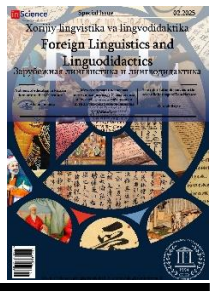
Xorijiy lingvistika va lingvodidaktika – Зарубежная лингвистика и лингводидактика – Foreign Linguistics and Linguodidactics
Journal home page:
https://inscience.uz/index.php/foreign-linguistics



Language corpora and their importance

Botir ELOV[1], Madina SAMATBOEVA[2]

Alisher Navo'i Tashkent State University of Uzbek Language and Literature

ARTICLE INFO

Article history:
Received December 2024
Received in revised form
10 January 2025
Accepted 25 January 2025
Available online
25 February 2025

Keywords:
Digitized world,
computational linguistics,
language corpora,
artificial intelligence,
sentiment analysis.

ABSTRACT

Language corpora are a crucial tool in the study of language, its processing, and the development of computational linguistics technologies. The process of their creation is based on clear principles and methods, and well-structured corpora hold significant importance for the advancement of linguistics and technology. In fields such as computational linguistics, artificial intelligence, and others, these corpora ensure the successful operation of modern technologies. This article explores language corpora, their creation, and their significance for language and society.

2181-3701/© 2024 in Science LLC.
DOI: https://doi.org/10.47689/2181-3701-vol3-iss2/S-pp203-209

Til korpuslari va ularning ahamiyati

Kalit so'zlar:
raqamlashgan dunyo,
kompyuter lingvistikasi,
til korpuslari,
sun'iy intellekt,
sentiment analiz.

ANNOTATSIYA

Til korpuslari tilni o'rganish, qayta ishlash va kompyuter lingvistikasi texnologiyalarini ishlab chiqishda juda muhim vosita hisoblanadi. Ularni yaratish jarayoni aniq tamoyillar va metodlarga asoslangan bo'lib, to'g'ri tuzilgan korpuslar tilshunoslik va texnologiyalarning rivojlanishida muhim ahamiyatga ega. Kompyuter lingvistikasi, sun'iy intellekt va boshqa sohalarda bu korpuslar zamonaviy texnologiyalarning muvaffaqiyatli ishlashini ta'minlaydi. Ushbu maqolada til korpuslari, ularning yaratilishi, til va jamiyatdagi ahamiyati to'g'risida fikr yuritiladi.

[1] PhD, Associate Professor, Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: elov@navoiy-uni.uz
[2] PhD student, Alisher Navo'i Tashkent State University of Uzbek Language and Literature.
E-mail: samatboyevamadina@navoiy-uni.uz

Языковые корпуса и их значение

Ключевые слова:
цифровой мир,
компьютерная
лингвистика,
языковые корпуса,
искусственный интеллект,
анализ тональности.

АННОТАЦИЯ

Языковые корпуса являются важным инструментом в изучении языка, его обработке и разработке технологий компьютерной лингвистики. Процесс их создания базируется на чётких принципах и методах, а правильно структурированные корпуса играют ключевую роль в развитии как лингвистики, так и современных технологий. В области компьютерной лингвистики, искусственного интеллекта и смежных дисциплин эти корпуса способствуют успешному функционированию инновационных систем. В данной статье рассматриваются языковые корпуса, методы их создания и значение для развития языка и общества.

KIRISH

Raqamlashgan dunyo va kompyuter lingvistikasi

Raqamlashgan dunyo – bu ma’lumotlar va bilimlarning raqamli formatda saqlanishi, ishlov berilishi va tarqatilishi jarayonlarini ifodalovchi atama. Bu dunyo texnologiyaning, ayniqsa axborot texnologiyalari va kompyuterlarning rivojlanishi bilan chuqur bog‘langan. Hozirgi kunda bizning kundalik hayotimizning ko‘plab qismlari raqamli formatda mavjud: internet, ijtimoiy tarmoqlar, onlayn ta’lim, raqamli moliya va boshqalar. Bularning barchasi raqamli dunyoning bir qismini tashkil etadi, ya’ni bu yerda ma’lumotlar tez va samarali tarzda almashiladi, saqlanadi va qayta ishlanadi.

