ISSN 2181-922X

UZBEKISTAN LANGUAGE & CULTURE

O‘ZBEKISTON TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 4 (6)

www.compling.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON

TIL VA MADANIYAT

KOMPYUTER LINGVISTIKASI

2023 Vol. 4 (6)

compling.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

Bosh muharrir: **Botir Elov**

Bosh muharrir o‘rinbosari: **Shahlo Hamroyeva**

Mas’ul kotib: **Oqila Abdullayeva**

Tahrir kengashi

Jurnal haqida ma’lumot

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi – Oliy attestatsiya komissiyasi ilmiy nashrlar ro‘yxatidagi “O‘zbekiston: til va madaniyat” akademik jurnalining ilovasi hisoblanib, unda professor-o‘qituvchilar, doktorantlar, stajor-tadqiqotchilar, mustaqil izlanuvchilar, magistrantlarning kompyuter lingvistikasi, jumladan, tabiiy tilga ishlov berish (NLP), o‘zbek tilining formal grammatikasi, korpus lingvistikasi, mashina tarjimasi, nutqni qayta ishlash tizimlari, intellektual tizimlar, kompyuter leksikografiyasi hamda lingvistik ontologiyalar kabi sohalarga oid tadqiqotlari nashr qilinadi.

Jurnal ilovasi bir yilda to‘rt marta chop etiladi.

O‘zbek, turk, rus va ingliz tillarida yozilgan maqolalar qabul qilinadi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e‘lon qilinadi.

Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

“O‘zbekiston: til va madaniyat. Kompyuter lingvistikasi” seriyasi 2023-yildan chiqa boshlagan.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti. O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103-uy.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

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

Information about the magazine

"Uzbekistan: language and culture. "Computer Linguistics" series is an appendix of the academic journal "Uzbekistan: Language and Culture" in the list of scientific publications of the Higher Attestation Commission, in which computer linguistics, including natural language processing (NLP) of professors-teachers, doctoral students, intern-researchers, independent researchers, master's students, researches related to formal grammar of the Uzbek language, corpus linguistics, machine translation, speech processing systems, intelligent systems, computer lexicography and linguistic ontologies are published.

The magazine supplement is published four times a year.

Articles written in Uzbek, Turkish, Russian and English languages are accepted.

The journal also publishes book reviews, literature reviews, conference reports, and research project results.

The opinion of the authors may differ from the editorial point of view.

"Uzbekistan: language and culture. "Computer Linguistics" series has been published since 2023.

Tashkent State University of Uzbek Language and Literature named after Alisher Navoi. Yusuf Khos Hajib street, 103, Yakkasaray district, Tashkent, Uzbekistan.

E-mail: kompling@navoiy-uni.uz
Website: compling.tsuull.uz

MUNDARIJA

CONTENT

ISSN 2181-922X 6-18

TIL KORPUSLARIDA MATNLARNI ANNOTATSIYALASH: AFZALLIK VA KAMCHILIKLARI

Mastura Primova[1]

Annotatsiya. Ushbu maqolada korpus annotatsiyasi, annotatsiya turlari, afzalliklari, kamchiliklari ko‘rib chiqiladi. Korpus ma’lum maqsadda yig‘ilgan matnlar majmuyini tashkil etuvchi til birliklari yig‘indisi, tabiiy tildagi elektron shaklda saqlanadigan yozma va og‘zaki, kompyuterlashtirilgan qidiruv tizimiga dasturiy ta’minot asosida joylashtirilgan online yoki offline tizimda ishlaydigan matnlar jamlanmasi. Tilshunoslikka oid tadqiqotda fakt bilan ish ko‘riladigan hollarda material yig‘ilishi, sistemaga solinishi lozim. Bunday katta hajmli ishni bajarishda korpus vaqt va mehnatni tejaydigan ish quroli vazifasini bajaradi. U texnik jarayonni tezlashtiruvchi vosita bo‘libgina qolmay, muayyan tilning zamonaviy shakliga xos axborot tizimi ham bo‘lib, kutilmagan savolga javob bera oladigan, til hodisasi bilan shug‘ullanadigan soha oldiga avval ko‘rilmagan dolzarb muammolarni qo‘ya oladigan tizim.

Kalit so‘zlar: *Post-tegging, lemmatizatsiya, sintaktik tahlil, annotatsiya, coreferens annotatsiya, pragmatik annotatsiya, stilistik annotatsiya.*

Kirish

Zamonaviy axborot texnologiyalari tilning funksional imkoniyatlaridan foydalanish borasida benihoya imkoniyatlar yaratadi. Kompyuter tarjimasi, avtomatik tahrir va tahlil, yozma matnni ovozlashtiruvchi nutq sintezatorlari, og‘zaki nutqni yozma matnga aylantiruvchi nutqni tanish dasturlari, elektron lug‘atlar, lingvistik mobil ilovalar, tezaurus (til xazinasi)lar va til ontologiyasi. Ayniqsa, zamonaviy elektron lug‘atlar tuzish va undan foydalanish madaniyatini shakllantirish til imkoniyatini egallashda samarador ekanligi o‘z isbotini topgan. Xususan, tilning imkoniyatini namoyon qilish va egallash borasida dunyo miqyosida tez sur’atlarda yaratilayotgan til korpuslarining roli beqiyos. Til korpuslari - til bo‘yicha tadqiqot va amaliy topshiriqlar yechimi uchun zarur ish quroli. U oddiy elektron kutubxonadan farqlanadi. Elektron kutubxonaning maqsadi -

