KORPUS TIZIMIGA FRAZEOLOGIK BIRLIKLAR TATBIQ ETILISHI MASALASI

Nurboyeva Maftuna Vahobjon qizi
Filologiya bo‘yicha falsafa doktori (PhD),
Millat umidi universiteti o‘qituvchisi

Annotatsiya. Maqolada tabiiy tilni avtomatik qayta ishlash tizimlarida qo‘llaniladigan morfologik, sintaktik va semantik analizatorlarning o‘ziga xos xususiyatlari hamda ularning til birliklarini avtomatik tahlil qilishdagi ahamiyati yoritiladi. Xorijiy va o‘zbek tilshunosligida semantik analizatorlarni yaratish va takomillashtirishga doir yondashuvlar tahlil qilinib, semantik tahlilda frazeologik birliklarning o‘rni va ularni lingvistik modellashtirish muammolariga alohida e’tibor qaratiladi.

Kalit so‘zlar: semantik analizator, morfologik analizator, sintaktik analizator, semantik tahlil, tabiiy tilni qayta ishlash, lingvistik modellashtirish, frazeologik birlik, frazeologik model, avtomatik tahlil.

Abstract. The article discusses the specific features of morphological, syntactic, and semantic analyzers used in natural language processing systems and their significance in the automatic analysis of linguistic units. Approaches to the development and improvement of semantic analyzers in foreign and Uzbek linguistics are analyzed, with particular attention paid to the role of phraseological units in semantic analysis and the problems of their linguistic modeling.

Keywords: semantic analyzer, morphological analyzer, syntactic analyzer, semantic analysis, natural language processing, linguistic modeling, phraseological unit, phraseological model, automatic analysis.

Аннотация. В статье рассматриваются особенности морфологического, синтаксического и семантического анализаторов, применяемых в системах автоматической обработки естественного языка, а также их значение для автоматического анализа языковых единиц.

Анализируются подходы к разработке и совершенствованию семантических анализаторов в зарубежной и узбекской лингвистике, особое внимание уделяется роли фразеологических единиц в семантическом анализе и проблемам их лингвистического моделирования.

Ключевые слова: семантический анализатор, морфологический анализатор, синтаксический анализатор, семантический анализ, обработка естественного языка, лингвистическое моделирование, фразеологическая единица, фразеологическая модель, автоматический анализ.

Korpus tizimida frazeologik omonimiyani lingvistik modellashtirish, u orqali yuzaga kelishi mumkin bo‘lgan yanglishlik va noaniqliklarni bartaraf etish xususida yaxlit tadqiqotni uchratmadik, ammo ayni paytda o‘zbek tilshunosligida 10 ga yaqin subkorpuslar yaratilganligini ta’kidlash joiz, xususan, O‘zbek tili ta’limiy korpusi, Parallel korpus, O‘zbek tili milliy korpusi va yana bir qator mualliflik korpuslari mavjud[9–12]. Ularning faoliyat qamrovi va jarayoni bilan tanishganimizda iboralar ham qisman izohlanganining guvohi bo‘ldik. Jumladan, milliy korpus tizimiga *bosh, qo‘l, dars, til* kabi so‘zlar tizimga qidiruvga berilganda avval shu so‘zning semantikasi, namunalar va shu so‘z ishtirokidagi iboralar ham biriktirilganini ko‘rdik. Lekin bu iboralarning omonim yoki polisemantikasiga oid punktlari ochiq qolgan. Tizimga butun iboraning o‘zi qidiruvga berilganda ham sanoqli iboralarni aniqlay olmoqda. Va uning semantik va formal imkoniyatlari ochiqlanmagan. Shuningdek, milliy korpusga “Oila ma’naviyati” konseptining elektron tezaurusi ham kiritilgan bo‘lib, mazkur baza bo‘yicha ham qidiruv olib borganimizda konseptga oid tushunchalar ishtirokidagi iboralar va ularning izohlari hamda ular ishtirokidagi namunalar keltirilgan (*nasldan naslga o‘tmoq, onasini Uchqo‘rg‘ondan ko‘rsatmoq, ota go‘ri qozixonami?, tirnoq orasidan kir qidirmoq* kabi)[13], shu qatorida, leksik omonimiyani bartaraf etish, korpus tizimi omonim xarakterli leksik birliklarni ajrata olishi bo‘yicha; frazeologizmlar miqyosida esa umumiy sath bo‘yicha va muayyan tillarning o‘zaro frazeologik

