

FE'L SO'Z TURKUMIDAGI LEKISIK OMONIMLARNI ANIQLASH YONDASHUVLARI

Malikov Elbek

malikovelbekk@gmail.com

ToshDO'TAU Kompyuter lingvistikasi
mutaxassisligi magistranti

Annotatsiya. Ushbu maqola omonimlikni aniqlash metodlarini ikki asosiy yondashuvga – statistik va qoidalarga asoslangan usullarga ajratadi. Statistik usullar matnlardagi omonimlarni morfologik tahlil orqali aniqlashda qo'llaniladi, unda grammatik xususiyatlarning birgalikda yuzaga kelish statistikasi ishlatiladi. Qoidalarga asoslangan tizimlar esa sintaktik tahlil bosqichida omonimlikni aniqlaydi va ko'proq mutaxassislar ishtirokini talab qiladi. Gibrid yondashuvlar, statistik metodlar va kam sonli qoidalar kombinatsiyasidan tashkil topgan. Tadqiqotlar rus va ingliz tillaridagi matnlarda turli metodlarning samaradorligini ko'rsatadi. Maqola, shuningdek, o'zbek tilida omonimiyani semantik farqlashga qaratilgan yangi yondashuvlarni muhokama qiladi.

Abstract. This article discusses methods for resolving homonymy, categorizing them into statistical and rule-based approaches. Statistical methods are used for morphological analysis to identify homonymy in texts, relying on the co-occurrence statistics of grammatical features. Rule-based systems address homonymy at the syntactic analysis stage and require significant input from specialists. Hybrid approaches combine statistical methods with a limited number of rules. Research demonstrates the effectiveness of various methods in texts in Russian and English. The article also explores new approaches for semantic differentiation of homonyms in the Uzbek language.

Аннотация. Данная статья рассматривает методы устранения омонимии, разделяя их на два основных подхода – статистические и основанные на правилах. Статистические методы используются для морфологического анализа, чтобы идентифицировать омонимию в текстах, опираясь на статистику совместного появления грамматических характеристик. Системы, основанные на правилах, устраняют омонимию на этапе синтаксического анализа и требуют значительного участия специалистов. Гибридные подходы состоят из комбинации статистических методов и небольшого количества правил. Исследования показывают эффективность различных методов на текстах на русском и английском языках. Статья также обсуждает новые подходы к семантическому различению омонимов в узбекском языке.

Kalit soʻzlar: omonimiya, tilshunoslik, statistik tahlil, morfologik tahlil, semantik farqlash, algoritmlar, tilni qayta ishlash, grammatik tahlil.

Kirish. Omonimlar tahlili tilshunoslikning dolzarb masalalaridan biridir, chunki ular tilning koʻplab aspektlarini, xususan, semantik va sintaktik oʻziga xosliklarini oʻrganishda katta ahamiyatga ega. Bitta soʻzning bir nechta maʼnoga ega boʻlishi nafaqat matnlarni tahlil qilish, balki avtomatik tarjima, tilni qayta ishlash (NLP) va kompyuter lingvistikasida muhim muammo hisoblanadi. Omonimiyalarni toʻgʻri farqlash, matnlarning aniq va toʻgʻri talqin qilinishini taʼminlash uchun zamonaviy metodlar va algoritmlar zarur. Statistik usullar, qoidalarga asoslangan tizimlar va gibril yondashuvlar bu masalani samarali hal qilishda asosiy vositalar hisoblanadi.

Oʻzbek tilida omonimlarni tahlil qilishda yuzaga keladigan muammolar va ularni bartaraf etish uchun qoʻllaniladigan metodlar haqida tadqiqotlar davom etmoqda. Bu masala nafaqat tilshunoslar, balki kompyuter olimlari va NLP mutaxassislari uchun ham jiddiy eʼtibor talab qiladi. Shuningdek, ingliz va rus tillaridagi tajribalar asosida oʻzbek tilida ham samarali yondashuvlar ishlab chiqilishi kerak. Omonimlarni aniq farqlash va tahlil qilishda ilmiy yondashuvlar, metodologiyalar va ilgʻor texnologiyalar birlashtirilishi lozim.

Asosiy qism. Omonimlarni farqlash metodlari qoidalarga tayanadigan statistik metodlar va gibril yondashuvlarga boʻlinadi.