Kompyuter lingvistikasi esa – kompyuter texnologiyalari yordamida tilni tushunish, qayta ishlash va ishlab chiqish bilan shug‘ullanuvchi ilmiy soha. Bu soha kompyuterlar va sun’iy intellekt tizimlarining inson tilini o‘rganish va ishlov berishga asoslangan. Kompyuter lingvistikasi yordamida matnni avtomatik tarjima qilish, nutqni tanib olish, so‘z ma’nosini aniqlash, matnlarni analiz qilish, mashinaviy tarjima va boshqa ko‘plab vazifalar bajariladi.

Kompyuter lingvistikasi va raqamli dunyo o‘rtasidagi bog‘liqlik juda katta, chunki raqamli ma’lumotlar ko‘plab tilga oid xususiyatlarni o‘z ichiga oladi, masalan, matnli ma’lumotlar, audio va video fayllar, ijtimoiy tarmoqdagi postlar va boshqa ko‘plab axborotlarni o‘z ichiga oladi. Ushbu ma’lumotlar tahlil qilish uchun kompyuter lingvistikasi texnologiyalari kerak bo‘ladi. Kompyuter lingvistikasi esa tildan foydalanishning yanada samarali va intuitiv usullarini yaratishga yordam beradi. Misol uchun, matnni avtomatik tahlil qilish yoki so‘zlarni tushunish imkoniyatlari kompyuterlar tomonidan amalga oshiriladi va bu raqamli dunyo bilan aloqador bo‘lgan har bir sohada muhim o‘rin tutadi.

ASOSIY QISM

Til korpusi nima?

Til korpusi (yoki korpus) – bu tilni o‘rganish, tahlil qilish va qayta ishlashda foydalaniladigan katta miqdordagi, tizimli ravishda yig‘ilgan matnlardan iborat to‘plamdir. Korpuslar, asosan, tilshunoslik, kompyuter lingvistikasi, mashinaviy tarjima, nutqni tanib olish va boshqa sohalarda ishlatiladi. Til korpuslari yordamida kompyuterlar inson tilini ancha samarali tushunish va ishlov berish imkoniyatiga ega bo‘ladi.

Til korpusi – bu biror tilni to‘g‘ri va kompleks tarzda o‘rganish uchun zarur bo‘lgan ma’lumotlarni o‘z ichiga olgan matnlar to‘plamidir. Bunday matnlar, odatda, turli xil yozuvlar va nutq namunalarini o‘z ichiga oladi, masalan, kitoblar, maqolalar, she’rlar, onlayn matnlar, ijtimoiy tarmoq postlari va hokazo. Korpuslar tilshunoslar va kompyuter lingvistikasi mutaxassislari tomonidan tilni tahlil qilish, o‘rganish, semantik va sintaktik strukturalarni aniqlash hamda tilni qayta ishlashning turli texnologiyalarini ishlab chiqish uchun foydalaniladi.

Internetda matnlarning keng tarqalganligi va erkin kirish imkoniyati mavjudligi sababli, eng katta korpus deb hisoblash mumkin bo‘lgan narsa – bu o‘zini o‘zi ifodalovchi *Internet (Web as Corpus)* hisoblanadi. Unga kirish vositalari sifatida, masalan, *Google* kabi qidiruv tizimlari xizmat qiladi. Korpus – bu ma’lumotlarni izlash tizimi bo‘lib, u ma’lum bir tilda elektron shaklda matnlarni yig‘ish asosida tashkil etilgan. Biroq, internetdagi matnlar tartibsizdir. Ba’zida lingvistik jihatdan so‘rovni qidiruv tizimining so‘rov tilida sifatida shakllantirish qiyin yoki imkonsizdir [1]. Shu sababli, lingvistik qoidalarga tayangan holda til korpuslari ishlab chiqilmoqda.