muvofiqligini ochiqlovchi parallel korpus bo‘yicha tadqiqotlar talay ekanini ta’kidlash lozim. Xususan, rus tilshunosligida omonimiyani qisman bartaraf etish bo‘yicha ham “qo‘lda” morfologik tahlil qilish jarayoni ishlab chiqilgan. Unga ko‘ra, korpus tizimiga so‘zlarning turkumi, turkumdagi semantik sinfi xususida ma’lumotlar kiritiladi va so‘zlarning eng yuqori chastotada birikib keluvchi yondosh birliklar ham tizimga tatbiq etiladi, samaradorlik jarayoni uch bosqichda amalga oshirilib, birinchi bosqichda korpusdagi umumiy va tanib olinmagan so‘z qo‘llashlar soni hisoblab chiqiladi, lemmalizatsiya jarayoni 98% aniqlikda amalga oshadi, ikkinchi bosqichda leksemalararo omonimlikka ega bo‘lgan so‘zshakllar soni hisoblanadi, so‘nggi bosqichda esa semantik belgilashning sifati ustida ishlanadi, bunda faqat mustaqil so‘z turkumlariga oid lemmalar ishtirok etadi, yordamchi birliklar hisobga olinmaydi[1].

D.N.Gomon korpus tizimida omonimiyani bartaraf etish bu – omonim so‘zshakllarni bir-biridan ajratib turadigan belgini aniqlashdir, deya izohlab, barcha omonimlar qaysidir belgisi bilan hamrohidan farq qilishi va shu farq tizimga uni ajrata olish imkonini berishini ta’kidlaydi[2]. Deomonimizatsiya jarayonini amalga oshirishning esa quyidagi usullarini ko‘rsatib o‘tadi:

1) asosiy leksik va grammatik qurshov orqali;

2) kontekst tematikasi va nutq vaziyati yordamida. Bu usul haqida boshqa tadqiqotlarda ham ma’lumot berilgan bo‘lib, unga ko‘ra omonimiyadagi noaniqlikni hal qilish uchun uning o‘ng va chap tarafdagi leksik birliklar muhimdir, deyiladi. E.Reyfler esa bu jarayonda butun jumladan ham foydalanish mumkinligini aytadi, ya’ni matn tarkibidagi omonim quyidagi ko‘rinishlarda aniqlanishi mumkin: *preceding* – o‘zidan oldingi so‘z yordamida, *following* – o‘zidan keyingi so‘z yordamida, *both* – bitta oldingi va bitta keyingi so‘z yordamida, *preceding2* – o‘zidan oldingi ikkita so‘z yordamida, *following2* – o‘zidan keyingi ikkita so‘z yordamida, *both2* – o‘zidan oldingi va keyingi ikkita

so‘z yordamida, *sentence* – butun jumla tahlili orqali[3]. Bu usullar omonim birlikning turkumi va murakkablik darajasiga ko‘ra mos tarzda qo‘llaniladi.

D.N.Gomon omonimiyani bartaraf etish xususida tadqiqotlar ko‘p bo‘lsa-da, ularning aksariyatida avtomatik tarjima tizimlarining kerakli ma’noni mustaqil tanlay olishi masalasini hal qilmasligini ta’kidlaydi va belorus tilshunosi A.V.Zubovning ishidan o‘rnak keltiradi[2, 4]: tadqiqotchi ishida “qiya” so‘zining 6 xil omonimligini keltirib, ularning ikkitasi urg‘u bilan farqlanishini, so‘zning har bir ma’nosi uchun sakkiztagacha diagnostik belgilar – so‘zning muayyan semantikasini ochishga xizmat qiluvchi so‘zlarni belgilaydi, ammo muammo shundaki, mazkur omonimshaklning biror kontekstda qaysi ma’nosi ifodalanayotganini topish uchun tizim 8 so‘zdan iborat yuzlab jumlani so‘zma-so‘z tekshirishga majbur. D.N.Gomon omonimlarni matnning cheklangan qismi yordamida bartaraf etishga intilib, quyidagi bosqichlarda farqlash mumkinligini ko‘rsatadi: predlog (P – so‘z tartibi va morfologik analiz yordamida farqlash), urg‘u (U), grafik farqlar (G). Uning fikricha, bular ichida omonimiyani bartaraf etishda eng ko‘p xizmat qiladigani, shubhasiz, morfologik analizdir. Negaki, omoniyani kompyuter yordamida bartaraf etishda urg‘u, ovozni elektron tizimda tahlil qilishda grafik belgi bosqichlari o‘z ahamiyatini yo‘qotadi.