Statistik usullar asosida qurilgan tizimlar morfologik tahlil bosqichida matnlarda omonimiyani bartaraf etadi, bunda qoʻl bilan belgilangan holatlardan grammatik xususiyatlarning birgalikda yuzaga kelish statistikasi qoʻllaniladi.

Qoidaga asoslangan tizimlar omonimlarni sintaktik tahlil bosqichida bartaraf etadi. Bunday tizimlarning rivojlanishi juda uzoq vaqt talab qiladigan jarayon boʻlib, soha mutaxassislari jalb qilish talab etiladi.

Gibril yondashuvlar statistik usullarga asoslangan va juda kam sonli qoidalar bilan taʼminlangan boʻladi.

Tajriba uchun rus tili milliy korpusi (RTMK) ishlatilgan. Podkorpusning bir qismida modelni oʻqitish olib borilgan, boshqa qismi esa omonimlarni farqlash usullarini sinab koʻrish uchun ishlatilgan. Ish jarayonida ikkita tajriba oʻtkazildi. Tayanch vektor metodi uchun ham, YMM uchun ham omonimning lokal (mahalliy) konteksti ishlatilgan. Ushbu ishda lokal kontekst ostida ikkitadan ortiq boʻlmagan qoʻshni chap va oʻng soʻz teglari qabul qilinadi.

Birinchi tajriba omonimiyani qisman bartaraf etish uchun oʻtkazildi. Eslatib oʻtamiz, omonimiyaning bu turi ingliz tili matnlariga xos. Ushbu tajriba rus tilidagi

matnlarda omonimiyani bartaraf etish mumkin bo'lgan usullarni aniqlash uchun o'tkazildi.

Ikkinchi tajriba grammatik shakllar majmuini aniqlash uchun o'tkazildi. Bunda so'z turkumlariga kelishik, shaxs-son, zamon, nisbat grammatik shakllari qo'shilgan hamda tajriba amalga oshirilgan. Leksik omonimlarni farqlash uchun ma'no tushunchasi kiritilgan. Ma'no deganda, predmetni aniq tasvirlaydigan satr konstantasi tushuniladi.

Omonimlikni aniqlash usullari belgilangan korpusning bir qismida sinovdan o'tkazildi. Omonimlarga lug'at bo'yicha tegishli bo'lgan teglar va lemma tanlandi. So'ngra har bir usul bilan bir ehtimoliy tegni aniqlab, korpusdagi teg bilan solishtirib ko'rildi.

Birinchi tajriba uchun so'z qismi to'g'ri aniqlangan bo'lsa, omonimiya bartaraf etilgan deb hisoblandi. Ikkinchi tajriba uchun yuqorida bayon etilgan grammatik xususiyatlar va bartaraf etilgan leksik omonimiyani ko'rsatkichi sifatida urg'u to'g'ri kelsa, omonimiyani bartaraf etish natijasi to'g'ri hisoblangan bo'ladi. Tajriba davomida olingan natijalar 1 va 2-jadvallarda berilgan. 1-jadvalda qisman omonimiyani, 2-jadvalda esa to'liq omonimiyani bartaraf etish natijalari aks ettirilgan¹⁶:

Metod	Qisman to'g'ri mansubligini aniqlash, %		
	barcha so'zlar	leksik omonimiyani hisobga olgan holda barcha so'zlar	notanish so'zlar
YaMM	92,78 93,56	90,56	87,32
Support vektor metodi		90,98	89,13

2.1-jadval. Rus tili matnlarida qisman omonimiyani bartaraf etish metodi natijalari

Metod	Teglarning to'g'riligini aniqlash, %		
	barcha so'zlar	leksik omonimiyani hisobga olgan holda barcha so'zlar	notanish so'zlar
YaMM	87,73	84,39	64,25
Support vektor metodi	86,12	82,89	62,41

2.2-jadval. Rus tili matnlarida omonimiyani aniqlash metodlari natijalari

¹⁶ Порохнин А. Анализ статистических методов снятия омонимии в текстах на русском языке // <http://www.mathnet.ru/vagtu286>

Shunday qilib, har ikkala metod ham rus tilidagi matnlarda omonimiyani qisman bartaraf etish ishini uddalay oladi. Biroq rus tilidagi matnlarda qisman omonimiyani bartaraf etish natijalari ingliz tilidagi matnlarga qaraganda ancha past. Rus tilidagi matnlarda omonimiyani aniqlash uchun Yashirin Markov modeli Support vektor metodiga qaraganda aniqroq natija beradi.

