KORPUS MATERIALLARINI SEMANTIK TEGLASH PRINSIPLARI

Abjalova Manzura Abdurashetovna
Filologiya fanlari doktori (DSc),
Toshkent davlat o‘zbek tili va adabiyoti universiteti dotsenti
E-mail: abjalova.manzura@gmail.com

Turdialiyeva Muxlisa Dilshodbek qizi
ToshDO‘TAU O‘zbek filologiyasi fakulteti
Amaliy filologiya yo‘nalishi 2-bosqich talabasi
E-mail: turdialiyevamuxlisa24@gmail.com

Annotatsiya. Semantik teglash korpus matnlaridagi leksik birliklar semantikasini ochish va mavjud hollarda uning bir qancha ma’nolari yoki omonimligini aniqlash imkonini beradi. Semantik tahlil tabiiy tilni qayta ishlash (NLP)da asosiy jarayonlardan hisoblanib, korpusda bu semantik teglar yordamida amalga oshiriladi va foydalanuvchiga semantik annotatsiya shaklida taqdim etiladi. Ushbu maqolada o‘zbek tili korpuslari uchun semantik teglashning ahamiyati, til korpuslarini semantik teglash tizimini ishlab chiqish, semantik izoh berish xususida fikr yuritilgan.

Kalit so‘zlar: *korpus, matn birliklari, so‘z, semantik teg,semantika*.

Annotation: Semantic marking allows you to identify the semantics of lexical units in corpus texts and determine their multiple meanings or homonyms in existing cases. Semantic analysis is one of the main natural language processing (NLP) processes that is carried out on a corpus using these semantic tags and presented to the user in the form of a semantic annotation. This article discusses the importance of semantic tagging of linguistic units in Uzbek language corpora, the development of a system for semantic tagging of corpora and semantic annotation.

Keywords: *corpus, text units, word, semantic tag, semantics.*

Kirish. Bugungi kunda takomillashib borayotgan raqamli texnologiyalar asrida insonga ko‘plab qulayliklar yaratish maqsadida turli dasturlar ishlab chiqilmoqda, ixtirolar qilinmoqda. Jumladan, tarjimon dasturlari, matnlarni avtomatik tahlil dasturlari, matnlar generatsiyasi, matn mazmunini saqlagan holda uni avtomatik qayta yozish kabi dasturiy ta’minot va tizimlar yaratilgan, hozirda ular takomillashtirilmoqda. Bunday dastur va tizimlar inson vaqtini tejashga, katta hajmli matnlar ustida ishlashga, samarali natijaga tezkor erishishda muhim vosita hisoblanadi. Hozirda rivojlanayotgan sohalardan biri korpus lingvistikasi ham tadqiqotchilar maqsadiga tezroq erishishi uchun qulay imkoniyat yaratishga xizmat qilmoqda. Asosiysi, ta’lim jarayonida o‘qituvchi va o‘quvchilar uchun juda muhim o‘quv quroli sanaladi [7, 8]. Buning sababi korpus materiallari grammatik va semantik teglanadi. Til korpuslarini semantik teglash, uning ahamiyati haqida quyida batafsil fikr yuritamiz.

Semantik teglash – leksik birlik (so‘z yoki ibora)ning lug‘aviy ma’nosi bo‘yicha semantik guruh, aniqroq qilish maqsadida subsemantik guruhlarga mansubligini belgilash va ma’nosi (ma’nolari)ni aniqlash jarayoni. Semantik guruhlarga raqamli texnologiya ishi uchun maxsus belgilar beriladi, ular semantik teg uchun izohlar majmui hisoblanadi. Semantik teglash korpus materialida kontekstual ma’noni ochishga ham yordam beradi. Shuning uchun morfologik va

sintaktik teglash kabi semantik teglash ham juda muhim jarayon hisoblanadi. Masalan, kiyim ma'nosini bildiruvchi so'zlar ishtirok etgan kontekstlarni ajratib berish buyrug'i berilsa, semantik tegga ega matn bu so'rovni qayta ishlay oladi va so'ralgan semali lemma ishtirok etgan kontekst ajratib ko'rsatiladi. *Pidjak, ko'ylak, jemper* kabi so'zlarning qo'llanish chastotasini yillar kesimida o'rganish imkoni bo'lsa, *tulup, armyak, poddelka* kabi past chastotali so'zlarning turli tarjimonlar tilida qanday ifodalanishini kuzatish ham mumkin, chunki bunday so'zlar lug'atlarida kam uchraydi.

