O‘ZBEK TILI TEZAURUSI UCHUN SIFAT SO‘Z TURKUMI LINGVISTIK BAZASINI YARATISH

Po‘latova Gulhayo
Namangan davlat universiteti mustaqil tadqiqotchisi.

Annotatsiya: Jahonda inson faoliyatining bilim talab qiladigan zamonaviy sohalari rivojlanishi kompyuter texnologiyalari rolining oshishi bilan belgilanadi. Bugungi kunda axborot oqimi sezilarli darajada ko‘paymoqda, endi ularni saqlash, taqdim etish, formallashtirish va tartibga solish, shuningdek, avtomatik qayta ishlashning yangi usullarini izlash zarurati yuzaga kelmoqda. Shu bois turli amaliy maqsadlarda qo‘llanilishi mumkin bo‘lgan keng qamrovli bilim bazalariga qiziqish ortib bormoqda. Ayniqsa, inson omilisiz matndan har qanday ma’lumotni chiqarib oluvchi neyroto‘rlarga asoslangan tizimlarga ehtiyoj katta. Ushbu maqolada tezauruslar, sifat so‘z turkumi lingvistik bazasini yaratish bo‘yicha taklif va tavsiyalar berilgan.
Kalit so‘zlar: tezaurus, lingvistik baza, korpus, lug‘at, sifat so‘z turkumi, “oq” leksemasi.

Abstract: The development of modern fields of human activity that require knowledge in the world is determined by the increasing role of computer technologies. Today, the flow of information is increasing significantly, now there is a need to search for new methods of storing, presenting, formalizing and organizing them, as well as automatic processing. Therefore, there is a growing interest in comprehensive knowledge bases that can be used for various practical purposes. In particular, there is a great need for systems based on neural networks that extract any information from the text without the human factor. This article provides suggestions and recommendations for the creation of thesauruses and the linguistic base of adjective vocabulary.
Keywords: thesaurus, linguistic base, corpus, dictionary, adjective phrase, "oq" lexeme.

Аннотация: Развитие современных областей человеческой деятельности, требующих знаний, в мире определяется возрастающей ролью компьютерных технологий. Сегодня потоки информации значительно увеличиваются, теперь возникает необходимость поиска новых методов их хранения, представления, формализации и организации, а также автоматической обработки. Поэтому растет интерес к всеобъемлющим базам знаний, которые можно использовать для различных практических целей. В частности, существует большая потребность в системах на основе нейронных сетей, извлекающих из текста любую информацию без человеческого фактора. В данной статье даны предложения и рекомендации по созданию тезаурусов и лингвистической базы прилагательной лексики.
Ключевые слова: тезаурус, лингвистическая база, корпус, словарь, прилагательное словосочетание, «oq» лексема.

Kirish

Texnologiya asrida taraqqiyot sur’atining tezlashgani, intellektual salohiyat, texnikaning yuqori darajada yuksalishi kishilik jamiyati oldiga qator vazifalarni qo‘ymoqda. Mintaqamizdagi globallashuv jarayoni barcha sohalarda tezkorlikni, jadal taraqqiyotni talab etmoqda. Kompyuter tizimi qulayliklar, imkoniyatlar majmuyiga va tarkibiy qismiga aylandi [Пулатов А.К., Мухамедова С., 2008: 54].

Kishilik taraqqiyotining har bir davrida fan-texnika yangiliklari muhim ahamiyat kasb etib kelgan. Insoniyat uchinchi ming yillikka, axborot asriga qadam qo‘ydi. Inson uchun ma’lumot olish zaruriyatga aylandi. Axborot – ma’lumot oltindan ham qimmatroq deb topildi. Axborotni topish, saqlash, qayta ishlash va boshqalarga yetkazishning qulay usullariga bo‘lgan ehtiyoj kun sayin ortib bormoqda. Bu esa XX asrning buyuk kashfiyoti bo‘lgan kompyuter va kompyuter texnologiyalari sohasi uchun muhim vazifalarni belgilab berdi. Kompyuter texnologiyalari xalq xo‘jaligi, sport, san’at, tibbiyot, umuman, ijtimoiy hayotning barcha tarmoqlariga kirib keldi. Fan yo‘nalishlarini kompyuterlashtirish ilmning asosiy shartlaridan bo‘lgan obyektivlik va aniqlikka amal qilish asosida tez rivojlana boshladi va barcha sohalarda yutuqlarga erishildi [Xolmanova Z., 2020: 9]. Xususan, kompyuter lingvistikasi sohasida ham qator ilmiy izlanishlar olib borildi.

