TIL KORPUSI BIRLIKLARINI TEGLASH AHAMIYATI

Taxirova Marjona Olim qizi,
Buxoro davlat universiteti “Kompyuter lingvistikasi” magistranti
E-mail: marjonaxontaxirova@gmail.com

Annotatsiya. Bugungi kunda jahon va o‘zbek tilshunosligida zamonaviy yo‘nalish sifatida korpus lingvistikasi taraqqiy etib bormoqda. Fanning mahsuli o‘laroq bir qator til korpuslari yaratilmoqda. Ushbu maqolada mana shunday elektron til korpuslarini teglash masalasining dolzarbligi va amaliy ahamiyati haqida so‘z boradi.

Abstract. Today, corpus linguistics is developing as a modern direction in world and Uzbek linguistics. As a product of science, a number of language corpora are being created. This article discusses the relevance and practical significance of tagging such electronic language corpora.

Аннотация. Сегодня корпусная лингвистика развивается как современное направление в мировом и узбекском языкознании. В качестве продукта науки создаётся ряд языковых корпусов. В данной статье обсуждается актуальность и практическая значимость тегирования таких электронных языковых корпусов.

Kalit so‘zlar: korpus lingvistikasi, teg, lingvistik izohlash, lingvistik va ekstralingvistik razmetka, lingvistik teglash turlari.

Bugun tilshunoslik sohasida korpus lingvistikasi jadal sur’atlar bilan rivojlanib bormoqda. Korpus lingvistikasi kompyuter lingvistikasining tarkibiy qismi bo‘lib, kompyuter texnologiyasi yordamida tilni amaliy tadqiq qilish, til korpusini yaratish kabi muhim vazifani o‘taydi. Korpus – ma’lum maqsadda yig’ilgan matnlar majmuini tashkil etuvchi til birliklari yig’indisi. Korpus oddiy elektron matnlar bazasi bo‘libgina qolmay, muayyan tilning o‘ziga xos xususiyatlari va variantlarini aks ettiruvchi bir necha belgi asosida tanlab olingan elektron shakldagi matnlar yig’indisidir.

Korpusni boshqa elektron to‘plam, onlayn kutubxona va ensiklopediyalardan farqlovchi, matnda turli lingvistik amallarni bajarishga imkon beruvchi eng muhim amallardan biri korpus birliklarini **teglash** jarayonidir. Bunda matn birliklarining har biriga maxsus izoh beriladi va bu izoh matn *annotatsiyasi* deyiladi. Mazkur jarayonni amalga oshirish *lingvistik annotatsiyalash* yoki *lingvistik izohlash* deb nomlanadi. Lingvistik izohlash jarayonida har bir til birligini ramzlash esa *teglash* yoki *razmetkalash*, mana shu teglashda foydalaniladigan maxsus belgilar esa *teglar* deb ataladi. Teglashning keng tarqalgan va qamrovi jihatdan ko‘p tadqiq qilingan turlaridan biri bo‘lgan morfologik teglash asosida bir misol ko‘ramiz. Kecha bu kitobni o‘qidim. Ushbu gapdagi har bir birlikning so‘z turkumi va grammatik kategoriyalari maxsus teglar yordamida teglanadi. *Kecha (Adv – ravish), bu (P – olmosh), kitobni (N – ot, bir., tush.kel.q), o‘qidim(V – fe’l, I sh., o‘tg. am., aniq nis. ...)* .

Jahon tilshunosligi doirasida V.P.Zaxarov, B.Kutuzov, S.A.Sharovlar bu boradagi ishlarida teg, uning korpusdagi amaliy ahamiyatini atroflicha yoritib berishgan[1]. O‘tgan asrning 80-yillarida SGML (Standard Generalized Markup Language) nomi ostidagi elektron matnlar razmetkasi standarti qabul qilingan edi. Bu standart dastlab tipografiya sanoati uchun ishlab chiqilgan bo‘lsa-da, tez orada boshqa sohalarda ham qo‘llanila boshlandi. SGMLning mohiyati turli matn muharrirlarida terilgan hujjatni tahrirlash, tahlil qilish, o‘zgartirishdan iborat.

