

O‘ZBEKCHA MATNLARNI MORFOLOGIK TEGGLASH USULLARI МЕТОДЫ МОРФОЛОГИЧЕСКОЙ ТЕЙГИРОВАНИЯ УЗБЕКИНСКИХ ТЕКСТОВ

METHODS OF MORPHOLOGICAL TAGING OF UZBEKIN TEXTS

D.R.Maxammadiyeva¹

1.- O‘zbekiston-Finlandiya instituti, Samarqand, O‘zbekiston.

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.

Абстрактный. В данной статье описывается понятие языкового корпуса, корпуса аннотации и морфологической классификации, признанное одним из важных факторов дальнейшего развития узбекского языка.

Ключевые слова. Корпусная лингвистика, корпусная аннотация, морфологическая разметка, моделирование языка, проблемы морфологической разметки.

Abstract. This article describes the concept of language corpus, corpus annotation and morphological classification, recognized as one of the important factors in the future development of the Uzbek language.

Keywords. Corpus linguistics, corpus annotation, morphological tagging, language modeling, problems in morphological tagging.

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 parametr undan 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

D.R.Maxammadiyeva,
O‘zbekiston-
Finlandiya instituti,
Tyutor
diyoramaxammadiyeva@gmail.com

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, so'zshakl va konstruktsiyani topishni osonlashtiradi. Chunki oddiy elektron matnda uning xarakterini ochuvchi alohida ma'lumot (annotatsiya) mavjud emas. Bizga shunchaki so'z kerak bo'lsa, buni odatiy matn muharriri ham topib bera oladi. Lekin so'zning grammatik ma'nosi, tuzilishini "tushunadigan" matn bilan ishlash juda ko'p afzallik va qulaylikka ega bo'ladi. Faqat leksemani emas, balki so'zshaklni qidirish kerak bo'lsa, bunda lingvistik korpus tadqiqotchi yoki foydalanuvchiga juda katta yordam berishi mumkin. Annotatsiyaning mukammal, to'liqroq bo'lishi korpus ahamiyatini belgilovchi asosiy omildir.

Berilgan grammatik shaklni topish faqat oldindan ishlov berilgan matndagina bajarilishi mumkin, aks holda, qidiruv faqat qo'lda bajariladi. Bu esa foydalanuvchidan ko'p mehnat hamda vaqtni talab qiladi. Mukammal ishlov berilgan, katta hajmli, turli janrdagi matnlarni qamrab olgan korpus tadqiqotchi uchun bebaho yordamchi hisoblanadi. Chunki til materiallari o'rganilgan har qanday tadqiqot materialining boyligi, ishonarliligi, aniq va mukammalligi bilan baholanadi. Agar tadqiqotchi o'z tadqiqoti uchun misolni topish, ularni kartotekaga ko'chirish (kompyuter texnologiyalari rivojlanishidan oldingi davrdagi holat)ga oylab, ba'zan yillab vaqt sarflagan bo'lsa, bugun dunyo til korpuslari yordamida sanoqli daqiqada zarur deb hisoblangan yuzlab misolni topish, ular ustida ishlash imkoniyatiga ega bo'ldi[3]. Usmon Azim dramalari uchun tuzilgan korpusda ham so'zlar va so'zshakllar qo'l mehnati orqali morfologik annotatsiyalangan. Bu matnlarni lingvistik ma'lumotlar bilan boyitishning birinchi bosqichidir. Keyinchalik so'zning turkumi haqidagi ma'lumotlardan tilni modellashtirish, matnlarni tasniflash, fikrlarni aniqlash, shaxs va joy nomlarini avtomatik aniqlash, korpus bo'yicha qidirish, konkordans tuzish, mashina tarjimasi, grammatik tahlil, so'z ma'nosini ajratish, sintaktik tahlil va tokenlarga ajratish masalalarida foydalaniladi[4]. Matndagi har bir so'zga mos so'z turkumini qo'lda teglab chiqish ko'p vaqt va mashaqqatli mehnatni talab qiladi. Ko'plab tillarda matnni so'zshakllarga ajratish va avtomatik teglash amallarini bajaradigan dasturiy vositalar yaratilgan. Faqat dastur yordamida avtomatik teglangan matn lingvist mutaxassis tomonidan yana bir bor tekshiruvdan o'tkazilishi lozim. Chunki mashina har doim ham inson tafakkur tarzi, hayotiy tajribalari orqali baholaydigan masalalarga to'g'ri yondashavermaydi.