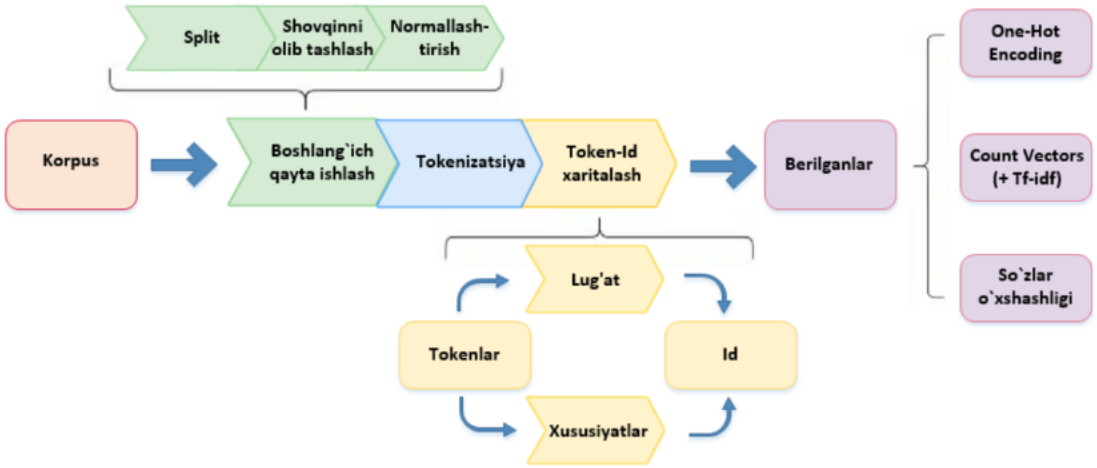
Birinchi lingvistik korpus 1960-yillarda Brayn universitetida (AQSh) yaratildi. U. Frensis va G. Kucher uni Amerika ingliz tilining amerikacha varianti bo‘yicha 500 dan 2000 so‘zgacha bo‘lgan prozatik bosma matnlar to‘plami sifatida loyihalashtirdilar. Matnlar amerikacha ingliz tilining eng ommabop o‘n beshta janriga tegishli edi. Mualliflardan biri, U. Frensis, “korpus” so‘zini “berilgan til, lahja yoki boshqa til qismiga tegishli va lingvistik tahlil uchun mo‘ljallangan matnlar to‘plami” ma’nosida ishlatgan. 1963 yilda Brayn lingvistik korpusining paydo bo‘lishi ommada keng qiziqish uyg‘otdi va jonli munozaralarga sabab bo‘ldi [2].

Kompyuter texnologiyalarining korpus lingvistikasi sohasida qo‘llanilishi elektron lug‘atlarni yaratish va rivojlantirish davrini belgilaydi. Bu esa olimlarga til korpusi asosida so‘zlarning ishlatilish chastotasini kuzatish, terminologiyaning rivojlanishini o‘rganish, jumla tuzilishini tahlil qilish, tarjima dasturlari uchun til bazalarini yaratish va metodologiyalarni o‘rganish imkonini beradi. Lug‘atlarni yaratish qobiliyati indeksdan avtomatlashtirilgan jarayonga o‘tdi, bu esa lug‘atlarni tuzishda ilgari ko‘rilmagan qulayliklarni yaratdi. Til korpuslarining, ayniqsa chastota lug‘atlarini yaratishning ahamiyati, ma’lum bir sohadagi terminologiyani shakllantirishda muhimdir. Bilingvistik yoki ko‘p tilli parallel korpuslarning tarjima ma’lumotlar bazasidan foydalanadigan tarjima lug‘atlari, avtomatik tarjima dasturlarining lingvistik ma’lumotlar bazalarini shakllantirishda asosiy manba bo‘lib xizmat qiladi.

Umuman olganda, ko‘p maqsadli va ko‘p funksiyali til korpusi, ma’lum bir tilning *Milliy Korpusi* hisoblanadi. Bunday korpusning asosida quyidagi talablarga javob berilishi kutiladi:

1. Tilning barcha uslublariga mansub matnlarning katta hajmi mavjudligi;
2. Matnlarning lingvistik, morfologik va sintaktik teglar bilan belgilanishi;
3. Ishlatilayotgan ma’lumotlar to‘plamining mavjudligi [3].

Ma’lum bir til doirasida “til korpusi”ni yaratish murakkab va mashaqqatli jarayon. Bunda juda ko‘p va turli shakl va mazmunga ega bo‘lgan matnlar jamlanmasi ma’lumotlar bazasi (data base) sifatida kerak bo‘ladi. Matnlar tozalanadi va qayta ishlanadi. Yaratilayotgan korpus turi va uning qo‘llanilish ahamiyatiga ko‘ra korpus hajmi va uning interfeysi ishlab chiqiladi. Qarang: 1-rasm.