[1]*Primova Mastura Hakim qizi* – Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti Kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi o‘qituvchisi.
E-pochta: *primovamastura@navoiy-uni.uz*
ORCID: 0000-0002-0241-4659

xalqning ijtimoiy-siyosiy, ma'naviy, iqtisodiy hayotini aks ettiruvchi badiiy va publitsistik asarlarni nisbatan to'liq qamrab olish. Elektron kutubxona matnlari til nuqtayi nazaridan ishlov berilmaganligi sababli tadqiqotlar 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. Til korpusi esa elektron kutubxonadan farqli o'laroq, tilni o'rganish va tadqiq qilish uchun foydali va qiziqarli matnlarni to'plashni nazarda tutadi. Ko'p hollarda korpus annotatsiyalari korpusning belgilari bilan bog'liq hisoblanadi. Shuningdek, tilshunoslik tadqiqotlarda korpuslardan foydalanish orqali lingvistik ma'lumotlarni olishda ishlatiladi. Har doim ham korpusdan ma'lumotlarni ajratib olib bo'lmaydi. Bunday hollarda lingvistik tahlilni korpusga kodlash orqali amalga oshiriladi. "Elektron korpusga og'zaki va/yoki yozma til ma'lumotlarining izohlovchi lingvistik ma'lumotlarni qo'shish" jarayoni **korpus annotatsiyasi** deb ataladi [Leech, 1997. 2]. Korpus annotatsiyasi korpusga qo'shimcha qiymat beradi, ya'ni, korpus osonlikcha hal qila oladigan tadqiqot savollari doirasini sezilarli darajada kengaytiradi. Keng ma'noda ta'riflangan korpus annotatsiyasi ham matnli/kontekstual ma'lumotni, ham izohli lingvistik tahlilni kodlashni nazarda tutishi mumkin bo'lsa-da, adabiyotda tez-tez uchraydigan ikki tushunchaning o'zaro bog'liqligi ko'rsatiladi. Bu yerda atama tor ma'noda qo'llanilib, faqat lingvistik kodlashda ishlatiladi. Masalan, nutq qismini (POS) teglash va korpus matnida sintaktik tahlil qilishda.

Asosiy qism

Annotatsiya - Korpus lingvistikasi doirasida qaralganda berilgan matnga bevosita aloqasi bo'lmagan, ammo uning qaysidir qismi haqida lingvistik yoki ekstralingvistik axborot beruvchi umumiy ma'lumot. Annotatsiya o'z ichiga metama'lumot va teglarni qamrab olishi mumkin. Ayrim manbalarda [Leech, Vilson, 1994] korpus annotatsiyasi deganda, til korpusida matnning elektron shakliga kodlash qo'shilgan izohlovchi, lingvistik ma'lumotlarni qo'shish amaliyoti ham tushuniladi. McEnery annotatsiyalashni uch usulda amalga oshirish mumkinligini yozadi: to'la avtomatlashtirilgan, yarim avtomatlashtirilgan va qo'l mehnati yordamida. Ammo hech bir usul mukammal ish bermasligi, mutlaqo xatosiz natija olib bo'lmasligini aytgan.

Razmetka – McEnery razmetka va meta ma'lumotni annotatsiyalash jarayonining bir qismi sifatida baholagan [McEnery, Hardie,

2012. 47]. Boshqa manbalar bilan tanishganimizda esa razmetkaga annotatsiyalash jarayonining o‘zi sifatida baho berilganligiga guvoh bo‘ldik [Myfilology.ru]. Ingliz manbalarida razmetka markup termini bilan yuritiladi [McEnery, Hardie, 2012. 47]. Jahon amaliyotida razmetkalashning standart prinsiplari ishlab chiqilgan. U SGML yoki Standard Generalized Markup Language [https://www.techtarget.com/] deb ataladi. E’tibor qaratish lozim bo‘lgan jihat shundaki, SGMLda standart annotatsiya emas, balki annotatsiyalash jarayonini tashkil etish metodologiyasi aks etadi. SGML asosida ishlab chiqilgan va keng ommalashgan tillarga HTML yoki XMLni misol qilish mumkin.