DOI: 10.5281/zenodo.19905440

Lingvistik izohlashda har bir so‘z ma’lum bir kodga ega bo‘ladi. Bugungi kunda matnga lingvistik, boshqa ma’lumotlarni qo‘shishning umume’tirof qilingan standarti mavjud emas. Lekin Text Encoding Initiative (TEI) maxsus xalqaro loyihasi razmetkaning standart vositasini ishlab chiqishga mo‘ljallangan. Buning uchun hujjat razmetkasining butun xalqaro qabul qilingan tili – SGML va XML mavjud[2]. Dastavval korpus matnlari birliklari mutaxassislar tomonidan qo‘lda teglangan. Ushbu jarayonda aniqlik yuqori bo‘lsa-da, uzoq vaqt va og’ir mashaqqat talab etardi. Keyinchalik inson omili ishtirokida avtomatik teglash yo‘lga qo‘yildi.

Korpus lingvistikasida razmetkaning *lingvistik* hamda *ekstralingvistik* turlari ajratiladi. Ekstralingvistik teg bu – matn haqidagi tashqi ma’lumot. Ekstralingvistik razmetkaning quyidagi turlari farqlanadi:

1. Matn formatining o‘ziga xosligini aks ettiruvchi (bob, xatboshi, qism va h.) teg.
2. Matn, uning muallifiga tegishli ma’lumotni ifodalovchi teg. Chunki muallif haqidagi ma’lumot nafaqat nom, balki yosh, jins, u yashagan yil kabi ma’lumotlarni ham bildirishi mumkin. Matn haqidagi ma’lumot esa matn (asar) nomidan tashqari tili, yozilgan hamda nashr etilgan yilini qamrab oladi. Bunday ma’lumotlarning mavjudligi bazada detallashtirilgan qidiruvni amalga oshirish imkonini beradi.

Lingvistik razmetkaning *morfologik teglash* (POS-tagging), *sintaktik teglash* (Parsing), *semantik teglash*, *anaforik teg* (ma’nosi olmosh bilan ifodalanayotgan so‘z alohida kodlanib, keyingi o‘rinda shu so‘zga ishora qilayotgan olmosh yoniga shu kod biriktiriladi va natijada olmoshning matndagi ma’nosi aniqlanadi), prosodik teg (ovoz transkripsiya qilingan korpusda urg‘u, ohangni ifodalovchi izoh) kabi ko‘rinishlari mavjud.

O‘zbek tilida morfologik teglashga doir bir qancha tadqiqot ishlari olib borilgan. Ammo lingvistik izohlash jarayonida til birliklari orasida yuz beradigan omonimiya va polisemiya kabi muammolar tufayli semantik teglashga doir qilingan tadqiqot ishlari ozchilikni tashkil qiladi. Semantik teglarning yagona standart shakli yo‘q. Harf, raqam yoki faqat raqamlardan iborat kodlardan foydalaniladi. Semantik teg korpusdagi so‘z ma’nosining ixtisoslashuvi, omonimlik, sinonimlik, ma’noviy guruhga ajratish kabi muammolarni hal qiladi. V.P.Zaxarov, S.YU.Bogdanovalar rus tili milliy korpusini tuzishda semantik razmetkalashning o‘z variantini taklif qiladi[3].

Xulosa o‘rnida aytish joizki, korpus birliklarini teglash tilning amaliy ahamiyatini ochib berishda, leksik birliklarning ma’nolarini aniqlashda, omonimlik va polisemantik hodisalarga aniqlik kiritishda muhim ahamiyat kasb etadi.

FOYDALANILGAN ADABIYOTLAR:

1. Курс “Корпусная лингвистика” (А.Б. Кутузов) Лицензия Creative commons Attribution Share-Alike 3.0 Unported (Электрон ресурс) - //lab314.brsu.by/kmp-lite/kmp-video/CL/CorporeLingva.pdf.
2. Zaxarov V., Mengliyev B., Hamroyeva Sh. Korpus lingvistikasi – o‘quv qo‘llanma. – Toshkent, 2023. – B.125.
3. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2011. – Б.45.
4. Ҳамроева Ш.М. Корпус лингвистикаси атамаларининг қисқача изоҳли луғати. – Тошкент: Камалак, 2018. – B.96.
5. Abjalova M. Korpus birliklarini teglash va annotatsiyalash masalasi. // Zamonaviy ilm-fan va ta’lim istiqbollari. Respublika ko‘p tarmoqli ilmiy- amaliy konferensiya to‘plami. – Toshkent, 2023. – 345-351b.