Morfologiya o'zbek tilshunosligida atroflicha o'rganilgan sohalardan biridir. U.Tursunov, J.Muxtorov, Sh.Rahmatullayev, R.Sayfullayeva, B. Mengliyev, Abuzalova, A.Hojiyev, S.Usmonov, E.Fozilov ishlarida o'zbek tilining morfologik sathi o'rganilgan, ta'riflangan, tavsiflangan. Yuqoridagi ilmiy adabiyotlar ham milliy korpusdan joy olsa, grammatika bilan shug'ullanuvchi lingvist ma'lumotlarni kam vaqt sarflab solishtirish, qiyosiy tahlil qilish va xulosalash imkoniyatiga ega bo'lar edi. N. Abduraxmonovaga ko'ra korpus annotatsiyasi quyidagi ma'lumotlardan iborat bo'lishi mumkin [1]:

- Metama'lumot, sarlavha, paragraf;
- Tokenlar;
- Lemmalar;
- So'z turkumi tegi(part of speech-POS);
- Morfologik belgilar;
- Sintaktik tahlil;

Dramalar matnidagi so'z turkumlarini teglash jarayonida ba'zi munozarali holatlarga duch keldik:

- bir so'z turli kontekstda kelganda boshqa-boshqa turkumga tegishliligi ya'ni omonimlik holatlari;
- ibora, qo'shma fe'l, ko'makchi fe'lli so'z qo'shilmasi, so'z birikmalarini teglashdagi holatlar va boshqalar. Matndagi so'zni teglash jarayonida bu holatlarga duch kelinganda, teglashni bajarayotgan tilshunosning bilimli va tajribali bo'lishi muhim va bu matnli korpus sifatini belgilovchi omillardan biridir [4].

B.Mengliyev o'z 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. O'zbek tili ayni vaqtda internet tili bo'la olmas ekan ona tilimizning tom ma'noda axborot-kommunikatsiya texnologiyalari, internet tili darajasiga ko'tarilmayotgani ayniqsa, bu borada tegishli ilmiy-tadqiqot ishlarining olib borilmayotganligi uning ham yo'qolish xavfidan butkul xoli emasligini, dunyo tillari ro'yxatida esa yo'qolishdan ehtiyot qilish lozim bo'lgan tillar qatorida qolib ketayotganligini ko'rsatadi. Afsuski, tilimizning ichki imkoniyatlari axborot texnologiyalari tizimida yetarli darajada ishga solinmayapti. Yurtimizda o'zbek tili korpuslarini yaratish ishlarini qoniqarli deb bo'lmaydi. Bu ona tilimizning axborot texnologiyalari tili sifatida, shuningdek, milliy korpus shaklida global korpuslar to'riga ulanib, jahonga chiqishiga monelik qilyapti. Bu sohada olib borilayotgan ilmiy tadqiqot ishlari o'zbek korpus lingvistikasining nazariy asoslarini ishlab chiqish, namuna sifatida bir-ikki korpus lavhasi yaratish bosqichida turibdi, xolos. O'zbekistonda amaliy tilshunoslikning bir yo'nalishi sanalgan o'zbek 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 o'rnida aytish mumkinki, bugun amaliy tilshunosligimiz oldida turgan eng muhim vazifalardan biri – bu turli janrlarni qamrab olgan o'zbek 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 ro'yxati

1. Abduraxmonova N. O'zbek 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.bo'y. PhD diss. – Qarshi. 2018.
4. Rabbimov I., Umirova S., Xolmuxamedov B. O'zbek tili korpusida so'z turkumlarini teglash masalasi. //“O'zbek milliy va ta'limiy korpuslarini yaratishning nazariy hamda amaliy masalalari”. 97-bet.
5. Mengliyev B. “Globallashuv: tillar taraqqiyoti va tanazzuli”. //“Ma'rifat” gazetasi, 2017-yil, 19-oktabr, 76-son.