Teg – Kompyuter yordamida matn tahlilini amalga oshirish jarayonini tezlashtirish va osonlashtirishga xizmat qiluvchi shartli belgi yoki maxsus kod. Teglar bir necha turlarga bo‘linadi [https://ucrel.lancs.ac.uk/]: semantik teg, sintaktik teg va grammatik teg. Grammatik teg, shuningdek, PoS (Part of speech) tegging nomi bilan ham mashhur. Annotatsiya (razmetka) va tegning farqini quyidagi jadvalda aniqroq ko‘rish mumkin [https://knowledge.autodesk.com/]:

1-jadval. Annotatsiya va teg terminining ta’riflanishi

Annotatsiya	Teg
Muayyan komponent yoki segment haqida jadval yoki grafik ko‘rinishidagi ma’lumot	Muayyan komponent yoki segment uchun unikal identifikator
Komponent yoki segment uchun bir nechta annotatsiya bo‘lishi mumkin	Komponent yoki segment uchun yagona nom beriladi va u teg deb ataladi
Bir qancha ta’riflar to‘plamini matn shaklida o‘zida jamlaydi va unda teglar ham aks etadi.	Gap emas, uning bo‘laklari haqida ma’lumot beruvchi muayyan formatdagi data hisoblanadi

Korpus belgisi korpusning tarkibiy qismlari va har bir matnning matn tuzilishi haqida nisbatan obyektiv tekshiriladigan ma’lumotlarni beradi. Shuningdek, korpus annotatsiyasi izohlovchi lingvistik ma’lumotlar bilan bog‘liq. “Annotatsiyani “*tarjimon*” deb atash orqali izohlash hech bo‘lmaganda ma’lum darajada inson ongining matnni tushunish mahsuli ekanligini bildiramiz” [Leech, 1997. 2]. Misol uchun, so‘zning nutq qismi noaniq bo‘lishi mumkin, buni korpus belgisidan ko‘ra korpus annotatsiyasi sifatida aniqlash osonroq hisoblanadi. Boshqa tomondan, ma’ruzachi yoki yozuvchining jinsi odatda obyektiv tekshiriladi va bu izohlash emas, balki belgilash bo‘ladi.

Korpus annotatsiyasi

Korpus belgisi kabi izoh korpusga qiymat qo‘shadi. Leech [1997. 2] korpus annotatsiyasiga shunday ta’rif beradi: “korpusni kelajakdagi tadqiqot va ishlanmalar uchun lingvistik ma’lumot manbayi sifatida boyitib, korpus keltiradigan foydaga hal qiluvchi hissadir”. McEnery [2003. 454-455] korpus annotatsiyasi kamida *to‘rtta afzalliklarga* ega ekanligini ko‘rsatadi.

• **Annotatsiya korpusdan ma’lumotlarni bir necha xil usullar orqali olish** ancha oson hisoblanadi. Leachning fikriga ko‘ra, nutqning bir qismini teglashsiz, xom korpusdan sifat sifatida chapni ajratib olish qiyin, chunki uning turli ma’nolari va qo‘llanishlarini faqat orfografik shakli yoki kontekstidan aniqlash mumkin emas. Masalan, o‘ngning teskari ma’nosiga ega chap orfografik shakli sifatdosh, qo‘shimcha yoki ot bo‘lishi mumkin. U o‘tgan zamon qo‘shimchasi yoki o‘tgan zamon shakli bo‘lishi ham mumkin. Nutqning tegishli qismiga izohlar bilan chapning bu turli xil qo‘llanilishini bir-biridan osongina ajratish mumkin. Shuningdek, korpus annotatsiyasi, inson tahlilchilari va mashinalariga o‘zlari qodir bo‘lmagan tahlillardan foydalanish va olish imkonini beradi [McEnery, 2003. 454].

Misol uchun, Xitoy tilini bilmasangiz ham, agar sizda to‘g‘ri izohlangan Xitoy korpusi bo‘lsa, ushbu korpusdan foydalangan holda Xitoy tili haqida ko‘p narsalarni bilib olishingiz mumkin. Korpusdan ma’lumotlarni olish tezligi - izohli korpusning yana bir afzalligi hisoblanadi. Agar biror kishi kerakli lingvistik tahlilni amalga oshirishga qodir bo‘lsa ham, agar biror kishi korpusning o‘ziga izoh berishdan boshlash kerak bo‘lsa, xom korpusni izohli korpusni o‘rganish kabi tez va ishonchli tarzda tekshira olishi dargumon.

• **Korpus annotatsiyasi qayta foydalanish mumkin bo‘lgan resursdir**, chunki annotatsiya korpus ichidagi lingvistik tahlillarni qayd etadi, keyinchalik ularni qayta ishlatish mumkin hisoblanadi. Korpus annotatsiyasi odatda qimmat va vaqt talab qilishini hisobga olsak, qayta foydalaniladi [Leech, 1997. 5].

• **Korpus annotatsiyasi ko‘p funksiyalilik uchun uni qayta foydalanish** mumkin hisoblanadi. Korpus dastlab ma’lum bir maqsad uchun izohlangan bo‘lishi mumkin. Biroq, korpus tahlili turli ilovalar uchun va hatto dastlab mo‘ljallanmagan maqsadlarda ham qayta ishlatilishi mumkin.

• **Korpus annotatsiyalari lingvistik tahlilni aniq qayd qilib boradi.** Unda tahlil va tanqidga ochiq bo‘lgan aniq obyektiv rekordni ta’minlaydi [McEnery, 2003].