V.P.Zaxarov fikricha, lingvistik tegning barcha (morfologik, sintaktik, semantik, anaforik, prosodik) turlari quyidagi tamoyillar asosida amalga oshiriladi [11]:

1) teg sxemasini tavsiflash (asoslash);
2) umumiy lingvistik tushunchalar tizimini aniqlash;
3) foydalanuvchi uchun ma'lum bo'lgan tahlil sxemasini shakllantirish;
4) teg sxemasining nazariy an'anaviyligiga erishish;
5) xalqaro andozalarga amal qilish.

O'zbek tili korpusi materiallari morfologik, sintaktik, semantik, anaforik, prosodik, diskurs va stilistik jihatdan annotatsiyalash [Abjalova, 2022:30-31], [9] mumkin. Buning uchun korpus matnlaridagi leksik birliklarni teglashda quyidagi prinsiplarga amal qilish maqsadga muvofiq sanaladi:

1. Nazariy jihatdan **neytral (an'anaviy) izohlash sxemasi** – har bir korpusga o'ziga xos izohlash sxemalari, ya'ni elementlaridan foydalanib, chigalliklarni yuzaga keltirgandan ko'ra, yirik lingvistik korpuslarni teglash spetsifikasidagi elementlarni umumfoydalanish uchun asos qilib olinishi o'zbek tili korpuslarining jahon talabidagi zamonaviy korpus deya e'tirof etilishiga asos bo'ladi va o'z o'rnida bunday korpus standart korpus vazifasini o'taydi. Modomiki ko'pchilikka ma'lum izohlash sxemasidan foydalanilmasdan mualliflik nazariyasi yaratib olinsa, korpusdan foydalanuvchi izohlash tizimini chuqurroq o'rganib chiqishga majbur bo'ladi. Tabiiyki, bunday ortiqcha izlanish foydalanuvchiga ma'qul kelmaydi.
2. **Lingvistik tushunchalarning umumiy qabul qilingan tizimi** – lingvistik korpuslarning dunyo miqyosida ahamiyatga ega bo'lishi uchun teglar xalqaro belgi va ramzlardan iborat bo'lishi o'rinli. Bu, asosan, lingvodidaktikada ahamiyatli hisoblanib, til o'rganish va o'rgatish jarayonini yanada qulaylashtiradi.
3. **Parametrlarni samarali kiritish** – har bir lug'aviy birlikning grammatik va semantik parametrlari iyerarxiyali tarzida mukammal kiritilishi nafaqat ma'lumotlarning aniq chiqishining garovi hisoblanadi, balki omonimlik [3, 5], ko'pma'nolilik va polifunksionallik [6] holatlarini aniqlash imkoni ham yuzaga keladi. Buning uchun lingvist va dasturchilarning o'zaro hamfikr ishlashi muhim sanaladi. Juda katta miqdordagi lingvistik birliklarni to'g'ri teglashda inson omili va yarim avtomat jarayoni ishonchlidir. Buning uchun fidoyi mutaxassislar jamoasining sermahsul mehnati talab qilinadi.
4. **Xalqaro standartlarga rioya qilish** – teglashning TEI xalqaro standartiga rioya etish ulkan tajribaga tayanish hisoblanadi.

G.I.Kustova Rus tili milliy korpusi semantik razmetkasini tahlil qilar ekan [12], asosiy e'tiborni semantik razmetkaga asos bo'lgan "Leksikograf" bazasiga qaratadi. Ushbu tizim Rossiya Fanlar akademiyasi rus tili ilmiy tadqiqot institutida 1992-yilda ishlab chiqilgan. Xususan, bazaning eng asosiy qismi sanalgan fe'llarning lingvistik ta'minoti E.V.Paducheva, narsa otlari E.V.Raxilina tomonidan tuzilgan [13, 14]. Ushbu tizim oldindan semantik izohlar bilan

shakllantirilgan so‘zlarni ro‘yxatga olish uchun mo‘ljallangan lug‘atdan iborat yuqori texnologiyali lingvistik baza hamda rus tili milliy korpusining avtomatik-semantik razmetkalash asosi sanaladi [https://ruscorpora.ru/page/instruction-semantic/].