Omonimlarni semantik farqlashning ko'plab zamonaviy usullari bilan tajriba o'tkazishda nafaqat teglarni, balki o'ziga xos iboralarni ham qo'llash va omonimlarni turiga qarab bosqichma-bosqich bartaraf etishdan foydalanish ko'zda tutiladi.

M.Abjalova dissertatsiyasida omonimlarni tahlil qilish texnologiyasiga to'xtalib o'tadi. U Brill usuli va Yashirin Markov modelining vazifasi hamda afzalliklarini yoritadi. O'zbek milliy korpusining yaratilmaganligi, rasmiy va ilmiy uslubdagi matnlar bazasi to'la shakllantirilmaganligi e'tiborga olinib, o'zbek tilidagi omonim shakllarni tahlil qilishda N-grammga bog'liq ravishda omonim-tahlil usuli qo'llanishini ta'kidlaydi. Omonim-tahlil usuli muayyan omonim so'zga boshqa so'zning ma'lum grammatik ko'rsatkich orqali bog'lanishi bilan tushuntiriladi va bu jarayon tahlil imkoniyati hamda tahlilning aniqligini oshirishi ko'rsatiladi¹⁷.

Ingliz tilidagi matnlarda omonimiyani bartaraf etishni qiyoslash maqsadida ikki metod tanlab olingan: Support vektor metodi(SVM) va Yashirin Markov modeli (Hidden Markov Model HMM). Bu metodlar lug'atda bo'lmagan so'zlar uchun omonimiyani aniqlash imkoniyatini beradi.

Mashhur Yashirin Markov modeli (HMM) omonimlarni ehtimoliy yondashuv asosida aniqlashga qaratilgan. Algoritmning asosiy g'oyasi shundan iboratki, jumlagi kiritilgan har bir so'z uchun quyidagi funktsiyani maksimallashtiradigan grammatik sinfni (tegni) tanlash kerak:

$$P(\text{word}|\text{tag}) * P(\text{tag}|\text{previous } n \text{ tags})$$

$P(\text{tag}|\text{previous } n \text{ tags})$ – shartli ehtimollik, ushbu teg oldingi n teg aniqlangandan keyin hosil bo'ladi.

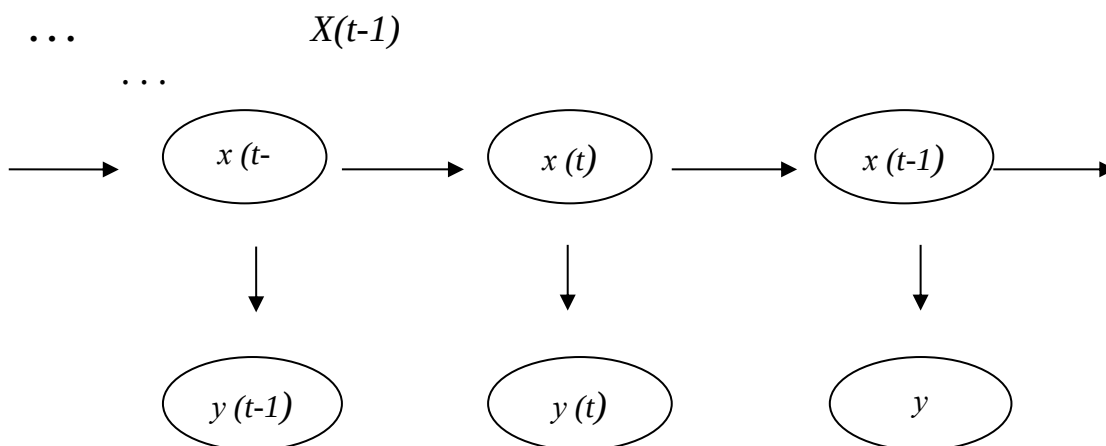
$P(\text{word}|\text{tag})$ – shartli ehtimollik, berilgan joyda hosil bo'lgan *word* so'zi, berilgan grammatik sinf tegiga ega.

Yashirin Markov modeli – noma'lum parametrlarga ega bo'lgan jarayonning ishlashini simulyatsiya qiluvchi statistik model. Noma'lum parametrlarni o'zgaruvchan kuzatuvlar asosida hal qilish vazifa qilib qo'yiladi¹⁸.

