“O‘ZBEK BOSHLANG‘ICH SINF MAKTAB KORPUSI”DAN BAZIS SO‘ZLARNI AVTOMATIK AJRATIB OLISH

Madatov Xabibulla Axmedovich,
Fizika-matematika fanlari nomzodi, dotsent,
habi1972@mail.ru
Urganch davlat Universiteti

Xajibaeva Surayyo Maxmudjonovna
o‘qituvchi
surayyo.khajiboyeva@gmail.com
Urganch davlat Universiteti

Xujamov Elyor Jumanazarovich
UrDU talabasi

Annotatsiya – bazis so‘zlarni ajratib olish masalasi til o‘rganuvchilar, maktab o‘quvchilari, shuningdek, lingvistlar uchun ham bir qator qulayliklar yaratadi. Ushbu maqolada “O‘zbek boshlang‘ich sinf maktab korpus”idan bazis so‘zlarni ajratib olish masalasi ko‘rib chiqilgan. Bazis so‘zlar to‘plami, shunday to‘plamki, ushbu to‘plam yordamida o‘zbek tilidagi barcha so‘zlarni ifodalash mumkin bo‘lsin. Ya’ni, bazis so‘zlar o‘zbek tilining asosini tashkil etuvchi so‘z yasalishi va lingvistik tahlil jarayonlarida ishlatiladigan so‘zlardir. Odatda, bunday so‘zlarga kundalik muloqotning asosini tashkil etuvchi umumiy otlar, fe’llar, sifatlar va boshqa nutq qismlari kiradi. Bundan tashqari, bazis so‘zlar asosan ilmiy va badiiy adabiyotlarda qo‘llaniladi. Ushbu ishda yuqori chastotali so‘zlarni aniqlash va solishtirish usullari qo‘llanilgan. Biz o‘zbek tili sinonimlarining izohli lug‘ati va “O‘zbek boshlang‘ich sinf maktab korpus”i asosida yaratilgan yuqori chastotali so‘zlarni aniqlash va bazis so‘zlarni ajratib olish usulini taqdim etamiz.

Abstract – the issue of extracting basis words creates a number of conveniences for language learners, schoolchildren, as well as linguists. This article presents a set of basis words, such a set that can be used to represent all words in the Uzbek language. That is, basis words are words used in the processes of word formation and linguistic analysis that form the basis of the Uzbek language. Typically, such words include common nouns, verbs, adjectives and other parts of speech that form the basis of everyday communication. In addition, basis words are used mainly in scientific and literary literature. In this work, high-frequency method and comparison method were used. We present a method for identifying high-frequency words and extracting basis words, created on the basis of the Explanatory Dictionary of Uzbek Synonyms and Uzbek Primary School Corpus.

Аннотация – вопрос извлечения базовых слов создает ряд удобств для изучающих язык, школьников, а также лингвистов. В данной статье

представлен набор базовых слов, такой набор, с помощью которого можно выразить все слова узбекского языка. То есть базовые слова — это слова, используемые в процессах словообразования и лингвистического анализа, составляющие основу узбекского языка. Обычно к таким словам относятся нарицательные существительные, глаголы, прилагательные и другие части речи, составляющие основу повседневного общения. Кроме того, основные слова в основном используются в научной и художественной литературе. В данной работе используются методы высокочастотной идентификации и сравнения. Мы представляем метод выявления высокочастотных слов и извлечения базовых слов, созданный на основе Толкового словаря узбекских синонимов и Корпуса узбекского языка для начальной школы.

Kalit so‘zlar – *bazis so‘z, sinonimlar lug‘ati, maktab korpusi, yuqori chastotalar metodi, agglyutinativ til.*

I. KIRISH

O‘zbek tili go‘zal va juda boy til. Bugungi kunda, o‘zbek tilini axborot kommunikatsion texnologiyalardan foydalanib, turli xil online izohli lug‘atlar yaratish, tarjima masalalari, umuman, olganda tabiiy tilni qayta ishlash masalalari ustida turli xil ishlar olib borilmoqda. Hissiyotlarni tahlil qilish, nomuhim so‘zlarni topish, matnlarni tasniflash, lemmatizatsiya masalasi, niqoblangan tilni modellashtirish (MLM) kabi jarayonlar o‘zbek tilini tahlil qilishning masalalari hisoblanadi.