Semantikada ma’nolar va ularning obyektlari bilan bog‘liqlik e’tiborga olinadi. Unda so‘z(shakl)larning obyekt ifodalagan tushunchani o‘zida aks etish imkoniyati, so‘zlarning obyektlar bilan bog‘liqlik darajasi va ularni qo‘llash jarayonida yuzaga kelgan qo‘shimcha (pragmatik) ma’nolari tadqiq etish bilan shug‘ullanadi. Shuningdek, semantikada semema (leksik ma’no) va sema (leksik ma’noning tarkibiy qismi) munosabati, integral va differensial semalarning obyekt ta’rifini berishdagi ahamiyati ham o‘rganiladi. Korpusda leksik-semantik guruhlar sememalardagi semalarning shu sememalar uchun umumiy yoki xususiy ekaniga ko‘ra tashkil qilinadi. Bu holat semantik annotatsiyaning aniqligini va yuqorida tilga olingan muammolarni bartaraf etishda muhim sanaladi. Yuzaga kelgan semantik guruhlar maxsus teglanadi. Natijada korpus menejeri (maxsus qidiruv tizimi)da filtrli qidiruv imkoniyati ham yuzaga keladi. Semantik teg bir turdagi ko‘plab so‘z(shakl)larni bitta semantik guruhga birlashtiradi. Masalan: *ipak, pilla, tut* so‘zlari semantik munosabatlari asosida *ipak* mazmuniy guruhini tashkil etsa, *isiriq, tutatmoq, o‘simlik, o‘smoq* so‘zlari *o‘simlik* mazmuniy guruhining *isiriq* subguruhini tashkil etadi.

IPAK – 1. Pilla qurti chiqargan juda ingichka, pishiq, yaltiroq, tovlanib turadigan muloyim tola va undan yigirilgan ip. *Xom ipak. Pishitilgan ipak.*

2. Shunday iplardan to‘qilgan mato, shoyi va undan tikilgan kiyim yoki buyum. *Ipak gilam. Ipak ro‘mol.* [O‘TIL, 2-jild, 224]

PILLA I – [fors – ipak qurti pillasi, g‘umbak] 1. Ipak qurti g‘umbakka aylanishdan oldin o‘z ustiga o‘rab hosil qilgan qobiq, ipak qurtining mahsuli; tabiiy ipak olinadigan xomashyo. 2. Ko‘pgina hasharotlar g‘umbagining himoya qobig‘i. [O‘TIL, 3-jild, 258]

TUT – [a. – tut daraxti] 1. Barglari ipak qurti uchun asosiy oziq bo‘ladigan daraxt. 2. Shu daraxtning iste’mol qilinadigan oq, qora yoki qizil tusli totli mevasi. [O‘TIL, 4-jild, 205]

ISIRIQ – Isiriqdoshlar oilasiga mansub ko‘p yillik yovvoyi o‘simlik – [Tarkibida alkaloidlar bo‘lib, xalq tabobatida shifobaxsh vosita sifatida turli yo‘llar bilan (jumladan, tutatib) foydalanib kelinadi]. [O‘TIL, 2-jild, 229]

TUTATMOQ – 1. Tutamoq fe’lining orttirma nisbatdagi shakli. *Pechkani tutatmoq*. 2. O‘t oldirib, yondirib olmoq; yondirib olib chekmoq. [O‘TIL, 4-jild, 206]

O‘SIMLIK – Havo va tuproqdagi organik va anorganik moddalar bilan oziqlanadigan, anorganik moddalarni organik moddalarga aylantirish xususiyatiga ega bo‘lgan, odatda, biror joyga o‘rnashgan holda rivojlanadigan organizm. [O‘TIL, 5-jild, 173]