¹⁷ Абжалова М.А. Ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг лингвистик модуллари (Расмий ва илмий услубдаги матнлар таҳрири дастури учун). Филол. фан. бўйича фалсафа доктори (PhD) диссер. – Фарғона, 2019. – 164 б.

¹⁸ Ветров Д. Скрытые марковские модели //

http://www.machinelearning.ru/wikiimages/8/83/GM12_3.pdf.



3.1-rasm. YMMning umumiy strukturasi.

Bunda $x(t)$ yashirin o'zgaruvchining qiymati faqat $x(t-1)$ yashirin o'zgaruvchining qiymatiga bog'liq ekanligi ravshan bo'ladi. Bu bog'liklik (tobelik) “Markov xususiyati” deb ataladi. Shu bilan birga, kuzatilgan o'zgaruvchining qiymati $y(t)$ faqat yashirin o'zgaruvchining $x(t)$ qiymatiga bog'liq bo'ladi. Lug'atda bo'lmagan so'zlarni tahlil qilish uchun uch harfli postfikslar ishlatiladi. Natijada, morfologik xususiyatlarni aniqlash muammosini hal qilish yashirin qiymatlarning eng munosib qiymatlari ketma-ketligini hisoblash uchun kamayadi:

$$u'(t) = \operatorname{argmax} P(y|x)$$

YMM algoritmi juda yuqori hisoblash murakkabligiga ega (klassik Viterbi dinamik dasturlash usuli bilan amalga oshiriladi, shunga o'xshash Levenshteyn masofani hisoblash uchun ishlatiladi) va amalda, odatda, uni kamaytirish uchun turli soddalashtiruvchi taxminlar ishlatiladi (masalan, uch so'zli ketma-ketliklardan ortiq bo'lmashligi kerak). Ingliz tili uchun YaMM algoritmining aniqligi 96%.

Morfologik noaniqlikni bartaraf etishning yana bir yondashuvi deterministik va ehtimoliy taggerlarning eng yaxshi xususiyatlarini birlashtiradigan yondashuvdir. Bunday turdagi eng mashhur yondashuvlardan biri – TBL (Transformation-Based Learning) tagger yoki Brillning morfologik analizatori (Brill tagger)¹⁹.

Bu usul mashinasozlik g'oyalariga asoslangan bo'lib, unda grammatik sinflarni so'zlarga tayinlash qoidalari oldindan belgilangan korpusdan avtomatik ravishda chiqariladi va keyin deterministik algoritmlar qayta ishlangan matnga qo'llaniladi.

Ushbu algoritmnı o'rgatishda, avvalo, korpuslarning belgilangan barcha so'zlariga eng ehtimoliy teg qiymatlari beriladi. Keyin barcha mumkin bo'lgan teglar (ishlab chiqarish qoidalari) o'rganiladi; (allaqachon belgilangan korpus nuqtai

¹⁹ Cristopher D. Manning, Hinrich Schutze. Foundation of Statistical Natural Language Processing, 1999.

nazaridan) tahlil sifatini maksimal yangilaydiganlari qoldiriladi. O'qitish natijasida yangi holatlarni qayta ishlashda yana qo'llash mumkin bo'lgan qoidalarning tartiblangan ro'yxati (ajralish tartibi) yaratiladi²⁰.

Ko'rinadiki, ingliz tilida matnlarda omonimiyani bartaraf etish masalasida ehtimoliy metodlardan yuqori darajada foydalanilgan. Ingliz tilida gapda so'zlar ketma-ketligi aniq bo'lib, mazkur usullarni qo'llash soddalashtirilgan. Tadqiqotlardan kuzatiladiki, ingliz tili matnlarida omonimiyani bartaraf etish metodlaridan Yashirin Markov modeli va tayanch vektor metodi yetakchilik qiladi.

Sh. Gulyamova dissertatsiyasida omonimlarni turli so'z turkumlari doirasidagi guruhlariga ajratgan. Ularni ikki so'z turkumi doirasidagi omonim so'zlarni farqlashda “Qoidalarga asoslangan teglash usuli”dan foydalanishni taklif qilgan²¹. Ushbu teglash usuli aniq qoidalarga tayanadi. Ya'ni ikkita so'z turkumi doirasidagi omonim so'zlarning ma'nolarini farqlash ma'lum qoidalar orqali amalga oshiriladi.