“Hozirgi kunda mamlakatimizda yashayotgan, O‘zbekistonimizni yagona va ahil oila deb biladigan turli millat va elatlar vakillari ham o‘zbek tilini o‘rganishga katta qiziqish va istak bildirayotganlari ayniqsa e’tiborlidir”, – deya ta’kidlaydi davlatimiz rahbari Shavkat Mirziyoyev. Ta’kidlanishicha, bugungi kunda jahon miqyosida 50 millionga yaqin kishi o‘zbek tilida so‘zlashadi. Ko‘plab mamlakatlarda bu go‘zal va jozibali til katta ishtiyoq bilan o‘rganilmoqda. Bundan tashqari, Prezidentimizning 2020-2030 yillarda o‘zbek tilini rivojlantirish va til siyosatini takomillashtirish konseptsiyasida xorijliklar uchun o‘zbek tilini o‘rgatuvchi dasturlar yaratish asosiy yo‘nalishlardan biri sifatida ta’kidlangan. Shunga asoslanib, o‘zbek tilini o‘rganuvchi mustaqil o‘rganuvchilar uchun yengillik yaratish bizning asosiy maqsadlarimizdan biridir.

O‘zbek tilini o‘rganuvchilar uchun tilni darajalarga bo‘lish asosida o‘rganish g‘oyasini ilgari sursak, har bitta darajada o‘rganuvchi bilishi zarur bo‘lgan so‘zlar lug‘ati mavjud bo‘lishi zarur. Ya’ni, o‘rganuvchi birinchi darajadagi aytaylik sodda so‘zlarni yod olsa va o‘rgansa, keyin o‘ziga mos va qiziqarli adabiyotni tanlay va o‘qiy oladi, shu asosida matn tuzadi, nutqida bayon qiladi, eshitsa anglaydi va javob qaytaradi. Masalan, 2-sinf o‘quvchisi uchun “kitob” so‘ziga izoh berish zarur bo‘lsa, unga “varaqlardan iborat o‘quv vositasi; darslik, ertak kabi turlari mavjud” deb

ta’riflanadi. Undan katta sinflarda esa “ma’lum matnli varaqlardan iborat, juzlab tikilgan, muqovalangan, hajmi 48 sahifadan kam bo‘lmagan bosma (qadim qo‘lyozma ham) asar” deya ta’rif beriladi. Shunday qilib, bosqichma-bosqich yangi va xilma-xil so‘zlar bilan uning lug‘at boyligi oshib boradi. Ammo, birdaninga o‘z darajasiga mos bo‘lmagan so‘zlarni yod oldirish yoki ulardan foydalanish ko‘nikmasini shakllantirishga harakat qilinsa, o‘rganuvchi ushbu tilni o‘rganishdan zerikadi, so‘zlardan foydanish ham qiyin kechadi. Natijada, tilni yaxshi o‘rgana olmaslik holati kelib chiqadi. Maktab o‘quvchilarida esa, adabiyotga nisbatan qiziqish pasayadi. Yuqoridagi fikrlardan kelib chiqib, o‘zbek tilida bazis so‘zlarni topishning ahamiyatli tomoni shundaki, o‘zbek tili darajalari asosida tayyorgan lug‘at bazis so‘zlarga asoslangan holda bo‘lishi zarur. Ya’ni, ushbu lug‘atni yaratishni bazis so‘zlarni kiritishdan boshlash kerak.