Yana bir afzalliklaridan biri *korpus annotatsiyasi korpusning o‘zi kabi standart ma’lumot manbasini taqdim etadi*. U o‘zi ifodalovchi til xilma-xilligi uchun standart ma’lumotnoga asoslangan. Korpus annotatsiyasi asosan lingvistik tahlilni asosini ta’minlab, obyektiv ravishda qayd etadi. Shuning uchun ketma-ket tadqiqotlar umumiy asosda taqqoslanishi va qarama-qarshi qo‘yilishi mumkin bo‘ladi.

So‘nggi o‘n yil ichida korpus annotatsiyalariga to‘rtta asosiy kamchiliklari tanqid qilingan:

• **Korpus annotatsiyalari korpusda tartibsizlikni keltirib chiqaradi.** Ular Hunstonning fikriga ko‘ra “matnga juda ko‘p izoh qo‘shilgan bo‘lsa-da, tadqiqotchi annotatsiya belgilaridan tozalangan oddiy matnni ko‘ra olishi muhim” (2002. 94) deb ta’kidlaydi

• **Korpus annotatsiyalari korpus foydalanuvchisiga lingvistik tahlilni yuklaydi.** Garchi korpus annotatsiyalari o‘z mohiyatiga ko‘ra izohli bo‘lsa-da, korpus foydalanuvchilari ushbu tahlilni qabul qilishga majbur emas. Agar xohlasalar annotatsiyani e’tiborsiz qoldirib, o‘zlarining talqinlarini yuklashlari mumkin. Korpus annotatsiyasida matnni talqin qilishdan boshlanadi [McEnery, 2003. 456]. Shuningdek, korpusni izohsiz qoldirish korpus tahlil qilinganda talqin qilish jarayoni sodir bo‘lmaydi degani emas. Aksincha, izohning yo‘qligi tadqiqotchilar xom korpusdan foydalanganda bunday ko‘p talqinlar hali ham sodir bo‘lishini yashiradi. Tahlil hali ham sodir bo‘ladi, u shunchaki aniq ko‘rinishdan yashiringan hisoblanadi. Shuningdek, korpus annotatsiyasi bu borada zaiflik emas, balki afzallik sifatida tan olinadi, chunki u tekshirish uchun ochiq bo‘lgan aniq tahlilning obyektiv rekordini ta’minlaydi - izoh bermaslik shunchaki tahlil qilmaslik emas. Biroq, izohning yo‘qligi tahlilni qayta qurish qiyin yoki hatto imkonsiz bo‘lishini ta’minlaydi.

• **Annotatsiya korpusni “*ortiqcha baholab*” qo‘yishi mumkin, bu esa uni kamroq kirish, yangilash va kengaytirish imkonini beradi** [Hunston, 2002. 92-93]. Shuningdek, annotatsiya korpusni kamroq kirishni talab qilmaydi. Misol uchun, ko‘plab tahlil qilingan [masalan, Lancaster Parsed Corpus va Suzanne korpusi] va prosodik ravishda izohlangan korpuslar [masalan, London-Lund Corpus va Lancaster/IBM Spoken English Corpus] hamma uchun ochiqdir. Ba’zi korpus yaratuvchilari odatda o‘z korpuslarini annotatsiya qilish uchun juda ko‘p harakat qilishlariga qaramay o‘z korpuslarini iloji boricha kengroq foydalanishga topshirishdan mamnun bo‘lishadi. Ko‘pincha tashkilotlar korpus qurulishini moliyaviylashtirishadi chunki, qimmatli annotatsiyalar resursi ommaga taqdim etiladi. Annotatsiyalangan korpusni (yoki hatto xom korpusni) ommaga

taqdim etmaslikning keng tarqalgan sababi, korpus ma'lumotlariga tegishli mualliflik huquqi bilan bog'liq muammolar uni taqiqlaydi. Bu cheklov mualliflik huquqiga qo'yiladi, annotatsiyalarga emas.

Namuna korpusi ma'lum bir vaqtda ma'lum bir til xilma-xilligini ifodalash uchun mo'ljallangan hisoblanadi. Masalan, LOB va Brown korpuslari 1960-yillarning boshidan yozma Britaniya va Amerika ingliz tilini ifodalaydi deb taxmin qilinadi. Bu ikkita korpus - FLOB va Frown uchun "*yangilanishlar*" mavjud. Unda 1990-yillarning boshidan yozma Britaniya va Amerika ingliz tillarini ifodalaydi va sekinroq til o'zgarishlarini kuzatish uchun ishlatiladi. Doimiy kengayish zarurati faqat dinamik monitor korpus modeli bilan bog'liq hisoblanadi. Bu namuna korpusiga argument sifatida qo'llanilishi shart emas. Aksariyat korpuslar namunaviy korpus ekanligini hisobga olsak, kengaytirilish argumenti unchalik muhim emas, chunki namunaviy korpus hajmi odatda korpus ishlab chiqilganda aniqlanadi. Odatda, korpus yaratilgandan so'ng, kengaytirishga hojat bo'lmaydi.

