INTERNATIONAL JOURNAL OF SCIENTIFIC RESEARCHERS
ISSN: 3030-332X Impact factor: 8,293 Volume 10, issue 2, March 2025
https://wordlyknowledge.uz/index.php/IJSR worldly knowledge
Index: google scholar, research gate, research bib, zenodo, open aire.
https://scholar.google.com/scholar?hl=ru&as_sdt=0%2C5&q=wosjournals.com&btnG
https://www.researchgate.net/profile/Worldly-Knowledge
https://journalseeker.researchbib.com/view/issn/3030-332X

Firuza Fozilova
Termiz Davlat universiteti,
Lingvistika (o'zbek tili) yo'nalishi, 1- kurs magistanti

O'ZBEK TILI KORPUSIDA SO'Z MA'NOLARI IZOHIGA DOIR MA'UMOTLAR BAZASINI SHAKLLANTIRISH

Annotatsiya: O'zbek tili korpusida so'z ma'nolari izohiga doir ma'lumotlar bazasini shakllantirish masalasi tilshunoslik va kompyuter lingvistikasi sohalarida dolzarb hisoblanadi. Maqola o'zbek tili korpusida so'zlarning ma'nosini talqin qilish bo'yicha ma'lumotlar bazasini shakllantirish jarayonini tavsiflashga bag'ishlangan. Bunday leksik ma'lumotlar bazalarini yaratishning asosiy prinsiplari, ish bosqichlari, ishlatilgan tahlil vositalari ko'rib chiqiladi. Konteksga qarab ko'p ma'noli so'zlarning turli semantik ma'nolarini aniqlashga imkon beradigan lingvistik usullarga alohida e'tibor beriladi. Hozirgi vaqtda tashkil etilgan o'zbek tili ma'lumotlar bazasining miqdoriy ko'rsatkichlari keltirilgan. Leksikografik, tarjima va ta'lim ishlarini olib borish uchun bunday bazalarning amaliy ahamiyati to'g'risida xulosa chiqariladi.

Kalit so'zlar: korpus lingvistikasi, ma'lumotlar bazasi, semantika, o'zbek tili, so'z ma'nolari izohiga oid ma'lumotlar bazasining amaliy qo'llanilishi.

Аннотация: Вопрос формирования базы данных объяснений значений слов в корпусе узбекского языка считается актуальным в области лингвистики и компьютерной лингвистики. Статья посвящена описанию процесса формирования базы данных по интерпретации значений слов в корпусе узбекского языка. Рассмотрены основные принципы построения подобных лексических баз данных, этапы работы, используемые инструменты анализа. Отдельное внимание уделено лингвистическим методам, позволяющим выявить различные семантические значения многозначных слов в зависимости от контекста. Приведены количественные показатели созданной на текущий момент базы данных узбекского языка. Сделан вывод о практической значимости подобных баз для проведения лексикографических, переводческих и образовательных работ.

Ключевые слова: корпусная лингвистика, база данных, семантика, узбекский язык, практическое применение лексико-семантической базы данных

Abstract: The issue of forming a database of word meaning explanations in the Uzbek language corpus is considered relevant in the fields of linguistics and computational linguistics. The article is devoted to the description of the process of forming a database on the interpretation of the meanings of words in the corpus of the Uzbek language. The basic principles of building such lexical databases, the stages of work, and the analysis tools used are considered. Special attention

is paid to linguistic methods that make it possible to identify different semantic meanings of polysemous words depending on the context. Quantitative indicators of the currently created database of the Uzbek language are given. The conclusion is made about the practical importance of such databases for carrying out lexicographic, translation and educational work.

Keywords:corpus linguistics, database, semantics, Uzbek language, the practical application of a Lexical Semantic Database

KIRISH

Matn korpuslari va tegishli leksik ma'lumotlar bazalarini yaratish hozirgi vaqtda kompyuter lingvistikasi va leksikografiyasining asosiy yo'nalishlaridan biri hisoblanadi. Matn korpuslari va tegishli leksik ma'lumotlar bazalari haqiqiy matnlarda ma'lum bir tilning leksik birliklaridan foydalanish to'g'risidagi ma'lumotlarni to'plash, tizimlashtirish va tahlil qilish imkonini beradi, bu esa katta ilmiy va amaliy ahamiyatga ega.[1] Bunday ma'lumotlar bazalarining yaratilishi lingvistik tadqiqotlar, terminologiya bazalarini shakllantirish, sun'iy intellekt tizimlari uchun til modellari yaratish hamda kompyuter lingvistikasi doirasidagi boshqa muhim ishlanmalar uchun mustahkam asos bo'lib xizmat qiladi. Shuningdek, ular korpus lingvistikasi metodlari yordamida til birliklarining qo'llanilish chastotasi, uslubiy xususiyatlari va semantik jihatlarini chuqur o'rganish imkoniyatini beradi.