Yana bitta ahamiyatli tarafi shundaki, bazis so‘zlar WordNet asoslarini qurish uchun zarur bo‘ladi. WordNet so‘zlarni semantik munosabatlarga, shu jumladan sinonimlar, giponimlar va meronimlar bilan bog‘laydigan so‘zlar orasidagi semantik munosabatlarning leksik ma’lumotlar ombori. Ya’ni, har bitta sinonimlar, giponimlar va meronimlar asosini bazis so‘zlar tashkil qiladi.

II. BOG‘LIQ ISHLAR

Ushbu maqola [1] ta’lim sifatini oshirish uchun muhim bo‘lgan boshlang‘ich sinf o‘quvchilarining intellektual salohiyatiga moslashtirilgan ta’lim korpusini yaratishni o‘rganadi. Darsliklarning o‘quvchilarning intellektual qobiliyatlari bilan mos kelmasligi tushunishni qiyinlashtiradi va o‘rganishga bo‘lgan qiziqishni kamaytiradi. Talabalarning intellektual salohiyatini baholash uchun korpusdan foydalangan holda, tegishli o‘quv materiallarini o‘quv jarayoniga integratsiyalash, faollikni oshirish mumkin. Tadqiqotda ushbu muammoni samarali hal etish maqsadida O‘zbekiston Respublikasi maktabgacha va maktab ta’limi vazirligi tomonidan tasdiqlangan 35 ta o‘zbek tilidagi darsliklardan ishlab chiqilgan korpusga e’tibor qaratilgan. Nomuhim so‘zlarni filtrlash matn so‘rovlarini qayta ishlashda hal qiluvchi protsedura bo‘lib, ular keng ma’lumotlar to‘plamlari ichida ma’lumotlarni qidirish uchun ishlatiladi. Bu jarayon semantik ma’noni yo‘qotmasdan qidiruv maydonini qisqartirish imkonini beradi. Faqat grammatik rolga ega bo‘lgan va so‘rovning axborot mazmuniga hissa qo‘shmaydigan nomuhim so‘zlar, shunga qaramay, so‘rovning umumiy murakkabligiga hissa qo‘shadi. Ushbu muammoni hal qilish uchun ishlatiladigan mavjud matematik modellar tabiiy tillarning barcha toifalariga taalluqli emas[2]. Tabiiy tilni qayta ishlashda ma’lumotlarni yig‘ish va matnni tahlil qilish jarayonida nomuhim so‘zlarini bilish kerak. Bu usul agglyutinativ xususiyatga ega bo‘lgan tillarga nisbatan qo‘llaniladi. O‘zbek tili ham agglyutinativ til hisoblanadi. Bu ish[3] o‘zbek matnlaridagi nomuhim so‘zlarni yoki ko‘pchilik nomuhim so‘zlarni baholash usulini aniqlashga bag‘ishlangan. Axborot

almashuvi va axborotlar tez sur’atlar bilan o‘sib borayotgan bugungi davrda biz turli sohalarga oid barcha ma’lumotlarni o‘qish imkoniyatiga ega emasmiz. Muayyan sohaga oid ma’lumotlarni o‘qish va tahlil qilish ham ko‘p vaqt talab etadi. Ushbu muammoni hal qilish uchun maqolada matnni umumlashtirish taklif etiladi [4]. Unda o‘zbek tili uchun tajriba sifatida umumlashtirish topshirig‘i mavjud bo‘lib, metodologiya TF-IDF algoritmi asosida matnni abstraktlashtirishga asoslangan. O‘zbek tili bilan bog‘liq ishlar va manbalar nuqtai nazaridan shuni e’tirof etish kerakki, uning tabiiy tilni qayta ishlash (NLP) nuqtai nazaridan kam resursli til maqomiga ega bo‘lishiga qaramay, so‘nggi paytlarda ilmiy-tadqiqot ishlari jadal sur’atlarda bormoqda. Bu tilning NLP qamroviga ko‘plab yangi qo‘shimchalar, jumladan transliteratsiya vositalari[5], hissiyotlarni tahlil qilish vositalari[6] va o‘zbekcha matnni umumlashtirish algoritmi[7] qo‘shilishi natijasida paydo bo‘ldi.

