

O'ZBEKCHA MATNLARNI MORFOLOGIK TEGGLASH USULLARI

D.R.Maxammadiyeva ¹*Annotatsiya:*

Ushbu maqolada o'zbek tilining kelgusi taraqqiyotida muhim omillardan biri sifatida e'tirof etilgan til korpusi tushunchasi, korpus annotatsiyasi va morfologik razmetkasi haqida mulohazalar bayon etilgan.

Kalit so'zlar: korpus lingvistikasi, korpus annotatsiyasi, morfologik razmetka, tilni modellashtirish, morfologik teglashdagi muammolar.

doi: <https://doi.org/10.2024/9vemzr51>

Texnologiyalar shu qadar jadal rivojlanib, o'zgarib bormoqdaki, yashab turgan olamimizdagi barcha sohalar IT tillari orqali o'zining yangi taraqqiyot yo'lini boshladi. Jumladan, tilshunoslikda ham korpus lingvistikasi deb nomlangan soha dunyo tillarining yangi umr yo'lini boshlab berdi. Shu kungacha dunyodagi mavjud korpuslar turli jihatlariga ko'ra annotatsiyalangan. Til korpuslarida bugungi kunda annotatsiyaning morfologik, sintaktik, semantik, aktsentik kabi bir necha ko'rinishi mavjud. Masalan, Broun korpusi, hind korpusi so'z turkumlari jihatidan lingvistik annotatsiyalangan. Penn Treebanki gaplarni sintaktik tahlil qilish bo'yicha annotatsiyalangan, SemCor korpusida esa so'zlarning ma'nolariga ko'ra semantik farqlari [1] annotatsiyalanganini kuzatish mumkin. 1993 yilda Lankaster-Oslo/Bergen (LOB) va Britaniya milliy korpusi (BNC) muallifi J.Lich tomonidan tuzilgan annotatsiyalash usuli e'tiborga molik. Uning fikriga ko'ra, umumfoydalanishga mo'ljallangan korpusning razmetkasi uch prinsipga muvofiq kelishi kerak [2].

1. Razmetka (korpus annotatsiyasi) foydalanuvchi uchun qo'llanma yoki ko'rsatma shaklida mavjud bo'lgan tahlil sxemasiga asoslangan bo'lishi hamda har bir parametrlardan joy olishi kerak.

2. Foydalanuvchi uchun ochiq korpus razmetkasi "nazariy jihatdan neytral" bo'lishi lozim: razmetka parametrlari barcha uchun tushunarli bo'lgan tushunchalar tizimidan iborat bo'lishi talab etiladi. Agar korpus aniq bir loyiha uchun mo'ljallangan bo'lsa, uni razmetkalashda maxsus, aynan muallifga xos hamda umumqabul qilingan tasnifdan foydalanish lozim: bunda ham tuzuvchidan u yoki bu til nazariyasiga tayanish talab qilinadi.

3. Korpus annotatsiyasi sxemasi kim tomonidan, qaysi auditoriyaga mo'ljallanganligi aniq, ravshan ko'rsatilishi lozim, chunki korpusdan foydalanishda turli yuridik va texnik chegaralar mavjud [1].

Annotatsiya - bu matnga lingvistik va ekstralingvistik qo'shimcha ma'lumot bilan ishlov berilishidir. Masalan, matn va uning komponentlariga so'z turkumi haqida ma'lumot berilishi mumkin. Bu izoh matn tegi (belgi, ishora) yoki annotatsiya deb ataladi. Mukammal annotatsiya tadqiqotchiga qidirilayotgan so'z,

¹ D.R.Maxammadiyeva, O'zbekiston-Finlandiya instituti, tyutor

soʻzshakl va kontruktsiyani topishni osonlashtiradi. Chunki oddiy elektron matnda uning xarakterini ochuvchi alohida maʼlumot (annotatsiya) mavjud emas. Bizga shunchaki soʻz kerak boʻlsa, buni odatiy matn muharriri ham topib bera oladi. Lekin soʻzning grammatik maʼnosi, tuzilishini “tushunadigan” matn bilan ishlash juda koʻp afzallik va qulaylikka ega boʻladi. Faqat leksemani emas, balki soʻzshaklni qidirish kerak boʻlsa, bunda lingvistik korpus tadqiqotchi yoki foydalanuvchiga juda katta yordam berishi mumkin. Annotatsiyaning mukammal, toʻliqroq boʻlishi korpus ahamiyatini belgilovchi asosiy omildir.