Turkiy tillar, xususan o'zbek tillari uchun to'laqonli lingvistik ma'lumotlar bazalarini yaratish bo'yicha ishlar hali boshlang'ich bosqichda. Shu bilan birga, bunday resurslardan foydalanish potensiali tabiiy tilni qayta ishlash, mashina tarjimasi, tilni o'qitish va boshqalar uchun juda yuqori. O'zbek tili uchun sifatli va keng qamrovli matn korpuslari va leksik ma'lumotlar bazalarining ishlab chiqilishi til texnologiyalarini rivojlantirishga katta hissa qo'shadi hamda o'zbek tilining raqamli muhitda yanada keng qo'llanilishiga imkon yaratadi.

USULLAR VA ADABIYOTLAR TAHLILI

Tadqiqotda analiz, sintez, induksiya, deduksiya, qiyosiy tahlil va ilmiy bilishning boshqa usullaridan foydalanildi.

Korpus lingvistikasi – bu tilni matn korpusi orqali o'rganishning empirik usuli.[2]

Xorijiy va mahalliy tilshunoslikda turli tillar uchun korpuslar va leksik ma'lumotlar bazalarini yaratishda ma'lum tajriba to'plangan. Xususan, shunga o'xshash manbalar **ingliz**[3]**, nemis, fransuz**[4]**, yapon**[5] va boshqalar kabi tillar uchun ishlab chiqilgan. **Ingliz tili** uchun Brown Corpus,

---

[1] Захаров В.П., Хохлачева С.Н. Корпусная лингвистика: Учебно-методическое пособие. СПб.: С.-Петерб. гос. ун-т, 2013. 48 с.

[2] Meyer, Charles F.. English Corpus Linguistics, 2nd, Cambridge: Cambridge University Press, 2023 — 4-bet

[3] Davies M. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights // International journal of corpus linguistics. 2009. Т. 14. No. 2. С. 159-190.

[4] Jouis C. Contributions françaises au développement de la linguistique de corpus // La Linguistique de corpus / P. Corbin, ed. - Paris : Armand Colin, 2000. p. 71-91.

[5] Котов А.А. Создание и использование корпусов японского языка // Вестник РГГУ. Серия «Языкознание». 2018. No4. С. 94-106.

British National Corpus, COCA kabi yirik resurslar mavjud bo‘lib, ular lingvistik tadqiqotlar, avtomatik tarjima va tabiiy tilni qayta ishlash jarayonlarida keng foydalaniladi. **Nemis** va **fransuz tillari** uchun ham shunga o‘xshash korpuslar ishlab chiqilgan bo‘lib, ularning asosiy xususiyatlari – katta hajm, annotatsiya qilingan leksik birliklar va zamonaviy matnlar bilan doimiy ravishda yangilanib borishidir. **Yapon tilida** esa BCCWJ (Balanced Corpus of Contemporary Written Japanese) kabi muhim korpuslar mavjud bo‘lib, ularning tarkibi turli janrdagi matnlardan tashkil topgan. Ushbu tajribalar shuni ko‘rsatadiki, zamonaviy tilshunoslikda korpus va leksik bazalar faqat lug‘at shaklida emas, balki keng qamrovli, kontekstual ma’lumotlarni o‘z ichiga olgan tizimli resurslar sifatida yaratilishi kerak.

O‘zbek tili uchun korpuslar va lug‘atlar yaratishga ham urinishlar qilingan. Masalan, **A.Gatiatullin, N. Abduraxmonova**[6], **N. Mahmudov** kabilar asarlarida semantik chastotalarning o‘zbek-turk va o‘zbek-rus lug‘ati tasvirlangan. Ushbu lug‘atlar asosan ikkita til orasidagi semantik bog‘liqliklarni ochib berishga qaratilgan bo‘lsa-da, ular zamonaviy korpus lingvistikasi talablariga to‘liq javob bermaydi. Ayniqsa, haqiqiy matnlardagi leksik birliklarning qo‘llanilish chastotasi, ularning sintaktik va semantik xususiyatlarini o‘z ichiga olgan ma’lumotlar bazasi hali mavjud emas.