O‘zbekiston aholisi orasida 2 millionga yaqin kishi resurslari kam bo‘lgan qoraqalpoq tilida so‘zlashadi, bu esa nomuhim so‘zlarini ajratib olishda turli qiyinchiliklar tug‘dirmoqda. Ushbu muammo ushbu maqolada tasvirlangan ma’lumotlar to‘plami tomonidan hal qilinadi[8]. Shunga qaramay, u o‘zbekcha nomuhim so‘zlari ro‘yxati yordamida sinovdan o‘tkazildi. Mahalliy olimlar o‘zbek tilidagi matnlarning til korpusini yaratish borasida ham muayyan ishlarni amalga oshirdilar. Ushbu maqolada [9] biz o‘zbek tilining sintaktik va morfologik tegli korpusini yaratish uchun “Nutqning bir qismi” (POS) va sintaktik teglar to‘plamini ishlab chiqish orqali bu bo‘shliqni to‘ldirishga harakat qildik [10]. Tabiiy tilni qayta ishlash texnologiyasidan samarali foydalanish kam resursli tilda so‘zlashuvchi aholi uchun juda muhimdir. Aytish joizki, o‘zbek tili uchun Aspect-Based Sentiment Analysis (ABSA) vositalarini yaratish uchun ommaga ochiq, yaxshi o‘rnatilgan lingvistik manbalar mavjud emas. Ushbu ishda yuqorida aytib o‘tilgan bo‘shliqni yopish maqsadida UzABSA, birinchi yuqori sifatli izohli ABSA ma’lumotlar to‘plami taqdim etilgan. O‘zbek tili aglyutinativ til bo‘lganligi uchun qo‘shimchalar soni ko‘p bo‘lib, qo‘shimchalar qo‘shilishi bilan so‘zlar yasaladi. Shu sababli biz so‘zning o‘zagini topishda bir qancha muammolarga duch kelamiz. Ushbu maqolada[11] o‘zbek tilidagi oddiy so‘z shakllarining hech qanday ma’lumotlar bazasini o‘z ichiga olmasdan, affikslarni ajratish usuli bilan o‘zbekcha so‘zlarga o‘zak yasash metodologiyasi taklif qilingan.

III. NAZARIY QISM

Ushbu bo‘limda o‘zbek tili sinonimlarining izohli lug‘ati va O‘zbek boshlang‘ich sinf maktab korpusi(O‘BSMK) yordamida bazis so‘zlarni ajratib olish hamda uning bosqichma-bosqich metodologiyasi haqida ma’lumot berilgan.

A. Ma’lumotlarni tavsiflash va to‘plash

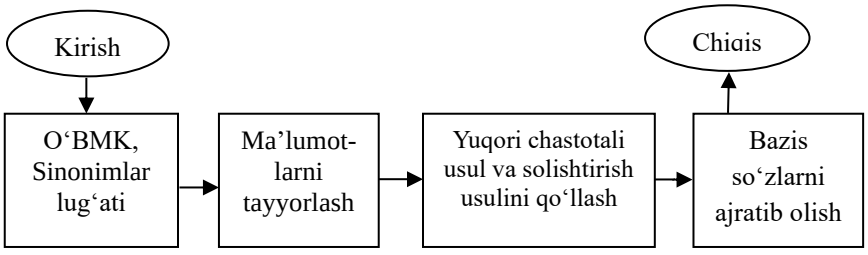
Juda ko‘plab xorijiy tillarda, jumladan, ingliz tilida ham tilni bilish darajasi orqali til o‘rganuvchilar ma’lum bir toifalarga bo‘linadi. Ularning har bir toifasiga qarab so‘zlar bazasi majud bo‘lib, ular yordamida o‘z fikrini, yozishda matn mazmunini to‘liq ifoda qila oladi. Masalan, ingliz tilida ushbu darajalar A1, A2, B1, B2, C1, C2 yoki beginner, pre-intermediate, intermediate, upper-intermediate, advanced, mastery kabilar ko‘rinishida ifodalanadi. O‘zbek tilida tilni bilish darajalari uchun so‘zlar bazasini shakllantirish masalasini ko‘rib chiqdik. Ushbu jarayonni 1-,2-,3-,4- sinf o‘quvchilari uchun mos holatga keltirdik. Ya’ni, birinchi sinf o‘quvchisi nechta so‘z(token)ni bilishi zarur degan savolga javob berdik.