Berilgan grammatik shaklni topish faqat oldindan ishlov berilgan matndagina bajarilishi mumkin, aks holda, qidiruv faqat qoʻlda bajariladi. Bu esa foydalanuvchidan koʻp mehnat hamda vaqtni talab qiladi. Mukammal ishlov berilgan, katta hajmli, turli janrdagi matnlarni qamrab olgan korpus tadqiqotchi uchun bebaho yordamchi hisoblanadi. Chunki til materiali oʻrganilgan har qanday tadqiqot materialining boyligi, ishonarliligi, aniq va mukammalligi bilan baholanadi. Agar tadqiqotchi oʻz tadqiqoti uchun misolni topish, ularni kartotekaga koʻchirish (kompyuter texnologiyalari rivojlanishidan oldingi davrdagi holat)ga oylab, baʼzan yillab vaqt sarflagan boʻlsa, bugun dunyo til korpuslari yordamida sanoqli daqiqada zarur deb hisoblangan yuzlab misolni topish, ular ustida ishlash imkoniyatiga ega boʻldi [3]. Usmon Azim dramalari uchun tuzilgan korpusda ham soʻzlar va soʻzshakllar qoʻl mehnati orqali morfologik annotatsiyalangan. Bu matnlarni lingvistik maʼlumotlar bilan boyitishning birinchi bosqichidir. Keyinchalik soʻzning turkumi haqidagi maʼlumotlardan tilni modellashtirish, matnlarni tasniflash, fikrlarni aniqlash, shaxs va joy nomlarini avtomatik aniqlash, korpus boʻyicha qidirish, konkordans tuzish, mashina tarjimasi, grammatik tahlil, soʻz maʼnosini ajratish, sintaktik tahlil va tokenlarga ajratish masalalarida foydalaniladi [4]. Matndagi har bir soʻzga mos soʻz turkumini qoʻlda teglab chiqish koʻp vaqt va mashaqqatli mehnatni talab qiladi. Koʻplab tillarda matnni soʻzshakllarga ajratish va avtomatik teglash amallarini bajaradigan dasturiy vositalar yaratilgan. Faqat dastur yordamida avtomatik teglangan matn lingvist mutaxassis tomonidan yana bir bor tekshiruvdan oʻtkazilishi lozim. Chunki mashina har doim ham inson tafakkur tarzi, hayotiy tajribalari orqali baholaydigan masalalarga toʻgʻri yondashavermaydi.

Morfologiya oʻzbek tilshunosligida atroflicha oʻrganilgan sohalardan biridir. U.Tursunov, J.Muxtorov, Sh.Rahmatullayev, R.Sayfullayeva, B. Mengliyev, Abuzalova, A.Hojiyev, S.Usmonov, E.Fozilov ishlarida oʻzbek tilining morfologik sathi oʻrganilgan, taʼriflangan, tavsiflangan. Yuqoridagi ilmiy adabiyotlar ham milliy korpusdan joy olsa, grammatika bilan shugʻullanuvchi lingvist maʼlumotlarni kam vaqt sarflab solishtirish, qiyosiy tahlil qilish va xulosalash imkoniyatiga ega boʻlar edi. N. Abduraxmonovaga koʻra korpus annotatsiyasi quyidagi maʼlumotlardan iborat boʻlishi mumkin [1]:

- Metamaʼlumot, sarlavha, paragraf;
- Tokenlar;
- Lemmalar;
- Soʻz turkumi tegi(part of speech-POS);
- Morfologik belgilar;
- Sintaktik tahlil;

Dramalar matnidagi soʻz turkumlarini teglash jarayonida baʼzi munozarali holatlarga duch keldik:

- bir soʻz turli kontekstda kelganda boshqa-boshqa turkumga tegishliligi yaʼni omonimlik holatlari;

- ibora, qoʻshma feʼl, koʻmakchi feʼlli soʻz qoʻshilmasi, soʻz birikmalarini teglashdagi holatlar va boshqalar. Matndagi soʻzni teglash jarayonida bu holatlarga duch kelinganda, teglashni bajarayotgan tilshunosning bilimli va tajribali boʻlishi muhim va bu matnli korpus sifatini belgilovchi omillardan biridir [4].

