

ELEKTRON TEZAVURUS YARATISHNING LINGVISTIK ASOSLARI VA SINSET SHAKLLANTIRISH METODOLOGIYASI

Muhammadjon Najmiddinov
Qo‘qon universiteti, DSc, dotsent

DOI: <https://doi.org/10.5281/zenodo.21929637>

Annotatsiya. Ushbu maqolada elektron tezaurus yaratishning lingvistik asoslari, xususan, tezaurus birliklarini tanlash va sinsetlarni shakllantirish metodologiyasi tahlil qilinadi. Tadqiqot maqsadi elektron tezaurusning an'anaviy lug'atdan farqli tuzilish xususiyatlarini aniqlash hamda korpus statistikasi, distributiv semantika va BERT tipidagi transformer modellarni ekspert tekshiruv bilan uyg'unlashtirgan olti bosqichli metodikani ishlab chiqishdan iborat. “O'zbek tilining ta'limiy korpusi” va ARANEUM_UZBEKICUM korpusi materiallari asosida 5 ta leksik-grammatik toifaga oid 77 ta sinset yaratildi. Natijalar elektron tezaurusni axborot-qidiruv va tabiiy tilni qayta ishlash tizimlariga integratsiya qilish uchun lingvistik zamin bo'lib xizmat qiladi.

Kalit so'zlar: elektron tezaurus, sinset, korpus lingvistikasi, semantik munosabatlar, BERT, deskriptor, sinonimiya, ontologiya, o'zbek tili.

Аннотация. В данной статье анализируются лингвистические основы создания электронного тезауруса, в частности, методология отбора тезаурусных единиц и формирования синсетов. Цель исследования состоит в выявлении структурных особенностей электронного тезауруса, отличающих его от традиционного словаря, а также в разработке шестиступенчатой методики, объединяющей корпусную статистику, дистрибутивную семантику и трансформерные модели типа BERT с экспертной верификацией. На материале «Учебного корпуса узбекского языка» и корпуса ARANEUM_UZBEKICUM создано 77 синсетов, относящихся к 5 лексико-грамматическим категориям. Результаты служат лингвистической основой для интеграции электронного тезауруса в системы информационного поиска и обработки естественного языка.

Ключевые слова: электронный тезаурус, синсет, корпусная лингвистика, семантические отношения, BERT, дескриптор, синонимия, онтология, узбекский язык.

Abstract. This article analyses the linguistic foundations of electronic thesaurus construction, focusing on the methodology for selecting thesaurus units and forming synsets. The aim of the study is to identify the structural features that distinguish an electronic thesaurus from a traditional dictionary and to develop a six-stage methodology that integrates corpus statistics, distributional semantics and BERT-type transformer models with expert verification. Drawing on the “Uzbek Educational Corpus” and the ARANEUM_UZBEKICUM corpus, 77 synsets were constructed across 5 lexical-grammatical categories. The results provide a linguistic basis for integrating the electronic thesaurus into information-retrieval and natural-language-processing systems.

Keywords: electronic thesaurus, synset, corpus linguistics, semantic relations, BERT, descriptor, synonymy, ontology, Uzbek language.

Kirish. Raqamli axborot hajmining tobora ortib borishi va semantik veb texnologiyalarining rivojlanishi matnli manbalarni semantik jihatdan tizimlashtirishni zamonaviy tilshunoslikning dolzarb vazifalaridan biriga aylantirdi. An'anaviy, chiziqli tuzilgan lug'atlar katta hajmdagi matn to'plamlarining semantik mazmunini tizimlashtirish uchun yetarli emas [7. 1976:B.14]. Shu sababli lug'aviy birliklar o'rtasidagi semantik munosabatlarni izchil aks ettiruvchi raqamli lingvistik resurs — elektron tezaurus — semantik veb va sun'iy intellekt texnologiyalarining asosiy tarkibiy qismiga aylanmoqda [3. 2001:B.36].

Tezaurus yaratishning nazariy-metodologik asoslari G'arb va rus tilshunosligida keng ishlab chiqilgan: Yu.N. Karaulov va B.V. Dobrovning tezaurus modeli, shuningdek, G.A. Miller rahbarligida yaratilgan WordNet loyihasi zamonaviy tezaurus nazariyasining tayanch negizini tashkil etadi [10. 1995:B.39; 5. 1998:B.58]. Biroq o'zbek tili materialida elektron tezaurus yaratishning lingvistik asoslari, xususan, birliklarni tanlash va sinset shakllantirish metodologiyasi alohida monografik tadqiqot predmeti bo'lmagan.

O'zbek tili aglyutinatil til sifatida boy so'z yasash tizimi, keng ko'p ma'nolilik va sistemali sinonimik-antonimik munosabatlarga ega bo'lib, bu xususiyatlar hozirgacha raqamli shaklda izchil modellashdirilmagan. Mavjud izohli va sinonimlar lug'atlari leksik ma'noni qayd etadi, biroq tezaurus talab

qiladigan semantik munosabatlar tarmog'ini — sinonimiya, giperonimiya, giponimiya, meronimiya va assotsiativlikni — formallashtirmaydi.