1-rasm. Til korpusi matnlarini boshlang‘ich qayta ishlash bosqichlari [4]

Korpus asosida ish ko‘radigan eng birinchi soha – leksikografiya, u katta hajmli lug‘atlarni tuzish uchun asosiy, betakror manba hisoblanadi. Yaratilayotgan barcha zamonaviy, so‘nggi lug‘atlar korpusga asoslangan, ular misollarining haqiqiyligi, ishonarliligi bilan baholanadi. Chunki korpusda til jamiyatda qanday yashasa, shunday aks etadi, natijada, lug‘atdagi misol ishonarli, asosli bo‘ladi [4]

Til korpuslari ahamiyati

Til korpuslarining ahamiyati juda keng va ko‘plab sohalarda ulardan foydalanish imkoniyatlari mavjud. Korpuslar turli kontekstlarda turli xil muhim rollarni o‘ynaydi. Til korpuslarining ma’lum bir til doirasida qanchalik ahamiyatli ekanligi, korpuslar turlarining hosil bo‘lishiga sabab bo‘ladi. Til korpuslari jamiyatda quyidagi ahamiyatga ega:

Lingvistik ahamiyati

Lingvistik nuqtai nazardan, til korpuslari tilni o‘rganish va tahlil qilishda asosiy vosita hisoblanadi. Korpuslar tilning sintaktik, morfologik va semantik tuzilmalarini, jumladan, so‘zlarning o‘zaro munosabatlarini tahlil qilish imkonini beradi. Korpuslar lingvistik tahlil uchun haqiqiy matnlar asosida ishlaydi, bu esa nazariy tahlillarni tajriba bilan bog‘laydi. Korpuslar lingvistik tadqiqotlarda quyidagi jihatlarni o‘rganish uchun qo‘llaniladi:

- Til tuzilmalari va grammatikasining shakllanishi.
- So‘zlar va iboralar o‘rtasidagi semantik munosabatlar.
- Tilning evolyutsiyasi va yangi so‘zlar paydo bo‘lishi.

Texnik ahamiyati

Kompyuter lingvistikasi va sun’iy intellekt (AI) sohalarida til korpuslari texnik nuqtayi nazardan juda katta ahamiyatga ega. Ular tilni avtomatik qayta ishlash, mashinaviy o‘rganish va nutqni tanib olish tizimlarini yaratish uchun asos bo‘lib xizmat qiladi. Korpuslar yordamida quyidagi texnologiyalar rivojlanadi:

- **Avtomatik tarjima tizimlari:** Korpuslar, parallel matnlar asosida mashinaviy tarjimani yaxshilashda ishlatiladi. Shuningdek, korpuslar avtomatik tarjima tizimlarining asosini tashkil etadi. Parallel korpuslar (bir tildan boshqa tilga tarjima qilingan matnlar) yordamida mashinaviy tarjima tizimlari til o‘rganishini, so‘z va iboralarni to‘g‘ri tarjima

qilishni o‘rganadi. Korpuslar orqali tarjima tizimlarining aniqligi va samaradorligini oshirish mumkin, chunki ular yangi til shakllarini va jumlalarni o‘rganish imkoniyatini yaratadi. Korpuslar yordamida tilni qayta ishlash uchun algoritmlar va modellarning ishlash samaradorligini baholash mumkin.

- **Sentiment tahlil:** Ijtimoiy tarmoqlarda yoki boshqa matnlarda his-tuyg‘ularni aniqlash va tahlil qilish uchun korpuslar kerak. Masalan, katta hajmdagi korpuslar yordamida neural tarmoqlarni o‘rgatish va yaxshilash mumkin. Bu tizimlar, korpusdagi matnlardan o‘rganib, yangi matnlar ustida to‘g‘ri qarorlar qabul qilish imkonini beradi.
- **Nutqni tanib olish:** Korpuslar, nutqni to‘g‘ri tushunish va tanib olish tizimlarining samaradorligini oshirishda ishlatiladi. Nutqni tanib olish tizimlari korpuslarsiz ishlay olmaydi. Nutq korpuslari yordamida kompyuter tizimlari inson nutqini tushunish va unga javob berish qobiliyatiga ega bo‘ladi. Tilning fonetik, morfologik va sintaktik xususiyatlarini tahlil qilish uchun katta hajmdagi nutq korpuslari kerak. Shuningdek, tilning nozik xususiyatlari (ohang, urg‘u va his-hayajon)ni tushunish ham muhimdir, buni amalga oshirish uchun ham korpuslar ishlatiladi.