Buning uchun, bizga avvalo asosiy elementlardan biri bo‘lgan maktab korpusi zarur bo‘ldi. Ushbu jarayon ustida O‘zbekiston Respublikasi Maktabgacha va maktab ta’limi vazirligi tomonidan tasdiqlangan 35 ta boshlang‘ich sinf darsliklari qo‘l mehnati asosida tokenlarga ajratildi, turli xil belgilar olib tashlandi va so‘zlar unikal holatga keltirildi. Keyin ularning chastotalari aniqlangan holatda korpusga joylandi.

Yana bitta asosiy elementlardan biri bu sinonimlar lug‘atidir. Biz buning uchun Azim Hojayevning “O‘zbek tili sinonimlar lug‘ati” kitobini oldik, unda sinonimlar bizga kerakli holatda berilgan. Ya’ni, ular sinonimlar guruhlarga ajratilgan shaklda berilgan. Ushbu kitobni pdf formatdan txt formatga o‘tkazdik. Ushbu fayl ustida tozalash ishlari bajarildi, so‘zlar unikal holatga keltirildi.

B. Metod algoritmi

Ushbu ishda biz bazis so‘zlar jadvalini tuzish modelini taklif qildik, u quyidagi bosqichlarni o‘z ichiga oladi: (1) O‘BSMK, Sinonimlar lug‘ati olish; (2) Ma’lumotlarni tayyorlash; (3) Yuqori chastotali usul yordamida yuqori chastotali so‘zlarni topish va solishtirish usuli yordamida korpus uchun so‘zlar guruhini ajratish; (4) Bazis so‘zlar sifatida maksimal chastotali so‘z yoki so‘zlar tanlanadi



Shakl 1. Taklif etilayotgan modelning umumiy ko‘rinishi

C. Ma’lumotlarni tayyorlash.

Har bir sinf uchun bazis so‘zlarning alohida lug‘atini yaratish masalasini ko‘rib chiqildi. O‘zbek tilida O‘BSMK uchun 1,2,3 va 4-sinflar uchun tokenlar

to‘plamini yaratildi. Ushbu korpusda so‘zlar chastotalari bilan keltirildi. Sinonimlar lug‘atida sinonimlar guruhlarga bo‘linadi. Masalan, chiroyli, go‘zal, sohibjamol so‘zlari bir guruhga mansub so‘zlardir. Bir guruhdan so‘z olinib, uni O‘BSMK bilan taqqoslandi. Ushbu lug‘atda a - so‘z bilan boshlangan barcha so‘zlarning chastotalari qo‘shilgan so‘zning chastotasi. Shu tarzda sinonimlar lug‘atidan bir guruhga mansub so‘zlar ichidan eng katta chastotali so‘z olinadi va yangi lug‘at tuziladi. Agar bir nechta maksimal va teng chastotali so‘zlar bir guruhga kirsa, ularning barchasi yangi lug‘atga kiritilgan. Yuqoridagi amallar yordamida bazis so‘zlarni alohida ro‘yxatlar(1-sinf, 2-sinf, 3-sinf, 4-sinf)ga yozib qo‘yamiz.

Olingan natijalar 1-sinf, 2-sinf, 3-sinf va 4 sinflar kesimida [https://zenodo.org/records/13340539] sillkasiga yuklangan.

IV. OLINGAN NATIJALAR

BAZIS SO‘ZLAR SONI

Boshlang‘ich maktab uchun sinflar kesimida bazis so‘zlar soni			
1-sinf	***2-sinf***	***3-sinf***	***4-sinf***
2336	2462	4186	5112

I jadval sinflar kesimida darsliklarda qo‘llanilgan bazis so‘zlarning umumiy sonini ko‘rsatadi.