• **Oxirgi tanqidda korpus annotatsiyasining aniqligi va izchilligi bilan bog'liq.** Korpusga annotatsiya berishning uchta asosiy usuli mavjud - *avtomatik*, *kompyuter* va *qo'lda*. Hunston fikriga ko'ra, "avtomatik izohlash dasturi inson tadqiqotchisi oladigan natijalarga 100% mos keladigan natijalarni berishi dargumon; boshqacha qilib aytganda, xatolar ehtimoli bor". Bunday xatolar matnlar faqat odamlar tomonidan matnlarni tahlil qilinganda ham sodir bo'ladi - hatto eng yaxshi tilshunos ham ba'zida xato qiladi. Shuning uchun annotatsiyalarga inson omillarini kiritish boshqa natijalarga olib kelishi mumkin; Sinkler (1992) ta'kidlashicha, "qo'lda yoki kompyuter yordamida korpus annotatsiyasiga inson omillarini kiritish annotatsiya izchilligini pasayishiga olib keladi". Bu ikki fikrni bir joyga jamlagan holda, nima uchun har qanday tilshunos tahlil qiladi, degan savol tug'ilishi mumkin. Chunki tahlillardagi nomuvofiqlik va noaniqlik haqiqatan ham kuzatilishi mumkin bo'lgan hodisalar bo'lsa-da, ularning ekspert inson tahliliga ta'siri bo'rttirilgan. Bundan tashqari, kompyuter noaniqliklar yoki nomuvofiqliklarning oldini olishning ishonchli vositasi emas: bu ikki nuqta mashina tahliliga ham tegishli bo'lishi mumkin. Avtomatlashtirilgan annotatsiyalar xatolari bo'ladi va bir-biriga mos kelmaydi. Agar annotatsiya dasturi uchun resurslar o'zgartirilsa - leksika o'zgartirilsa, qoidalar qayta yozilsa, vaqt o'tishi bilan dastur natijasi inson tahlilchilari tomonidan ko'rsatilganidan ancha yuqori darajada oshib ketishi mumkin bo'lgan shkalada nomuvofiqlikni ko'rsatadi. Korpus annotatsiyasida nimadan foydalanishimiz kerak: inson tahlilchilarimi yoki kompyu-

termi? Korpus annotatsiyasining ahamiyati keng e'tirof etilganligini hisobga olsak, inson tahlilchisi va mashina bir-birini to'ldirishi kerak, bu esa korpus hal qilish uchun mo'ljallangan tadqiqot savoli uchun noaniqlik va nomuvofiqlikni kamaytirishga qaratilgan aniqlik va izchillikka muvozanatli yondashuvni ta'minlashi kerak.

Yuqoridagi to'rtta tanqidlarni korpus annotatsiyasi rad etish mumkin. Unda annotatsiya faqat lingvistik tahlilni amalga oshirish va uni amalga oshirishni anglatadi.

Korpus annotatsiyasiga qanday erishiladi?

Korpus annotatsiyasi to'liq avtomatik ravishda ishlatiladi: ya'ni yarim avtomatlashtirilgan inson va mashina o'zaro ta'siri orqali yoki inson tahlilchilari tomondan butunlay qo'lda amalga oshiriladi. Uchalasini o'z navbatida qamrab olish uchun avtomatik izohlashda kompyuter dasturchi tomonidan oldindan belgilangan qoidalar va algoritmlarga rioya qilgan holda annotator sifatida yakka o'zi ishlaydi, ammo qoidalar oldindan belgilangan ML algoritmi yordamida mashinani o'rganish (ML) orqali ham mashina tomonidan tanlashi mumkin. Biroq, avtomatik izohlash vositasini ishlab chiqish vaqt va pul talab qilib, ma'lumotlar bazasida katta hajmdagi ma'lumotlarga tez va doimiy ravishda izohlanadi (resurs o'zgarmagan holda). Ba'zan, bu ish allaqachon boshqa joyda amalga oshirilganligini va kerakli izohni bajara oladigan dastur tekin mavjud ekanligini ko'rish mumkin.

Annotatsiyalarning ayrim turlari, masalan, Ingliz, fransuz va ispan tillari uchun lemmatizatsiya va POS belgilarini belgilash, xitoy tili uchun segmentatsiyasi va POS belgilarini ishonchli tarzda mashina yordamida amalga oshirilishi (odatda xatolik darajasi 3%), annotatsiyasiga to'liq avtomatlashtirilgan yondashuvdan iborat. Avtomatlashtirilgan jarayondan olingan ma'lumotlar tahlil qilinganligi kabi bo'lmasa yoki chiqish ishonchli bo'lsada, lekin ma'lum maqsad uchun yetarlicha aniq bo'lmasa (masalan, izohni yaxshilash uchun foydalaniladigan o'quv korpusi), odatda insoniy tuzatish talab qilinadi (ya'ni post-tahrirlash). **Post-tahrirlash** odatda qo'lda izoh berishdan ko'ra tezroqdir. Ba'zi izohlash vositalari inson-mashina interfeysini ta'minlaydi, bu inson tahlilchisiga mashina aniq bo'lmagan noaniq holatlarni hal qilishga yordam beradi. Yarim avtomatik izohlash jarayoni to'liq avtomatlashtirilgan izohga qaraganda ishonchliroq natijalar beradi, lekin u sekinroq ishlaydi va qimmat hisoblanadi. Sof qo'lda izohlash foydalanuvchi uchun hech qanday izoh vositasi mavjud bo'lmaganda yoki mavjud tizimlarning aniqligi qo'lda tuzatishga sarflangan vaqtni sof qo'lda izohlashdan kamroq qilish uchun yetarli

bo‘lmagan hollarda yuzaga keladi. Qo‘lda izohlash qimmat va ko‘p vaqt talab qiladiganligi sababli, odatda faqat kichik korpuslar uchun amal qiladi. Yuqorida aytib o‘tilgan bir nechta istisnolardan tashqari, hozirda katta korpuslarda mavjud bo‘lgan izohlarning aksariyat turlari yarim avtomatik yoki qo‘lda kiritilgan.