Ingliz olimi Markov bir turkum doirasidagi omonimlikni aniqlashda so'z turkumlarini teglashtirish va ehtimollar nazariyasidan foydalanishni taklif qilgan. Bizning tadqiqotimizda o'zbek tilidagi fe'l so'z turkumi doirasidagi omonimlarni farqlashda mana shu metodlardan o'rinli va maqsadli foydalanildi.

O'zbek tilining xalqimiz ijtimoiy hayotida va xalqaro miqyosdagi obro'-e'tiborini tubdan oshirish, unib-o'sib kelayotgan yoshlarimizni vatanparvarlik, milliy an'ana va qadriyatlarga sadoqat ruhida tarbiyalash, mamlakatimizda davlat tilini to'laqonli joriy etishni ta'minlash, O'zbekistondagi millat va elatlarning tillarini saqlash va rivojlantirish, davlat tili sifatida o'zbek tilini o'rganish uchun shart-sharoitlar yaratish, o'zbek tili va til siyosatini rivojlantirishning strategik maqsadlari, ustuvor yo'nalish va vazifalarini hamda istiqboldagi bosqichlarini belgilash maqsadida²² bir qator chora-tadbirlar belgilanishi ona tilimiz mavqeini yanada ko'tarishga, uning amaliy qo'llanishini kengaytirishga qaratilgan.

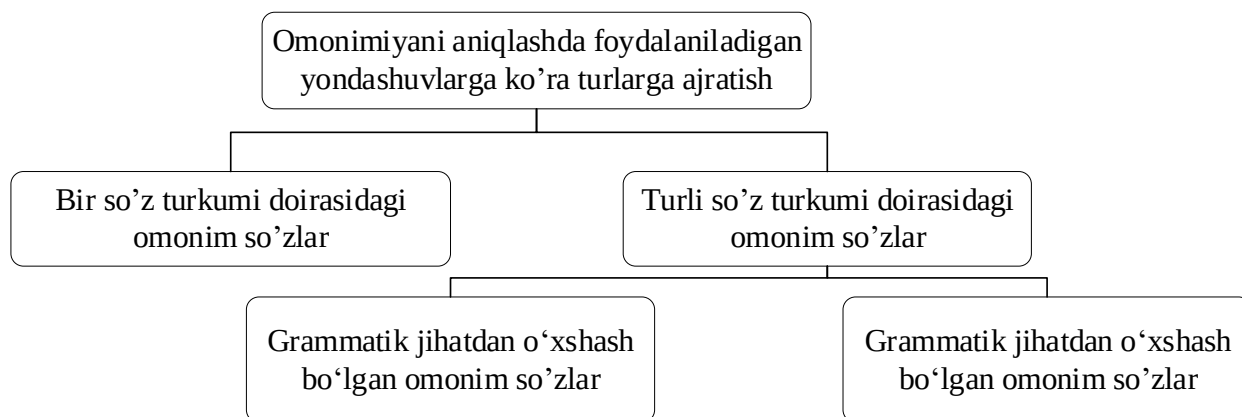
Gulyamova tomonidan taklif etilgan lingvistik ta'minot bo'yicha tadqiqotchi X.Axmedova o'zining “O'zbek tilidagi gaplarni semantic analiz qilishning modeli, algoritmlari va axborot tizimini ishlab chiqish” deb nomlangan PhD dissertatsiyasi avtoreferatida omonim so'zlarni semantik farqlashga alohida

²⁰ Зеленков Ю.Г., Сегалович И.В., Титов В.А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // http://www.dialog-21.ru/media/2444/zelenkov_segalovich.pdf

²¹ Gulyamova Sh. O'zbek tili semantic analizatori ishlab chiqishning lingvistik ta'minoti. Filol.fan.doktori (DcS) disser. –Farg'ona, 2022.

²² O'zbekiston Respublikasi Prezidenti Shavkat Mirziyoyevning “Mamlakatimizda o'zbek tilini yanada rivojlantirish va til siyosatini takomillashtirish chora-tadbirlari to'g'risida”gi PF-60-84-son farmoni // <https://lex.uz/docs/-5058351>

to'xtalgan²³. Olima omonim so'zlarni semantik farqlashda ularni so'z turkumlari doiraga ajratgan holda turlicha yondashuvlardan foydalangan. Ya'ni bir so'z turkumi, turli so'z turkumlari doirasidagi omonim so'zlarni farqlash uchun turlicha algoritmlarni taklif qilgan.