Asosiy qism

O‘zbek tilshunoslari tomonidan o‘zbek tili milliy korpusining yaratilganligi, til ixlosmandlari uchun yangi imkoniyatlar eshigini ochdi.

Korpus (korpus) lotincha “tana” degan ma’noni bildiradi. “Korpus so‘z, so‘z birikmasi, grammatik shakllarni, so‘z ma’nosini muayyan qidiruv tizimi orqali topishni anglatuvchi elektron ko‘rinishdagi matnlar jamlanmasidir” [http://rusorpora.ru]. Korpus tushunchasi bilan yonma-yon “matnlar korpusi” atamasi ishlatilmoqda. Matnlar korpusi elektron holda saqlanadigan fonema, grafema, morfemalar, leksema, gap va matnlardan tashkil topishi mumkin bo‘lgan yaxlit butunlikdir. Korpuslar aslida ma’lumotlar bazasi sifatida shakllantiriladigan, tilshunoslik masalalarini hal etish maqsadida va turli yo‘nalishdagi tadqiqotlarni amalga oshirish uchun material sifatida xizmat qiladigan jamlanmadir.

Jahon tilshunosligida korpusga doir ilk ma’lumotlar XX asrning 40-yillarida qayd etilgan. Korpuslar tarixi haqida so‘z borar ekan, birinchi navbatda, 1961-1964 yillarda yaratilgan Braun korpusi tilga olinadi. Bu korpus Braun universitetida yaratilgan, har biri 2000 so‘zli 500 ta matn fragmentini o‘z ichiga oladi. O‘zbek tilshunosligida A.Po‘latov kompyuter lingvistikasi, tabiiy tilni qayta ishlash, statistik tahlil masalalariga doir izlanishlarda korpus lingvistikasiga ham to‘xtalib o‘tgan. Kompyuter lingvistikasi yo‘nalishlari monografik tadqiqot obyekti sifatida o‘rganila boshlandi. Olimalarimiz Nilufar Abdurahmonova, Manzura Abjalova va boshqalar bu sohada tadqiqotlar olib bormoqdalar. Kompyuter lingvistikasining yo‘nalishi sifatida shakllangan, hozirda o‘z taraqqiyot yo‘nalishiga ega bo‘lgan alohida soha hisoblanuvchi korpus lingvistikasi masalalarini atroflicha o‘rgangan o‘zbek tilshunosi Shaxlo Hamroyeva monografik planda tadqiq etdi. Tezauruslarning ma’lumotlar bazasi sifatidagi qiymatlari yoritilgan tadqiqotlar ham korpuslar haqida muayyan tasavvur berishga xizmat qiladi. Tezauruslarning tuzilishi, ishlash tamoyillari, kompyuter bazasi sifatidagi imkoniyatlari, WordNet tezaurus bazasi haqidagi ma’lumotlar ham ilmiy-amaliy ahamiyatga ega.

Katta hajmdagi ma’lumotlarni izlash va qayta ishlashni nazarda tutuvchi o‘quv – axborot materiallarini yaratish asosi o‘qitishning zamonaviy axborot usullarini rivojlantirish va takomillashtirish bilan bog‘liq. Ushbu usullarni amalga oshirishning muhim va zarur sharti innovatsion loyihalar va ta’lim axborot qidiruv dasturlarini tayyorlash va yaratishdir. Ushbu ilmiy yo‘nalishdagi o‘qitish uslublarining keng miqyosda rivojlanishi innovatsion jarayonlarni o‘rganishga va yangi avtomatlashtirilgan axborot-qidiruv tizimlarini yaratishga olib keladi.

Har qanday tilning ma’lumotlar bazasini yaratishda mazkur tilning to‘liq lug‘atini yaratish muhim rol o‘ynaydi. Hozirgi zamon talablariga ko‘ra, samaradorligi katta bo‘lgan lug‘at

kompyuter lug‘atidir. Kompyuter lug‘ati, an’anaviy lug‘atdan farqli o‘laroq, quyidagi qismlardan iborat bo‘ladi:

1.An’anaviy lug‘atlarga kiritiladigan leksik ma’noga ega, o‘zbek tilida keng iste’molda bo‘lgan so‘z va frazeologik iboralar. Lug‘atga barcha tub so‘zlar (o‘tsimon, kompyuter, yaxshi va h.k.), yasama so‘zlar (badavlat, chiroyli, ishchi va h.k.), qo‘shma so‘zlar (gultojixo‘roz, oshqozon, kungabotar va h.k.), murakkab qisqartma so‘zlar (aeroport, mikroskop va h.k.) kiritiladi.