O‘zbek tilshunosligida ham korpus tizimida omonimiyani bartaraf etish bo‘yicha tadqiqotlar mavjud. Jumladan, M.Abjalova tomonidan matnni avtomatik tahlil qilish uchun lingvistik ta’minotni ishlab chiqish jarayonida omonimiya muammosini hal qilish bo‘yicha tavsiyalar berilgan va tizim omonim shakllarni ajrata olishi uchun qo‘llanilishi samara beradigan metodlar ham ko‘rsatib o‘tilgan[5]. Chunonchi, korpusda omonimiyani bartaraf etishdagi metodlarni grammatik me’yor va statistikaga asoslangan usullarga ajratilib, har bir sinfga kiruvchi metodlarning afzallik va kamchiliklarini sanab o‘tiladi. Masalan, Brill usuli[6] haqida to‘xtalib, bu usulda, avvalo, har bir so‘zning yondosh birliklari bog‘lanishi kerak bo‘ladi, jarayonga notanish so‘zlar uchrab qolsa, ot so‘z turkumi

sifatida qabul qilinadi, so‘ngra tez-tez uchrovchi xatoliklar uchun qayta ishlash mexanizmi yaratiladi va istalgan minimal xatoga erishgunga qadar ikkinchi bosqich takrorlanaveradi, tadqiqotchi metodning kamchiligi sifatida qoidalar sonining ko‘payishi aniqlik darajasining pasayishiga olib kelish holatini ko‘rsatib o‘tadi. Shuningdek, Yashirin Markov modeli haqida ham to‘xtalib, bu metod omonimlikni aniqlashning statistik metodi hisoblanib, jarayonda hosil bo‘ladigan barcha ehtimolliklarni nazarda tutishga intiladi va teglar chastotasi bo‘yicha kiritilgan ma’lumotlar orqali omonimiyani belgilashda aniqlikka erishishini ta’kidlaydi[7].

D.Ahmedova ham tadqiqotida o‘zbek tilidagi omonimlarni teglash muammolariga to‘xtalib o‘tib, korpus tizimida omonimshakllarni farqlash birmuncha murakkab jarayon ekanligini, murakkablik, asosan, bir xil so‘z turkumi doirasida omonimlik hosil qilayotgan birliklar o‘rtasida ro‘y berishini, chunki bunday so‘zshakllarda morfologik analiz natija bermasligini, ularni farqlashning birgina usuli korpus tizimida har bir ma’nosi izohlanishi kerakligi va bu yerda tizim foydalanuvchisidan ham biroz diqqat talab qilinib, shu kabi omonimshakllar qatnashgan kontekst izohi qidirilganda korpus tizimiga kiritilgan semantik tavsiflardan kontekstga mosini ajratib olishi lozimligini ta’kidlaydi[8].

Adabiyotlar ro‘yxati

1. Кобрицов Б.П., Ляшевская О.Н., Шеманаева О.Ю. Поверхностные фильтры для разрешения семантической омонимии в текстовом корпусе Microsoft Word – KobritsovBP.doc (murojaat sanasi: 19.04.2025).
2. Гомон Д.Н. Проблема снятия омонимии.pdf. (murojaat sanasi: 21.04.2025).
3. Рейфлер Э. Определение значения при машинном переводе // Машинный перевод. – М., 1957. – С. 185–222.
4. Зубаў А. В., Зубава І. І., Габіс А. А. Лінгвістычная інфарматыка (навучальны дапаможнік). – Минск, 1995. – 204 с.

5. Абжалова М.А. Лингвистик дастурлар учун омонимларни таҳлил қилиш технологияси. // Tilshunoslikdagi zamonaviy yo‘nalishlar: muammo va yechimlar. Xalqaro ilm.-amal. konf. –Andijon, 2020. – B. 666–668.
6. Brill E. A simple rule-based part of speech tagger // Proceedings of ANLS. 92. – C. 154–166.
7. Abjalova M. Tahrir va tahlil dasturlarining lingvistik modullari. Monografiya. – Toshkent. 2020. – 178 b.
8. Ахмедова Д. Атов бирликларини ўзбек тили корпуслари учун лексик - семантик теглашнинг лингвистик асос ва моделлари: Филол. фан. б. фалс. док. (PhD)... дисс. – Бухоро, 2020. – 247 б.
9. https://uzbekcorpus.uz/education
10. https://uzbekcorpus.uz/authorCorpus
11. https://uzbekcorpus.uz/parallel
12. http://uzbekcorpora.uz
13. http://thesaurus.uzbekcorpus.uz