Yuqoridagi jadvalda keltirilgan natijalarga ko‘ra, 1 sinf maktab o‘quvchisi 2336 ta bazis so‘zlarni bilish orqali gapira oladi, eshitib tushuna oladi, insho yoza oladi va o‘qish ko‘nikmasi shakllanadi. O‘quvchi bu bazis so‘zlarni o‘rganib boshqa so‘zlarni ham ifodalay oladi. Bu 2336 ta bazis so‘zlar O‘BSMKsidagi 1 sinflar uchun 3177 ta lemmalarning ichidan yuqori chastotali so‘zlarni aniqlash orqali kelib chiqqan. 2 sinf uchun 5119 ta lemmalarning ichidan 2462 ta bazis so‘z, 6710 lemmalarning ichidan 4186 ta bazis so‘z, 8029 ta lemmalarning ichidan 5112 ta bazis so‘z mos ravishda ajratib olingan.

V. XULOSA

Xulosa qilib aytganda, har bir sinf maktab o‘quvchilari va o‘zbek tilini o‘rganuvchilar uchun bazis so‘zlar asos bo‘lib xizmat qiladi. Biz bazis so‘zlarni boshlang‘ich sinf maktab darsliklaridagi lemmalar va o‘zbek tili sinonimlar lug‘atidan foydalangan holda yuqori chastotali so‘zlarni topish metodi va solishtirish metodini qo‘llab keltirib chiqardik.

Foydalanilgan adabiyotlar:

1. K. A. Madatov, S. Sattarova, Creation of a Corpus for Determining the Intellectual Potential of Primary School Students, *2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM)*, Altai, Russian Federation, 2024, pp. 2420-2423

2. Madatov, K.; Bekchanov, S.; Vičič, J. Automatic Detection of Stop Words for Texts in the Uzbek Language. Preprints 2022, 2022040234.

3. Madatov, K., Bekchanov, S., & Vičič, J. (2022). Accuracy of the Uzbek Stop Words Detection: a Case Study on “School Corpus” CEUR Workshop Proceedings, 3315, 107–115.

4. K. A. Madatov and S. K. Bekchanov, "The Algorithm of Uzbek Text Summarizer," 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, 2024, pp. 2430-2433, doi: 10.1109/EDM61683.2024.10615191.

5. U. Salaev, E. Kuriyozov, and C. Gómez-Rodríguez, “A Machine Transliteration Tool Between Uzbek Alphabets,” *CEUR Workshop Proceedings*, vol. 3315, pp. 42–50, 2022.

6. S. Matlatipov, H. Rahimboeva, J. Rajabov, and E. Kuriyozov, “Uzbek Sentiment Analysis Based on Local Restaurant Reviews,” CEUR Workshop Proceedings, vol. 3315, pp. 126–136, 2022.

7. K. A. Madatov and S. K. Bekchanov, "The Algorithm of Uzbek Text Summarizer," 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM), Altai, Russian Federation, pp. 2430-2433, 2024.

8. K. Madatov, S. Bekchanov, and J. Vičič, “Dataset of Karakalpak language stop words,” *Data in Brief*, vol. 48, pp. 109111, June 2023.

9. M. Sharipov, J. Mattiev, J. Sobirov, and R. Baltayev, “Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language,” *CEUR Workshop Proceedings*, vol. 3315, pp. 93–98, June 2022.

10. S. Matlatipov, J. Rajabov, E. Kuriyozov, and M. Aripov, “UzABSA: Aspect-Based Sentiment Analysis for the Uzbek Language,” Proceedings of the 3rd Annual Meeting of the ELRA-ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2024) at LREC-COLING 2024, pp. 394–403, 2024.

11. M. Sharipov and O. Yuldashov, “UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language,” CEUR Workshop Proceedings, vol. 3315, pp. 137–144, 2022.