Biroq, haqiqiy matnlarda leksik birliklardan foydalanish statistikasini va ularning semantikasi to‘g‘risidagi ma’lumotlarni to‘playdigan kompleks resurs hali mavjud emas. Bunday resursning yo‘qligi o‘zbek tilining tabiiy tilni qayta ishlash (NLP), avtomatik tarjima, so‘z birikmalarining semantik tahlili kabi yo‘nalishlardagi rivojlanishiga to‘sqinlik qiladi. Masalan, Google Translate va boshqa tarjima tizimlari o‘zbek tilini yaxshi qo‘llab-quvvatlamaydi, chunki bu til uchun katta hajmdagi tekshirilgan korpuslar va ma’lumotlar bazasi mavjud emas. Shu sababli, ushbu tadqiqot doirasida o‘zbek tilining lingvistik xususiyatlarini hisobga olgan holda, aniq va tizimli ma’lumotlar bazasini yaratish asosiy vazifalardan biri sifatida belgilandi. Bu esa ushbu tadqiqotning yangiligi va amaliy ahamiyatini belgilaydi.

MUHOKAMA

Lug‘atning muhim xususiyatlaridan biri bu so‘zlarning polisemiyadir. Ishlab chiqilgan ma’lumotlar bazasidan foydalanib, tez-tez uchraydigan otlar misolida o‘zbek lug‘atining polisemantik tahlili o‘tkazildi.

Ko‘rib chiqilayotgan ismlarning taxminan uchdan ikki qismi 2 tadan 4 tagacha ma’noga ega ekanligi aniqlandi. Buning sababi shundaki, ushbu terminallar o‘zbek tilida faol ishlatiladigan eng muhim, asosiy tushunchalarni ifodalaydi.

Shu bilan birga, yuqori chastotali otlarning 10% dan ortig‘i ko‘p ma’noli - 5 yoki undan ortiq qiymatlarning yuqori darajasini ko‘rsatadi (masalan, “qo‘l”, “bosh”, “yo‘l”). Bu o‘zbek tilining leksik boyligini ko‘rsatadi.

Shunday qilib, leksik-semantik ma’lumotlar bazasidan foydalanish miqdoriy usullardan foydalangan holda o‘zbek leksikasi semantikasining xususiyatlarini chuqur tahlil qilishga imkon beradi.

---

[6] A. Rafizovich, N. Abduraxmonova, “O‘zbek tilining lingvistik ma’lumotlar bazasini shakllantirishda “turkiy morfema” portali instrument sifatida”, DOI: 10.53885/edinres.2021.96.64.071

Birinchi bosqichda kompyuter lingvistikasi usullari yordamida korpusni qayta ishlash, shu jumladan lemmatizatsiya, morfologik va sintaktik tahlil o‘tkazildi.
Bundan tashqari, olingan ma’lumotlar asosida k-means algoritmlari va ierarxik klasterlash usullari yordamida leksik-semantik guruhlar ajratildi.
Shu bilan bir vaqtda tilshunoslarning ko‘p ma’noli so‘zlarning turli ma’nolarini talqin qilish va ma’nolarning avtomatik klasterlash natijalarini tekshirish bo‘yicha ekspert ishlari olib borildi.
Natijalar tuzilgan va ma’lumotlar bazasiga kiritilgan, ularning har bir yozuvida quyidagilar mavjud:

- lemma
- nutqning bir qismi
- qiymatlar / ma’no
- foydalanish misollari
- chastota.

Hozirgi vaqtda ma’lumotlar bazasida o‘zbek tilining 3000 ta eng tez-tez uchraydigan leksik birliklari to‘g‘risidagi ma’lumotlar mavjud bo‘lib, ularda ma’nolar va real kontekstlardan foydalanish misollari ko‘rsatilgan.

So‘z ma’nolari izohiga oid ma’lumotlar bazasining amaliy qo‘llanilishi. So‘z ma’nolari izohiga doir ma’lumotlar bazasi turli sohalarda qo‘llanilishi mumkin. Uning asosiy amaliy qo‘llanilish yo‘nalishlari quyidagilar:

✔ **Tilshunoslik va leksikografiya**

- **Yangi lug‘atlar yaratish** – So‘z ma’nolarining korpus asosida avtomatik tahlil qilinishi natijasida aniq va mukammal lug‘atlar ishlab chiqish mumkin.
- **Sinonim, antonim va homonimlar bazasi** – So‘zlarning ma’no aloqalarini aniqlash orqali tilshunoslik tadqiqotlarini rivojlantirish.
- **Til korpusining boyitilishi** – So‘z ma’nolari izohlangan holda ma’lumotlar bazasi shakllansa, korpus aniqroq va tushunarli bo‘ladi.