O‘SMOQ – 1. Rivojlanib bo‘yiga yoki uzunasiga o‘zini qo‘ymoq, cho‘zilmoq, rivojlanmoq. 2. Ko‘kargan (o‘sgan) holda bo‘lmoq (daraxt, o‘simlik haqida). 3. Katta bo‘lmoq, balog‘atga yetmoq; ulg‘aymoq. 4. Son-miqdor, daraja va sh. k. jihati ortmoq, kuchaymoq. [O‘TIL, 5-jild, 173]

Semantik analizatorlar, matnlar va so‘zlar tizimida yashaydigan ma’nolarni tahlil qilishga yordam beruvchi dasturlar hisoblanadi. Bu analizatorlarning bir nechta afzalliklari biz uchun ancha asqotadi. Jumladan, semantik analizatorlar matnlarni tahlil qilish va so‘zlarning ma’nolarini aniqlashda yordam beradi. Semantik analizatorlar matnning konteksti bo‘yicha so‘zlar yoki so‘z birikmalarining ma’nolarini tahlil qiladi, aynan kontekstual tahlil ushbu analizatorning ahamiyatini yanada oshiradi. Bu analiz kontekstning o‘zida ma’no o‘zgarishlarini tushunishga yordam beradi. Semantik analizatorlar so‘zlar o‘rtasidagi mantiqiy bog‘liqliklarni tahlil qilishga ham yordam beradi. Bunda so‘zning yangi ma’nonisini aniqlash imkoni ham bo‘ladi. Semantik

analizatorlar ma'nolarni tushunish, matnning konteksti va mantiqiy bog'liqliklarini tahlil qilish uchun ham muhim.

Semantik teglashda leksik birlikning qo'shimcha yoki kontekstual ma'nosidan kelib chiqqan holda tavsiflovchi teglar qo'shish mumkin. Bu raqamli texnologiya Ishida material tarkibini hamda uning boshqa ma'lumotlarning qismlari bilan aloqalarini yaxshiroq tushunishga yordam beradi. Hozirgi kunda semantik teglar axborot qidiruvi, so'z turkumlarini teglash [2, 4, 6], so'zning yangi ma'nolarini aniqlashda tahlilni yaxshilash uchun matn, rasm yoki boshqa tur ma'lumotlardagi so'zlar, iboralar yoki obyektlarni belgilashni ham o'z ichiga olmoqda. Bu qidiruv tizimlari, tavsiyalar tizimlari va ma'lumotlar tahlili kabi turli ilovalarda aniqroq axborot olish, kontentni tashkil qilish va ma'lumotlarni qayta ishlash imkonini beradi.

Semantik tegning ham, boshqa teglarda bo'lganidek, yagona standart shakli bo'lmasa ham, harf, raqam yoki faqat raqamdan iborat kodlardan foydalaniladi. Birinchi harf yoki raqam umumiy semantik ma'noni, keyingi belgi esa so'z ma'nosini yanada maxsuslashtiruvchi kichik semantik guruhni ifodalaydi. Semantik teg nafaqat so'z, balki ko'plab birikmalarni ham semantik guruhlarga birlashtiradi, bunday paytda turli birikuvdagi bir ma'noni bildiruvchi birikmalar bitta belgi bilan kodlanadi. Ibora (idiomatik birlik) tarkibidagi so'zlar miqdorini bildiruvchi axborot ham tegdan joy oladi. Semantik teg korpusdagi so'z ma'nosining ixtisoslashuvi, omonimlik, sinonimlik, ma'noviy guruhga ajratish kabi muammolarni hal qiladi.

Til korpuslarini semantik teglash, semantik guruhlarga mansub so'zlarning izohlarini berish va shu bilan bir qatorda, asarlardan olingan misollar yordamida tushuntirish o'zbek tili korpusi foydalanuvchilari uchun keng imkoniyatlarni yuzaga chiqaradi. Korpusning semantik teglari so'z ma'no(lar)ining spetsifikasi, ya'ni o'ziga xos xususiyatlarini, ularning omonimlik, sinonimlik, antonimlik bilan bog'liq izohlar majmuini tuzish, so'zni kategoriyalash, semantik maydonga mansubligini belgilash, derivatsion xarakteristikasi, atash ma'nosi kabi belgilarni qamrab oladi. Shuning uchun semantik teglash korpusdagi so'zlarning ko'p ma'nolilik, omonimlik, sinonimlik xususiyatlarini hamda so'z ma'nosining ixtisoslashuvi, ularning ma'noviy guruhlarini aniqlash imkonini beradi va matndagi semantik muammolarni hal qiladi. Bu kabi imkoniyatlar nafaqat tadqiqotchilar uchun, balki boshqa ko'plab kasb egalari uchun ham qulaylik yaratadi.