2.Barcha dialekt (sheva)ga xos so‘zlar.

3.Ayrim sohalarda ishlatiladigan atamalar.

4.Vulgar so‘z va iboralar.

5.Kishi nomlari.

6.Geografik nomlar.

7.Qisqartma so‘zlar (abbreviaturalar) [Po‘latov A., 2011: 213].

Lingvistik bilimlar bazasi tabiiy tildagi matlarni analiz va sintez qiluvchi dasturiy majmuadir. Uning uch asosiy komponenti mavjud:

– lingvistik (tabiiy tilning formal modeli, lug‘at, grammatika, lingvistik jadvallar, qoidalar);

– matematik-algoritmik (formal til translyatori, matnni qayta ishlovchi algoritmlar);

– dasturiy ta’minot.

Lingvistik bilimlar bazasi mashina uchun matnlari qayta ishlashda zarur bo‘lgan barcha ma'lumotlarni o‘z ichiga qamrab oladi. Tabiiy tilining lingvistik sathlari quyidagilardan iborat: leksik, morfologik, sintaktik, semantik. Bularning har biri o‘zining struktur ma’lumotlar bazasiga ega. Har bir lingvistik birlik muayyan umumiy belgilariga ko‘ra paradigmalarga birlashadi. Lingvistik ma’lumotlar bazasi (ta’minoti) deyilganda, tilga oid barcha ma’lumotlar tushuniladi. Til tuzilishi hamda matnni analiz va sintez qilishning aniq vazifalariga, asosan, lingvistik ta’minotida quyidagilarning mavjud bo‘lishi talab qilinadi: matn – leksik asos – leksemashakl – so‘z birikmasi – sintaktik gap strukturalari[Abdurakhmonova, N.2020,2021,2022].

Tarjimadan avvalgi tahlil jarayonida uch asosiy ierarxik bosqich amalga oshiriladi: matn – jumla – leksema shakl. Bunda leksema shakl keng ma’noda matndan olingan istalgan segment birlik sifatida qaraladi. Gap va so‘z birikmalarining barcha strukturaviy hamda boshga xususiyatlari leksema shakl tushunchasida ifodalanadi.

Lingvistik bilimlar bazasiga quyidagilar kiradi:

Leksik resurslar:

– kompyuter lug‘atlari;

– tezaurus, ontologiya.

Matniy resurslar:

– matnlarning sohaviy majmuasi;

– matn korpusi [Abduraxmonova, 2021: 33].

Quyida lingvistik bilimlar bazasiga kiruvchi tezaurus haqida kichik bir tadqiqot o‘tkazamiz. Tezaurus (yun. “xazina”) muayyan so‘zning leksik-semantik, kontekstual ma’nolarini qamragan lug‘at hisoblanadi, umuman olganda, maxsus terminologiyadir. [Лукашевич, 2010: 89]. Zamonaviy ilmiy paradigmada tezaurusni bilimlar tizimi sifatida tushunishda bir nechta yondashuvlar mavjud. Jahon tilshunosligida tezaurus termini ostida ideografik lug‘at turi tushuniladi, dunyo haqidagi bilimlarning kognitiv tizimi (bu ma’noda tezaurus olam tasviri va mental leksika terminlariga muvofiq) hisoblanadi, keyingi yillarda esa xorij kompyuter