Psixologik ahamiyati

Til korpuslarining psixologik ahamiyati til va inson ongining o‘zaro bog‘liqligini o‘rganishda muhim rol o‘ynaydi. Korpuslar yordamida tilning psixologik aspektlari, odamlarning fikr yuritish jarayonlari, muloqotdagi hissiy munosabatlari va kognitiv jarayonlarni tushunish mumkin. Masalan:

- **Insonning tilni ishlatishi:** Odamlar turli holatlarda qanday so‘zlar yoki iboralardan foydalanishini tahlil qilish.
- **His-tuyg‘ularni ifodalash:** Qanday so‘zlar yoki ifodalar salbiy yoki ijobiy hissiyotlarni keltirib chiqaradi. Masalan, korpuslardan olingan ma’lumotlar yordamida, tilshunoslar yoki psixologlar qanday so‘zlar salbiy yoki ijobiy his-tuyg‘ularni ifodalash uchun ishlatilishini aniqlay olishadi. Bu esa, tilni ishlatishda his-tuyg‘ularning qanday ta’sirini o‘rganishda muhim bo‘ladi. Bunday tahlillar, shuningdek, hissiy tahlil (sentiment analysis) kabi texnologiyalarning rivojlanishiga olib keladi, bu esa biznes va kommunikatsiya sohalarida keng qo‘llaniladi. Bu orqali sentiment analizga asoslangan korpuslarni yaratish mumkin.
- **Kognitiv jarayonlar:** Til va fikrlash jarayonlari o‘rtasidagi bog‘liqlikni o‘rganish. Korpuslar yordamida, masalan, so‘zlarning ma’nosi, ularning tushunilishi va ularning ongda qanday saqlanishi haqida bilimlarni olish mumkin. Bu psixologik jihatdan, odamlarning fikrini shakllantirishdagi tilning rolini yaxshiroq tushunishga yordam beradi.

Ijtimoiy Ahamiyati

Til korpuslari ijtimoiy tilshunoslikda muhim ahamiyatga ega. Ular yordamida turli ijtimoiy guruhlar, madaniyatlar yoki tillar o‘rtasidagi farqlar va o‘xshashliklarni tahlil qilish mumkin. Korpuslar, shuningdek, ijtimoiy omillarning tilga ta’sirini o‘rganishda ishlatiladi. Masalan:

- **Lahjalarning tahlili:** Turli ijtimoiy guruhlar yoki hududlarga xos so‘zlashuv va iboralar. Bu usul orqali tilda frazeologik birliklar korpusini yaratish mumkin.
- **Ijtimoiy muloqotning shakllanishi:** Ijtimoiy, tarixiy asoslarga bog‘liq bo‘lgan til o‘zgarishlari.
- **Tillararo aloqalar:** Turli tillar va madaniyatlar o‘rtasidagi o‘zaro ta’sir va aloqalar.

Ta’limiy ahamiyati

Til korpuslari ta’lim sohasida ham katta ahamiyatga ega. Ular talabalarga tilni o‘rgatishda, xususan, lingvistik ko‘nikmalarni rivojlantirishda qo‘llaniladi. Korpuslar ta’lim tizimida quyidagicha ishlatiladi:

- **Til o‘rgatish:** Korpuslar yordamida grammatikani o‘rganish, so‘z boyligini oshirish, va nutqni yaxshilash mumkin.
- **Avtomatik sinovlar va baholash:** O‘quvchilarning bilimini baholash va tahlil qilish uchun korpuslardan foydalanish.
- **Tilni o‘rgatishda ma’lumotlar bazalari:** Korpuslar yordamida talaffuz va sintaktik xatolarni aniqlash. Imloviy xatolar ustida ishlash va natijani korpus orqali isbotlab berish.