Foydalanilgan adabiyotlar:

1. Abjalova M.A. Korpus lingvistikasi: uslubiy qo'llanma. – Toshkent: Bookmany print, 2022. – 110 b.
2. Abjalova M.A. Tahrir va tahlil dasturlarining lingvistik modullari. [Matn] : monografiya / M.A. Abjalova. – Toshkent: Nodirabegim, 2020. – 176 b. ISBN 978-9943-6939-0-6
3. Abjalova M. Milliy korpusi mavjud bo'lmagan tillarning lingvistik dasturlarida omonimlarni tahlil qilish texnologiyasi. // So'z san'ati. Xalqaro jurnal. – Samarqand, 1/2020. – B. 117-128. ISSN 2181-9297 DOI http://dx.doi.org/10.26739/2181-9297
4. Abjalova M., Elova D. Tabiiy tilni qayta ishlash (NLP)da so'z turkumlarini teglash masalasi. // O'zbekistonda til va madaniyat, – Toshkent: ToshDo'TAU, 1/2021. – B. 6-20.
5. Abjalova M. Omonimiya va lingvistik tizimlarda omonimlarni farqlash usullari. // "Oriental Renaissance: Innovative, Educational, Natural And Social Sciences (ORIENS). Vol.1, Issue 10. SJIF 2021-5.423. – pp. 1016-1021. www.oriens.uz ISSN 2181-1784.
6. Abjalova M., Iskandarov O. Methods of Tagging Part of Speech of Uzbek Language. // IEEE – UBMK – 2021: 6th International Conference on Computer Science and

Engineering. 15-16-17 September 2021. Ankara – Turkey. DOI: http://doi.org/10.1109/UBMK52708.2021.9558900 . – pp. 82-85. Impakt Factor 5.5

7. Abjalova M., Gulomova N. Author's Corpus of Alisher Navoi and its Semantic Database. // IEEE – UBMK – 2022: 7th International Conference on Computer Science and Engineering. 24-26 September 2022. – Diyarbakir, Turkey. – pp. 182-187. Impakt Factor 5.5. **DOI:** 10.1109/UBMK55850.2022.9919546
8. Abjalova M., E. Adalı and O. Iskandarov, "Educational Corpus of the Uzbek Language and its Opportunities," 2023 8th International Conference on Computer Science and Engineering (UBMK), Burdur, Turkiye, 2023, pp. 590-594, doi: 10.1109/UBMK59864.2023.10286682.
9. Abjalova M. Tagging and Annotation of Corpus Units. // International Journal of Language Learning and Applied LinguisticsISSN: 2835-1924 Volume 2 | No 12 | Dec -2023. – PP. 103-107. https://inter-publishing.com/index.php/IJLLAL/article/view/3228/2733
10. O'zbek tilining izohli lug'ati: 80 mingdan ortiq so'z va so'z birikmasi. / A.Madvaliyev tahriri ostida. – Toshkent: O'zbekiston Milliy ensiklopediyasi. 5 jildli. 2006.
11. Захаров В.П., Богданова С.Ю. Корпусная лингвистика. – Иркутск: ИГЛУ, 2011.
12. Кустова Г.И., Ляшевская О.Н., Падучева Е.В., Рахилина Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. Москва: Индрик, 2005.
13. Кустова Г.И., Падучева Е.В. Словарь как лексическая база данных // ВЯ, 1994. – №4.
14. Рахилина Е.В., Кобрицов Б.П., Кустова Г.И., Ляшевская О.Н., Шеманаева О.Ю. Лексико-семантическая разметка в национальном корпусе русского языка.