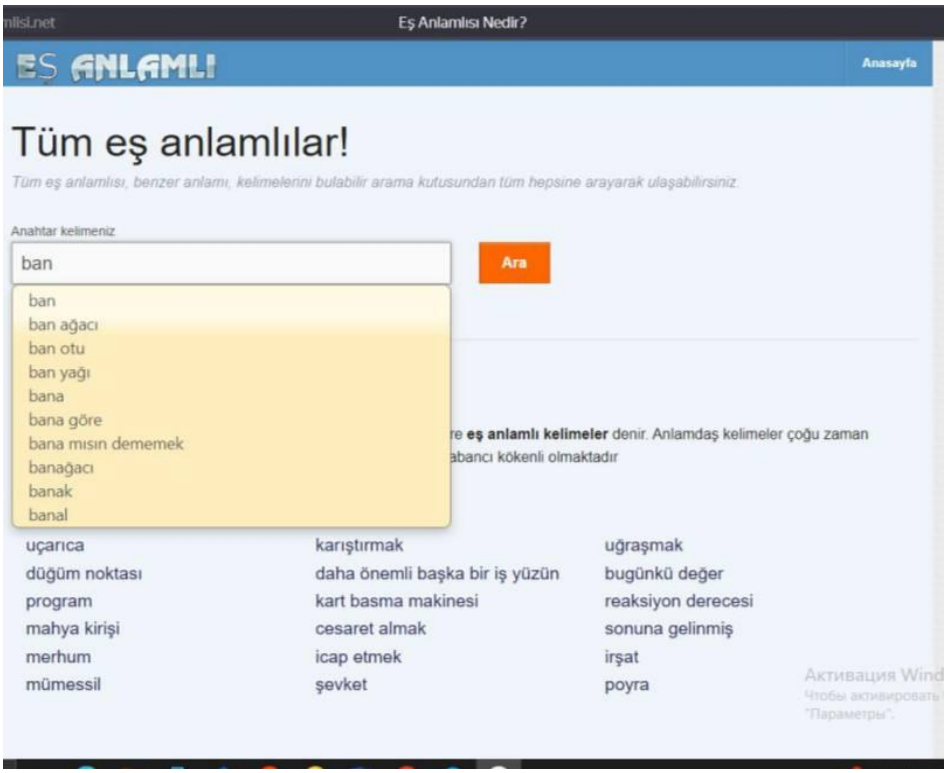
lingvistikasida axborot qidiruv tezauruslari va lingvistik ontologiyalar rivojlanib bormoqda, shuningdek, sun'iy intellekt tizimlarini ishlab chiqishda faol qo'llanilmoqda [Abjalova M., 2022: 117].

Tezaurus, WordNet va boshqa ingliz tili tezauruslari mualliflari ta'kidlaganlaridek, tabiiy til leksikasi yordamida ifodalanadigan bilimlar tizimi, tildagi so'zlarning muayyan ma'nosi bilan cheklanib qolmagan mantiqiy kategoriyalar yordamida boyitilgan ontologiya hisoblanishidan qat'iy nazar tezaurus axborot va ma'lumotlarni taqdim etuvchi semantik (kognitiv) tizim sifatida o'rganiladi. Bunda semantik tizim elementlaridan konsept, tushuncha va "madaniy konstantalar"ga e'tibor qaratiladi. Ontologiyada ushbu elementlardan tashqari semantik munosabatlar birlamchi o'rinda turadi[Abdurakhmonova, N.2020,2021,2022].

Tezauruslar matnlar to'plami yoxud til korpuslaridagi tushunchalarni, maxsus bilim sohasi yoki faoliyat sohasining tushunchalari, ta'riflari va terminlarini qamrab oluvchi leksikografik manbalar hisoblanadi.

Tezaurus – yaqin ma'noga ega so'zlar va so'z birikmalari tushunchalar, konseptlar yoki deskreptor (tavsiflovchi)lar deb ataladigan birliklarga birlashtirilgan hamda ushbu tushunchalar o'rtasida ularning iyerarxik shaklidagi semantik munosabatlari ko'rsatilgan lug'at hisoblanadi [Abjalova M., 2022: 195].

O'zbek tilshunosi Manzura Abjalova "O'zbek tezaurus lug'ati uchun sifat turkumi bazasini yaratish" bo'yicha kichik tadqiqot davomida https://live.babelnet.org, https://www.dictionary.com, https://www.thesaurus.com saytlarini farqli va o'xshash tomonlarini o'rgangan. Ushbu saytlarga qo'shimcha sifatida turk tilshunoslari tomonidan yaratilgan https://esanlamlisi.net saytini keltirishimiz mumkin. Bu saytning afzalligi turk tilidagi ayrim so'zlarning omonimlari kiritilgan bo'lib, lug'atdan har qanday inson onlayn foydalanish imkoniyatiga ega.



O‘zbek tili lingvistik bazasini yaratishda yuqorida tanishib chiqqan bir nechta bazalarni umumlashtirib, ularning har biridan foydalangan holda ish olib kerak. Sifat so‘z turkumi misolida “oq” leksemasining tezaurusi uchun baza tuzib olishimiz kerak.

Masalan:

“Oq” sifatining tezaurus lug‘atini yaratib ko‘ramiz:

Turkumi: Sifat;

Sinonimik qatori: bo‘z, oqish, tiniq

Antonimi: qora;

Omonimi: oqmoq – fe’l;

Paronimi: yo‘q;

Yasalishiga ko‘ra: sodda sifat;

Turiga ko‘ra: Asliy sifat;

Bog‘lanadigan so‘zlari: Ot so‘z turkumidagi barcha so‘zlar;

Qaysi tildan kirib kelgan: qadimgi turkiy tlda ham shunday ma’noni anglatgan bu sifat **a:q** tarzida talaffuz qilingan.