Korpus annotatsiyalarining turlari

Korpus annotatsiyasi turli darajalarda paydo bo‘lishi va turli shakllarga ega bo‘lishi mumkin. Masalan, fonologik darajada korpusga bo‘g‘in chegaralari (phonetic/phonemic annotation) yoki prozodik xususiyatlar (prosodic annotation) bilan izohlash mumkin; morfologik darajada korpusga prefikslar, qo‘shimchalar va o‘zaklar bilan izohlash mumkin morphological annotation); leksik darajada korpuslar nutq qismlari (POS tagging), lemmalar (lemmatizatsiya) (lemmatization), semantik maydonlar (semantic annotation) bilan izohlanishi mumkin; sintaktik darajada, korpuslarni tahlil qilish (parsing, treebanking or bracketing) yordamida izohlash mumkin; nutq darajasida korpusga anaforik munosabatlar (coreference annotation), nutq aktlari kabi pragmatik ma’lumotlar (pragmatic annotation) yoki nutq va fikrlarni ifodalash (stylistic annotation) kabi stilistik xususiyatlarni ko‘rsatish uchun izohlash mumkin.

Ulardan eng keng tarqalgan annotatsiya turi **POS tegi** bo‘lib, u ko‘plab tillarda muvaffaqiyatli qo‘llanilgan; **sintaktik tahlil** qilish ham jadal rivojlanmoqda, ayni paytda annotatsiyaning ayrim turlari (masalan, nutq va pragmatik annotatsiya) hozircha nisbatan rivojlanmagan.

POS teglari

POS tegi (grammatik teg yoki morfo-sintaktik izoh deb ham ataladi) korpusdagi har bir so‘zga POS tegi sifatida ham tanilgan nutqning bir qismi mnemonikasini belgilashni anglatadi. POS teglari korpus annotatsiyalarining birinchi keng tarqalgan turlaridan biri bo‘lib, bugungi kunda eng keng tarqalgan turi hisoblanadi. Shuningdek, tahlil va semantik izoh kabi keyingi tahlil shakllarining asosini tashkil etuvchi korpus annotatsiyasining eng asosiy turi hisoblanadi. Biroq, faqat nutq qismlari uchun izohlangan korpuslar keng ko‘lamli ilovalar uchun foydali bo‘lib, omograflarni ajratib ko‘rsatishdan tortib, korpusdagi so‘z sinflarining paydo bo‘lishini hisoblash kabi murakkabroq foydalanishgacha. Ko‘pgina lingvistik tahlillar, masalan, so‘z birikmasi ham POS teglariga tayanadi [Hunston, 2002. 81].

POS yorlig‘i rivojlanishining ilg‘or holatini hisobga olgan holda, u ko‘plab tillar uchun ko‘pgina tadqiqot savollari uchun yetarli aniqlik bilan avtomatik ravishda amalga oshirilishi mumkin.

Lug‘aviy elementlarga avtomatik ravishda POS teglarini tayinlaydigan izohlash vositasi tegger deb ataladi. Ingliz tili uchun eng mashhur va ishonchli teggerlardan biri Lankaster universitetida ishlab chiqilgan CLAWS (Clausibility Based Automatic Word Tagging System) hisoblanadi [Garside, Leech and Sampson, 1987]. Tizim gibrid statistik yondashuvdan foydalanadi, qoidaga asoslangan komponent bilan to‘ldiriladi, g‘ayrioddiy tarzda “idiomlar ro‘yxati” deb ataladi [Garside va Smith, 1997]. Bu tegger umumiy yozma ingliz tilida 97% aniqlikka erishgani xabar qilingan [Garside va Smith, 1997]. Tizim Britaniya Milliy Korpusini belgilash uchun ishlatilgan (BNC, 7.2-bo‘limga qarang). POS teggerlari fransuz [Gendner, 2002], ispan [Farwell, Helmreich, Kasper, 1995], nemis [Hinrichs, Kübler, Myuller va Uhle, 2002], shved [Kutting, 1994], Xitoy [Chang, Chen] kabi tillar uchun ham muvaffaqiyatli ishlab chiqilgan.