Korpus foydalanuvchilari doirasi biz o‘ylaganimizdan ko‘ra ancha keng. Lug‘at, grammatikadan faqat tilshunos foydalanmaganidek, korpus gumanitar fanlar tadqiqotchisi, adabiyotshunos hamda tarixchi uchun ham birdek zaruriy baza bo‘la oladi. Til korpusiga eng ko‘p ehtiyoj sezuvchi mutaxassis matnni avtomatik qayta ishlash (masalan, tarjima dasturi), turli qidiruv tizimlari bilan ishlaydigan dasturchi. Chunki bu mutaxassis tabiiy til bilan ish ko‘radi hamda ushbu tilda yozilgan barcha matnlar strukturasini (tabiiy, jonli tilda, akademik grammatika, darslik yoki o‘quv qo‘llanmalardagi misollarga tayanib emas!) mukammalroq “tushunishi”, his etishi lozim. Til korpusiga ehtiyoj sezadiganlardan yana biri kundalik turmush tarzida yozuv bilan shug‘ullanuvchi mutaxassislar: gazeta, jurnal muharriri, jurnalist, radio, televideniye xodimi hisoblanadi. Chunki bunday xodim ma’lum bir so‘z, ibora yoki konstruksiyaning qo‘llanilish holati, darajasini, kim, qachon ilk marta shu konstruksiyani qo‘llaganligi, qanday uslub uchun xoslanishini bilishga grammatika bilan shug‘ullanuvchi olimdan ko‘ra ko‘proq ehtiyoj sezadi. Korpusdan tashqari yana qaysi axborot banki shunday savollarga tez, oson javob berishi mumkinligini tasavvur etib bo‘lmaydi. Shuning uchun korpus lingvistikasi tadqiqotchilari muayyan korpuslar jurnalist, muxbir, muharrir, o‘qituvchi hamda dasturchilar uchun maxsus yaratila boshlangan, degan xulosaga ham kelishadi [5].

XULOSA

Raqamlashgan dunyo va kompyuter lingvistikasi o‘rtasidagi integratsiya, ayniqsa sun’iy intellekt va mashinaviy o‘rganish kabi texnologiyalarning rivojlanishi bilan yanada kuchaymoqda. Bu sohalar o‘zaro birlashgan holda, ma’lumotlarni yanada samarali qayta ishlash, ilg‘or tizimlar yaratish va insonlar uchun yangi imkoniyatlar ochish imkonini beradi.

Til korpuslari bilimlarni chuqurroq o‘rganishga, texnologiyalarni rivojlantirishga va insonlarning tilni ishlatish jarayonlarini tushunishga yordam beradi. Shuningdek, til korpuslarining ahamiyati juda katta. Ular tilni tahlil qilish, mashinaviy o‘rganish, nutqni tanib olish, avtomatik tarjima tizimlarini yaratish va boshqa ko‘plab sohalarda muhim vositalar bo‘lib xizmat qiladi. Korpuslar yordamida tilni avtomatik qayta ishlash va tushunish tizimlari ancha samarali ishlaydi. Bu esa, o‘z navbatida, til texnologiyalarining rivojlanishini va keng qo‘llanilishini ta’minlaydi.

FOYDALANILGAN ADABIYOTLAR RO‘YXATI:

1. Сысоев П.В. Лингвистический корпус в методике обучения иностранным языкам // Язык и культура. 2010. № 1. С. 99–111.

2. Т.А.Чернякова. Использование лингвистического корпуса в обучении иностранному языку. Удк 373.167.1.

3. B.Elov, M. Abjalova, R.Alayev. O‘zbek tili korpusi va uning imkoniyatlari. Informatika va energetika muammolari. 2023. № 2. 89-100b.

4. Кутузов А.Б. Корпусная лингвистика. – (Электрон ресурс): Лицензия Creative commons Attribution ShareAlike 3.0 Unported (Электрон ресурс).

5. V.Zaxarov, B.Mengliyev, Sh.Hamroyeva. Korpus lingvistikasi: Korpus tuzish va undan foydalanish. Globe Edit. Toshkent-2021. 19 b.