Misollar: Oq it, qora it – baribir it. *(Maqol)*. Xotin o‘tirmoqchi bo‘lgan edi, paranji ichidan uning atlas ko‘ylaklari va nafis oq qo‘llari ko‘rinib ketdi. *(Abdulla Qodiriy “O‘tgan kunlar”)*. Chapay askarlari keldilar bosib, Bosib keldi ular oqlar ustiga. *(A.Oripov “Yillar armoni”).*

Shu shaklda o‘zbek tilida mavjuda bo‘lgan sifatlarning har biriga to‘xtalib o‘tamiz. Ilmiy ishda asosiy manba sifatida “Izohli lug‘at” va https://uzbekcorpus.uz/ dan foydalansak, qo‘shimcha manbalar sifatida badiiy va ilmiy asarlardan foylaniladi.

Xulosa

O‘zbek tilida 50 000 dan ortiq leksemalar mavjud bo‘lib, korpus va lingvistik kompyuter dasturlari bazasida har bir leksema turkumining aniqlab berilishi o‘ta muhim masala hisoblanadi [Abjalova M., 2022: 5]. Korpus lingvistikasida so‘z turkumlarini teglash, grammatik kategoriyalarni teglash va so‘zlarni toifalashda noaniqliklarni bartaraf etish uchun so‘zni faqat uning lug‘atdagi shakliga asoslanib emas, balki matn (jumla)dagi ifodasi bo‘yicha uning turkumlik tegi va jumla (xatboshi, ibora)da boshqa so‘zlar bilan birikish imkoniyatini hisobga olish muhim sanaladi.

Foydalanilgan adabiyotlar:

1. Abduraxmonova N. Kompyuter lingvistikasi. Darslik Toshkent –2021. – B.33.
2. Abjalova M. O‘zbek tili ontologiyasi: yaratish texnologiyasi va konsepsiyasi Monografiya (qayta nashr) “Nodirabegim” nashriyoti Toshkent – 2022. – B.203.
3. Abjalova M., Sharipov E. O‘zbek tezaurus lug‘ati uchun sifat turkumi bazasini yaratish masalasi. “Kompyuter lingvistikasi: muammolar, yechim, istiqbollar” Respublika ilmiy-texnikaviy konferensiya. Vol. 1 №. 01 (2021). –B.189.
4. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М., 2010. – С.89.
5. Agostini, A., Usmanov, T., Khamdamov, U., Abdurakhmonova, N., & Mamasaidov, M. (2021, January). Uzwordnet: A lexical-semantic database for the uzbek language. In Proceedings of the 11th Global Wordnet conference (pp. 8-19).
6. Abdurakhmonova, N., Alisher, I., & Toirova, G. (2022, September). Applying Web Crawler Technologies for Compiling Parallel Corpora as one Stage of Natural Language

Processing. In 2022 7th International Conference on Computer Science and Engineering (UBMK) (pp. 73-75). IEEE.

7. Abdurakhmonova, N., & Ismailov, A. S. (2022). APPLYING WEB CRAWLER TECHNOLOGIES FOR COMPILING PARALLEL CORPORA AS ONE STAGE OF NATURAL LANGUAGE PROCESSING. In СОВРЕМЕННАЯ ФИЛОЛОГИЯ. СОЦИАЛЬНАЯ И НАЦИОНАЛЬНАЯ ВАРИАТИВНОСТЬ ЯЗЫКА И ЛИТЕРАТУРЫ (pp. 22-27).
8. Abduraxmonova, N., & Abduvaxobov, G. I. (2021). O ‘quv lug ‘atini tuzishning nazariy metodologik asoslari. СЎЗ САНЪАТИ ХАЛҚАРО ЖУРНАЛИ, 103.
9. Sulevmanov, D., Gatiatullin, A., Prokopyev, N., & Abdurakhmonova, N. (2020, November). Turkic morpheme web portal as a platform for turkology research. In 2020 International Conference on Information Science and Communications Technologies (ICISCT) (pp. 1-5). IEEE.
10. Po‘latov A. Kompyuter lingvistikasi – T.: Akademnashr, 2011-yil. – B.213.
11. Пулатов А.К., Мухамедова С. Компьютер лингвистикаси (укув кулланма). – Т., 2008. – B.56.
12. Xolmanova Z. Kompyuter lingvistikasi. Toshkent – 2020 “Asian Book House” – B.9.