3.2-rasm: Omonim so'zlarni ularni farqlovchi yondashuvlarga ko'ra guruhleri

X.Axmedova bir so'z turkumi doirasidagi omonim so'zlarni farqlashda, ularni birikuvchilari yordamida farqlash mumkinligi ta'kidlab, YMM modelini taklif qilgan. Turli so'z tukumlari doirasidagi omonim so'zlarni ham o'z navbatida ikki guruhga ajratgan: grammatik jihatdan o'xshash va o'xshash bo'lmagan so'z turkumlari doirasidagi omonimlarni alohida ajratgan. X.Axmedova grammatik jihatdan o'xshasha bo'lmagan omonim so'zlarni tilning grammatik qoidalari asosida farqlovchi matematik modellar ishlab chiqqan. Grammatik jihatdan o'xshash bo'lgan so'z turkumlari doirasidagi omonim so'zlarni farqlovchi qat'iy qoida mavjud emas deya baholaydi va bu turdagi omonimlarni farqlashda statistik usullardan foydalanishni taklif qilgan.

Xulosa. Omonimlarni aniqlash va ularni to'g'ri tahlil qilish tilshunoslik va NLP sohasida katta ahamiyatga ega. Ingliz va rus tillarida bu masala uchun samarali metodlar va algoritmlar ishlab chiqilgan bo'lsa-da, o'zbek tilida bu soha hali ham rivojlanish jarayonida. O'zbek tilida omonimlarni aniqlash va farqlash uchun ilg'or statistik metodlar, morfologik tahlil usullari, va gibrid yondashuvlarni qo'llash kerak. Ushbu metodlar tilshunoslik va kompyuter fanlarining integratsiyasini ta'minlaydi va tilning yangi imkoniyatlarini ochadi.

²³ Axmedova X. O'zbek tilidagi gaplarning semantik analiz qilishning modeli, algoritmlari va axborot tizimini ishlab chiqish// texnika fanlari falsafa doktori (PhD) disser. – Toshkent, 2023. 120-b.

Omonimlarni samarali bartaraf etish uchun ko'p jihatdan texnologik yondashuvlarning va ilmiy izlanishlarning davom ettirilishi muhim. Shuningdek, o'zbek tilining o'ziga xos xususiyatlarini hisobga olish orqali, yangi va innovatsion metodlar ishlab chiqilishi lozim. Bu jarayonda tilshunoslar, kompyuter olimlari va dasturchilarning hamkorligi tilshunoslik va NLP sohasida katta yutuqlarga olib kelishi mumkin. Omonimiyalarni to'g'ri tahlil qilish, tilni avtomatik qayta ishlashda yanada aniq va samarali natijalarga erishish imkonini beradi.

Foydalanilgan adabiyotlar

1. Hamroyeva Sh. O'zbek tili morfologik analizatorining lingvistik. – GlobeEdit, 2020;
2. Гулямова Ш. Ўзбек тили семантик анализаторининг лингвистик асослари: Филол.фан.док-ри...дисс... – Фарғона, 2022. – 215 б.;
3. Hamroyeva Sh. O'zbek tili mualliflik korpusini tuzishning lingvistik asoslari: Filol.fan. bo'yicha falsafa doktori (PhD)...diss. – Qarshi, 2018.–250 b;
4. Abjalova M.A. O'zbek tilidagi matnlarni tahrir va tahlil qiluvchi dasturning lingvistik modullari (Rasmiy va ilmiy uslubdagi matnlar tahriri dasturi uchun): Filol.fan.bo'yicha falsafa doktori (PhD)...diss. – Farg'ona, 2019. – 164b;
5. Axmedova D. Atov birliklarini o'zbek tili korpuslari uchun leksik- semzoantik teglashning lingvistik asos va modellari.Filologiya fanlari bo'yicha fal.dok-ri.diss.- Buxoro, 2020. – 163 b.;
6. Bakayev I.I. «O'zbek tili so'z shakllarini morfologik tahlil qilish modellari va algoritmlari»: texnika fan. fan. bo'yicha falsafa doktori (PhD)...diss.avtoref. – Toshkent, 2021. –12-22 b.;
7. X.Axmedova. O'zbek tilidagi gaplarni semantik analiz qilishning modeli, algoritmlari va axborot tizimini ishlab chiqish// texnika fanlari falsafa doktori (PhD)... diss.avtoref.– Toshkent, 2023.
8. Нурмонов А. Ўзбек тилшунослиги тарихи. – Тошкент: Ўзбекистон, 2002. – 177 б. – Б. 83.