✔ **Ta’lim va lingvodidaktika**

❖ **Elektron lug‘atlar va o‘quv resurslari** – O‘quvchilar va talabalarga so‘z ma’nolarini kontekstda tushuntirishga yordam beruvchi elektron vositalarni yaratish.
❖ **Chet tillarini o‘qitishda yordam** – Tarjima dasturlari va o‘qitish platformalarida so‘zning aniq ma’nosini berish orqali til o‘rganish jarayonini yengillashtirish.
❖ **Imtihon tizimlari** – So‘z ma’nolarini avtomatik aniqlash orqali test savollari yaratish va tahlil qilish.

✔ **Tarjima va tabiiy tilni qayta ishlash (NLP)**

- **Mashina tarjimasi** – O‘zbek tilining so‘z boyligini to‘g‘ri tarjima qilish uchun kontekst asosida ma’no izohlash tizimlarini ishlab chiqish.
- **Chatbot va ovozli yordamchilar** – So‘zlarning turli kontekstdagi ma’nolarini tushunish orqali tabiiy tilda muloqot qila oladigan sun’iy intellekt tizimlarini yaratish.
- **Avtomatik matn tahlili** – Ma’lumotlar bazasi yordamida hujjatlar yoki matnlarni to‘g‘ri indeksatsiya qilish va saralash.

ISSN: 3030-332X Impact factor: 8,293 Volume 10, issue 2, March 2025
https://wordlyknowledge.uz/index.php/IJSR worldly knowledge
Index:* *google scholar, research gate, research bib, zenodo, open aire.
https://scholar.google.com/scholar?hl=ru&as_sdt=0%2C5&q=wosjournals.com&btnG
https://www.researchgate.net/profile/Worldly-Knowledge
https://journalseeker.researchbib.com/view/issn/3030-332X

✔ **Adabiyotshunoslik va madaniy tadqiqotlar**

❖ **Matn tahlili** – Klassik va zamonaviy adabiyotdagi so‘zlarning qo‘llanilish xususiyatlarini tahlil qilish.

❖ **Shaxsiy uslubni aniqlash** – Mualliflarning so‘z tanlash uslubi va yozish uslubini o‘rganish.

❖ **Dialektologiya va tarixiy tilshunoslik** – So‘zlarning tarixiy rivojlanishi va dialektal farqlari bo‘yicha tadqiqotlar olib borish.

✔ **Raqamli texnologiyalar va sun’iy intellekt**

- **Tavsiya tizimlari** – Elektron kitob platformalari yoki onlayn ta’lim tizimlarida foydalanuvchiga mos tushadigan tavsiyalarni berish.
- **Ovozli buyruqlarni tushunish** – Nutqni matnga aylantirishda kontekst asosida so‘z ma’nolarini to‘g‘ri aniqlash.
- **Intellektual tahlil tizimlari** – Jurnalistik maqolalar, huquqiy hujjatlar yoki ilmiy ishlarni avtomatik baholash va tahlil qilish.

XULOSA

Tadqiqot o‘zbek tilining leksik-semantik ma’lumotlar bazasini shakllantirish uchun avtomatlashtirilgan va ekspert yondashuvlarini birlashtirish samaradorligini ko‘rsatdi. Unda to‘plangan ma’lumotlar tabiiy tilni qayta ishlash sohasidagi amaliy ishlanmalarda ishlatilishi mumkin, shuningdek, yangi avlodning to‘liq izohli va tarjima qilingan o‘zbek-turk va o‘zbek-rus lug‘atlarini yaratish uchun asos bo‘lib xizmat qilishi mumkin.

REFERCES:

1. Meyer, Charles F.. English Corpus Linguistics, 2nd, Cambridge: Cambridge University Press, 2023 — 4-bet
2. Захаров В.П., Хохлачева С.Н. Корпусная лингвистика: Учебно-методическое пособие. СПб.: С.-Петерб. гос. ун-т, 2013. 48 с.
3. Davies M. The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights // International journal of corpus linguistics. 2009. Т. 14. No. 2. С. 159-190.
4. Jouis C. Contributions françaises au développement de la linguistique de corpus // La Linguistique de corpus / P. Corbin, ed. - Paris : Armand Colin, 2000. p. 71-91.
5. Котов А.А. Создание и использование корпусов японского языка // Вестник РГГУ. Серия «Языкознание». 2018. No4. С. 94-106.
6. Махмудов Н., Менглиев Б. Ўзбек тилининг семантик частотали луғати. Ўзбек тилининг изоҳли луғати. Ж.1, Тошкент-Бишкек, 2017. 272 б.
7. A. Rafizovich, N. Abduraxmonova, “O‘zbek tilining lingvistik ma’lumotlar bazasini shakllantirishda “turkiy morfema” portali instrument sifatida”, DOI: 10.53885/edinres.2021.96.64.071