Xulosa

Tilni o‘rgatishda lug‘at boyligining ulkanligini ko‘rsata olish, so‘zning qo‘llanish imkoniyatini u yoki bu grammatik qurilma orqali tushuntirish uchun misollar massivini ko‘rsatishda korpus juda qo‘l keladi. Til birligini qidirish kerak bo‘lsa, bunday dasturiy ta’minot, ya’ni korpus tadqiqotchi yoki foydalanuvchiga juda katta yordam beradi. Ilgari tadqiqotchi o‘z ishi uchun misolni topish, ularni kartotekaga ko‘chirish (kompyuter texnologiyalari rivojlanishidan oldingi davrda)ga oylab, ba’zan yillab vaqt sarflagan bo‘lsa, bugun dunyo til korpuslari yordamida sanoqli daqiqada yuzlab misol topishga, tahliliy natijalarni olishga va ular ustida ishlash imkoniga ega bo‘ldi. Maxsus qidiruv tizimi korpusdan ma’lumot olishga mo‘ljallangan bir qancha dasturdan iborat, u statistik axborot va qidiruv natijasini foydalanuvchiga qulay shaklda taqdim eta oladi. Tilda qanday jarayon kechayotganligini aniq tasavvur qilish uchun korpus qamrovini yanada kengaytirish, nafaqat yozma nutq, balki og‘zaki nutq materialidan ham foydalanish maqsadga muvofiq. Bunday korpus yordamida taraqqiyot natijasida tilda sodir bo‘lgan va kutilayotgan o‘zgarish haqida aniq xulosa chiqarish mumkin. Tilshunoslikka oid tadqiqotda fakt bilan ish ko‘riladigan hollarda material yig‘ilishi, sistemaga solinishi lozim. Bunday katta hajmli ishni bajarishda korpus vaqt va mehnatni tejaydigan ish quroli vazifasini bajaradi. U texnik jarayonni tezlashtiruvchi vosita bo‘libgina qolmay, muayyan tilning zamonaviy shakliga xos axborot tizimi ham bo‘lib, kutilmagan savolga javob bera oladigan, til hodisasi bilan shug‘ullanadigan soha oldiga avval ko‘rilmagan dolzarb muammolarni qo‘ya oladigan tizim.

Korpuslar turli maqsadlarda turli sohalar uchun tuzilishi mumkin. Jumladan, og‘zaki yoki yozma matnlar, bir tilli yoki ko‘p tilli, uslubiga ko‘ra badiiy, publitsistik, folklor, bazasi o‘zgaruvchan yoki o‘zgarmas, matn hajmiga ko‘ra to‘liq matnli yoki fragmentli korpuslar bo‘ladi.

Til korpusi qurilishida lingvistik ta’minot masalasi muhim va murakkab hisoblanadi. Korpuslarda matnlardagi nutq bo‘laklariga mos identifakatorini belgilash jarayoni muammolidir, sababi tilni modellashtirish teglash qoidasi va tilda mavjud qonuniyat bilan bog‘liq. Teglash, xususan, grammatik teglash yoki POS tegging o‘zbek korpus lingvistikasi uchun ham dolzarb masaladir. Chunki maxsus “kodlangan” belgilar tizimi o‘zbek tili bilan bog‘liq NLP masalalarini yechishda birlamchi kalit bo‘lib xizmat qiladi.

Foydalanilgan adabiyotlar

Elov B., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlari uchun tf-idf statistik ko‘rsatkichni hisoblash. Science and innovation international scientific journal volume 1 issue 8 uif-2022: 8.2 ISSN: 2181-3337. 1774-1785 b.

O‘zbek tili ta’limiy korpusi - http://uzschoolcorpara.uz/

Elov B., Amirkulov M. Uzbek-English Parallel Corpus Algorithm and Alignment Problem. Central asian journal of literature, philosophy and culture eissn: 2660-6828 | Volume: 04 Issue: 06 June 2023. 71-78 p.

Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and Artificial Intelligence, Vol. 1 No. 3 (2023) 32–42. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v1i317. 117–129 b.

Elov B., Alayev R. O‘zbek tili korpusi va uning imkoniyatlari. O‘zbekiston Informatika va energetika mummolari jurnali. O‘zbekiston Jurnali. - Toshkent, 2023, - № 2. 89-100 b.

Elov B., Hamroyeva Sh., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlarini qayta ishlashda countvectorizer, tf-idf hamda co-occurrence matrix usullarining ahamiyati. ELEKTRON LUG‘ATLAR YARATISHNING NAZARIY VA AMALIY ASOSLARI mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari, 2023-yil 12-dekabr. Andijon. 78-88 b.

Захаров В.П. Корпусная лингвистика: Учебно-метод. Пособие. –СПб., 2005, -48с.

Boliang Zhang, Ajay Nagesh, Kevin Knight. 2020. Parallel Corpus Filtering via Pretrained Language Models. 10.18653/v1/2020.acl-main.756

Elaheh Rafatbakhsh, Alireza Ahmadi. 2019. A thematic corpus-based study of idioms in the Corpus of Contemporary American English. Asian-Pacific Journal of Second and Foreign Language Education,10.1186/s40862-0190076-4

Полицын, С.А., Полицына Е.В. 2018. Применение корпуса текстов для автоматической классификации в комплексе инструментов автоматизированного анализа текстов. Вестник ВГУ. Серия: Системный анализ и информационные технологии,10.17308/sait.2018.2/1224.