B.Mengliyev oʻz qarashlarida korpusning ahamiyatini shunday baholagan edi: “Bugungi kunda har qanday tilning yashab qolishi uchun bu tilga davlat tili maqomining berilganligi, bir necha million kishining bu tilda muloqat qilishi bu tilning yashab qolishiga kafolat bermaydi. Oʻzbek tili ayni vaqtda internet tili boʻla olmas ekan ona tilimizning tom maʼnoda axborot-kommunikatsiya texnologiyalari, internet tili darajasiga koʻtarilmayotgani ayniqsa, bu borada tegishli ilmiy-tadqiqot ishlarining olib borilmayotganligi uning ham yoʻqolish xavfidan butkul xoli emasligini, dunyo tillari roʻyxatida esa yoʻqolishdan ehtiyot qilish lozim boʻlgan tillar qatorida qolib ketayotganligini koʻrsatadi. Afsuski, tilimizning ichki imkoniyatlari axborot texnologiyalari tizimida etarli darajada ishga solinmayapti. Yurtimizda oʻzbek tili korpuslarini yaratish ishlarini qoniqarli deb boʻlmaydi. Bu ona tilimizning axborot texnologiyalari tili sifatida, shuningdek, milliy korpus shaklida global korpuslar toʻriga ulanib, jahonga chiqishiga monelik qilyapti. Bu sohada olib borilayotgan ilmiy tadqiqot ishlari oʻzbek korpus lingvistikasining nazariy asoslarini ishlab chiqish, namuna sifatida bir-ikki korpus lavhasi yaratish bosqichida turibdi, xolos. Oʻzbekistonda amaliy tilshunoslikning bir yoʻnalishi sanalgan oʻzbek korpus lingvistikasini rivojlantirish, uni ilmiy-nazariy jihatdan tadqiq etish, kompyuter va korpus lingvistikasi markazi yoki laboratoriyasini tashkil qilish hamda unga mutaxassislarni birlashtirish orqali amalga oshishi mumkin”[5].

Xulosa oʻrnida aytish mumkinki, bugun amaliy tilshunosligimiz oldida turgan eng muhim vazifalardan biri – bu turli janrlarni qamrab olgan oʻzbek tilining milliy korpusini yaratishdir. Bunda til birliklarini modellashtirish jarayonida umumiy standart namunani belgilab olish zarur. Korpus foydalanuvchisi milliy korpusdagi matnlar uchun tushunarli, ravon, bir xil annotatsiyadan fodalansa, maqsadga muvofiqdir.

Foydalanilgan adabiyotlar roʻyxati:

[1]. N.Abduraxmonova. *oʻzbek tili elektron korpusining kompyuter modellari. Monografiya. Toshkent – 2021.*

[2]. Хидиров О. *Миллий корпус учун парсинг дастури яратишнинг лингвистик асослари. PhD dissertatsiya avtoreferati. 2021.*

[3]. Хамроева Ш. *Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. filol. fan. boʻy. PhD diss. – Qarshi. 2018.*

[4]. I.Rabbimov, S.Umirova, B.Xolmuxamedov. *Oʻzbek tili korpusida soʻz turkumlarini teglash masalasi. “Oʻzbek milliy va taʼlimiy korpuslarini yaratishning nazariy hamda amaliy masalalari”. 97-bet*

[5]. Mengliyev B. *“Globallashuv: tillar taraqqiyoti va tanazzuli”. “Maʼrifat” gazetasi, 2017-yil, 19-oktabr, 76-son.*