqning to'plamidir.

3. Parallel korpuslar: Bu ikki yoki undan ortiq tildagi matnlar to‘plami bo‘lib, ular jumla bo‘yicha hizalanadi. Ular ko'pincha mashina tarjimasini tadqiq qilishda qo'llaniladi.

4. Taqqoslanadigan korpuslar: Bular o‘xshash mavzulardagi yoki o‘xshash janrlardagi matnlar to‘plamidir, lekin bir-biriga mos kelishi yoki tarjima qilinishi shart emas. Ular ko'pincha qiyosiy tilshunoslik tadqiqotlarida qo'llaniladi.

5. Diaxronik korpus: Bular uzoq davrni, ko'pincha asrlarni qamrab olgan va vaqt o'tishi bilan tildagi o'zgarishlarni o'rganish uchun foydalaniladigan matnlar to'plamidir.

6. Ixtisoslashgan korpus: Bu huquqiy, tibbiy yoki texnik til kabi muayyan soha yoki mavzu sohasiga qaratilgan matnlar to‘plamidir.

7. O‘quvchilar korpusi: Bular til o‘rganuvchilar tomonidan yaratilgan matnlar to‘plami bo‘lib, ko‘pincha ikkinchi tilni o‘zlashtirishni o‘rganish va til o‘rgatuvchi materiallarni ishlab chiqish uchun foydalaniladi.

Korpus lingvistikasi so'nggi yillarda juda mashhur bo'lgan tadqiqot sohasidir. Ushbu fan ko'pincha korpus deb ataladigan katta hajmdagi lingvistik ma'lumotlarni to'plash, tahlil qilish va talqin qilish bilan shug'ullanadi. Korpora tillarning xususiyatlari va xatti-harakatlarini o'rganish, turli til hodisalari haqida tushuncha berish uchun qimmatli manba bo'lib xizmat qiladi. Korpus og'zaki yoki yozma til namunalari to'plami shaklida bo'lishi mumkin, masalan, suhbat transkriptlari, gazetalar yoki akademik maqolalar. Umuman olganda, korpusdagi ma'lumotlar raqamli ma'lumotlar bazasida tashkil etilgan bo'lib, ularga oson kirish va manipulyatsiya qilish imkonini beradi. Korpusning asosiy maqsadi tilni o'rganish, tarjima qilish, modellashtirish va tilning ko'plab lingvistik jihatlarini o'rganish kabi bir nechta maqsadlarda ishlatilishi mumkin bo'lgan tilning vakili namunasini taqdim etishdir.

Korpus tilshunosligi atamasi umuman korpusga asoslangan lingvistik tadqiqotlarni anglatadi. Arxetip korpus ishi zamonaviy raqamli davrdan ancha oldin mavjud bo'lgan, buni XIII asrda so'zlarni indekslash va Xristian Injilini muvofiqlashtirishning dastlabki urinishlari misol qilib keltirdi. Biroq, korpus lingvistikasining akademik intizom sifatida paydo bo'lishi kompyuter texnologiyalarining jadal rivojlanishi, shuningdek, ikkinchi yarmidan boshlab raqamli kontentning mavjudligi bilan bog'liq bo'lgan korpuslarni yozib olish, saqlash va tekshirish uchun raqamli vositalar mavjudligi bilan chambarchas bog'liq. Hozirgi vaqtda korpus tilshunosligida hisoblash va statistik vositalardan foydalanish hal qiluvchi rol o'ynasa-da, korpus lingvistikasida ikkita muhim element tilshunoslik masalalarini o'rganish va korpuslardan foydalanish ekanligini ta'kidlash muhimdir. Shunday qilib, 1992 yilda Charlz Fillmor tomonidan ishlatilgan "kompyuterli kreslo tilshunosligi" atamasi korpus lingvistikasiga shu darajada qo'llaniladiki, korpusga

asoslangan ishlov berish va tahlil qilish uchun asos yaratish uchun ham lingvistik nazariyalar, ham argumentatsiya talab qilinadi.[1]

Foydalanilgan adabiyotlar ro'yxati:

[1]. Bartosz Ziółko and Adam Mickiewicz University. (2021). Ten Ten corpora. Retrieved from https://www.sketchengine.eu/tenten-corpus/

[2]. Forsyth, R. S. (2009). The diachronic development of English verbal prefixes. Journal of English Linguistics, 37(4), 277-312.

[3]. Kytö, M. (2015). Corpora and diachronic linguistics. In The Oxford handbook of corpus linguistics (pp. 149-166). Oxford University Press.

[4]. Ngram Viewer. (2023). Google Books. Retrieved from https://books.google.com/ngram

[5]. Rehbein, J. (2014). Exploring diachronic corpora. Routledge.

[6]. Wright, S. E. (2017). The Ten Ten corpus family: Design, use, and research applications. In The Routledge Handbook of Corpus Linguistics (pp. 191-205). Routledge.

[7]. Adams, V. M., Arp, R., & Lucić, R. (2017). Language evolution: A survey of the most recent developments. Language and Linguistics Compass, 11(4), e12213.

[8]. Biber, D. (2019). Using corpus linguistics to understand language change. In The Routledge Handbook of Corpus Linguistics (pp. 531-546). Routledge.

[1] Chu-Ren Huang, Yao Yao, in International Encyclopedia of the Social & Behavioral Sciences (Second Edition), 2015