ADVANTAGES AND DISADVANTAGES OF CORPUS ANNOTATION

Mastura Primova[1]

Abstract. This article discusses corpus annotations, types of annotations, advantages and disadvantages. A corpus is a collection of linguistic units constituting a collection of texts collected for a specific purpose, written and spoken in natural language, stored in electronic form, a collection of texts hosted in a computerized software-based search engine and operated in an online or offline system. When working with facts in linguistic research, material should be collected and systematized. When performing such large-scale work, the case acts as a working tool, saving time and labor. It is not only a tool for speeding up a technical process, but also an information system specific to the modern form of a particular language, a system that can answer unexpected questions and create unprecedented problems in the field of linguistic phenomena.

Key words: *Post-tagging, lemmatization, parsing, annotation, coreference annotation, pragmatic annotation, stylistic annotation.*

References

Elov B., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlari uchun tf-idf statistik ko‘rsatkichni hisoblash. Science and innovation international scientific journal volume 1 issue 8 uif-2022: 8.2 ISSN: 2181-3337. 1774-1785 b.

O‘zbek tili ta’limiy korpusi - http://uzschoolcorpara.uz/

Elov B., Amirkulov M. Uzbek-English Parallel Corpus Algorithm and Alignment Problem. Central asian journal of literature, philosophy and culture eissn: 2660-6828 | Volume: 04 Issue: 06 June 2023. 71-78 p.

Elov B., Hamroyeva Sh., Alayev R., Xusainova Z., Yodgorov U. O‘zbek tili korpusi matnlarini qayta ishlash usullari. Digital transformation and Artificial Intelligence, Vol. 1 No. 3 (2023) 32–42. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v1i317. 117–129 b.

Elov B., Alayev R. O‘zbek tili korpusi va uning imkoniyatlari. O‘zbekiston Informatika va energetika mummolari jurnali. O‘zbekiston Jurnali. - Toshkent, 2023, - № 2. 89-100 b.

[1]*Primova Mastura Hakim qizi* – Alisher Navo’i Tashkent State University of Uzbek Language and Literature Teacher Department of Computational Linguistics and Digital Technologies.
E-mail: *primovamastura@navoiy-uni.uz*
ORCID: https://orcid.org/0000-0002-0241-4659

Elov B., Hamroyeva Sh., Xusainova Z., Xudayberganov N. O‘zbek tili korpusi matnlarini qayta ishlashda countvectorizer, tf-idf hamda co-occurrence matrix usullarining ahamiyati. ELEKTRON LUG‘ATLAR YARATISHNING NAZARIY VA AMALIY ASOSLARI mavzusidagi xalqaro ilmiy-amaliy anjuman materiallari, 2023-yil 12-dekabr. Andijon. 78-88 b.

Zaxarov V.P. Korpusnaya lingvistika: Uchebno-metod. Posobiye. – SPb., 2005. -48s.

Boliang Zhang, Ajay Nagesh, Kevin Knight. 2020. Parallel Corpus Filtering via Pretrained Language Models. 10.18653/v1/2020.acl-main.756

Elaheh Rafatbakhsh, Alireza Ahmadi. 2019. “A thematic corpus-based study of idioms in the Corpus of Contemporary American English”. Atsian-Pacific Journal of Second and Foreign Language Education,10.1186/s40862-0190076-4

Politsin, S.A., Politsina Ye.V. 2018. “Primeneniye korpusa tekstov dlya avtomaticheskoy klassifikatsii v komplekse instrumentov avtomatizirovannogo analiza tekstov”. Vestnik VGU. Seriya: Sistemniy analiz i informatsionnie texnologii,10.17308/sait.2018.2/1224

Jurnal 2017-yil 26-oktyabrda O‘zbekiston Respublikasi Matbuot
va axborot agentligi tomonidan 0936-raqam bilan ro‘yxatdan o‘tgan.

Jurnal O'zbekiston Respublikasi Oliy Attestatsiya Komissiyasi tomonidan
filologiya fanlari bo'yicha falsafa doktori (PhD) va fan doktori (DSc)
dissertatsiyalari asosiy ilmiy natijalari chop etilishi lozim bo'lgan ro'yxatga
kiritilgan (30.10.2021. № 308/6).

Tahririyatga kelgan maqolalar mualliflarga qaytarilmaydi.

Manzil: Toshkent shahri, Yakkasaroy tumani, Yusuf Xos
Hojib ko‘chasi 103-uy.
Telefonlar: +99871 281-45-11, +99871 281-41-93.
Website: compling.tsuull.uz
E-mail: kompling@navoiy-uni.uz

Bosishga 25.12.2023-yilda ruxsat etildi.
Bichimi 70x100 1/16, Ofset bosma. “Cambria” garniturasi.
Shartli b.t. 7,51. Nashr b.t. 7,62.

“O‘zbekiston: til va madaniyat” jurnali tahririyatida
tayyorlandi va sahifalandi.
“YASHNOBOD NASHR” bosmaxonasida chop etildi.
Adadi 300 nusxa. Buyurtma №2.
Bosmaxona manzili: Toshkent shahar Yashnobod tumani,
58-a harbiy shaharcha.