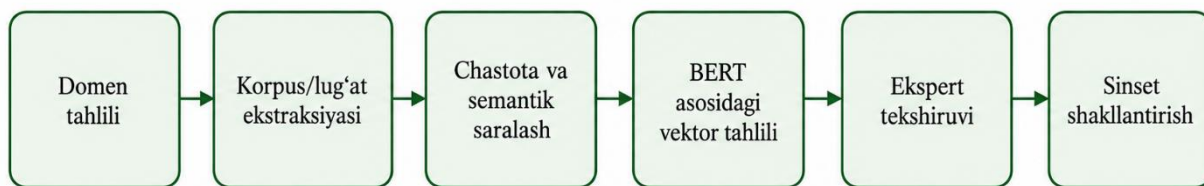
Maqolaning maqsadi — elektron tezaurusning tuzilishi va tarkibini lingvistik-texnologik nuqtayi nazardan tavsiflash hamda BERT tipidagi transformer modellariga asoslangan, tezaurus birliklarini tanlash va sinsetlarni shakllantirish uchun bosqichma-bosqich metodikani ishlab chiqishdan iborat [4. 2019:B.4172; 13. 2017:B.5999].

Tadqiqot metodologiyasi. Tadqiqotning empirik materialini “O'zbek tilining ta'limiy korpusi” va ARANEUM_UZBEKICUM korpusi, shuningdek, o'zbek tilining izohli va sinonimlar lug'atlari tashkil etadi. Ushbu manbalardan predmet nomlari, belgi bildiruvchi sifatlar, harakat/holat fe'llari, holat sifatleri va ravishlarga oid 77 ta leksik birlik tanlab olindi.

Tadqiqotda tavsifiy, komponentli, ideografik, ontologik, onomasiologik va semasiologik, lingvostatistik, miqdoriy, distributiv hamda illyustrativ tahlil metodlaridan foydalanildi. So'zlar orasidagi ma'no munosabatlari qo'lda ekspert tahlili va so'z vektorlariga asoslangan avtomatik distributiv-semantik tahlilning uyg'unlashuvi orqali aniqlandi; nomzod sinonimlar o'rtasidagi kontekstual o'xshashlik BERT tipidagi kontekstual vektor tasvirlari orasidagi kosinusli o'xshashlik sifatida hisoblab chiqildi.

Ko'rib chiqilgan manbalar va korpus materiali asosida tezaurus birliklarini tanlash va sinsetlarni shakllantirishning olti bosqichli, mezonga asoslangan metodikasi ishlab chiqildi (1-jadval). Jarayon domenni tahlil qilishdan boshlanib, chastotaga asoslangan ekstraktsiya, semantik bog'lanishlarni aniqlash, transformer asosidagi vektorli tahlil (kosinusli o'xshashlik chegarasi $\geq 0,75$), ekspert tekshiruvi va yakuniy sinset shakllantirish bosqichlaridan iborat.

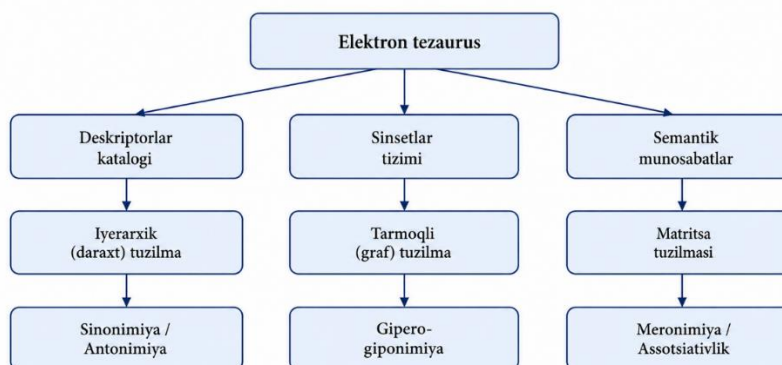
2-rasm. Tezaurus birliklarini tanlash va sinset shakllantirish metodikasi bosqichlari



1-rasm. Tezaurus birliklarini tanlash va sinset shakllantirishning olti bosqichli metodikasi

Elektron tezaurusda axborotni tashkil etishning uchta tuzilma modeli formallashtirildi: (1) turkum-tur va bo'lak-butun munosabatlarini ifodalovchi iyerarxik (daraxt) tuzilma; (2) giperonim-giponim munosabatlarini yo'naltirilgan semantik tarmoq sifatida ifodalovchi tarmoqli (graf) tuzilma; (3) meronimik va assotsiativ munosabatlarni deskriptorlar birgalikda uchrashining o'zaro jadvali sifatida ifodalovchi matritsa tuzilmasi (2-rasm).

1-rasm. Elektron tezaurusning iyerarxik tuzilishi

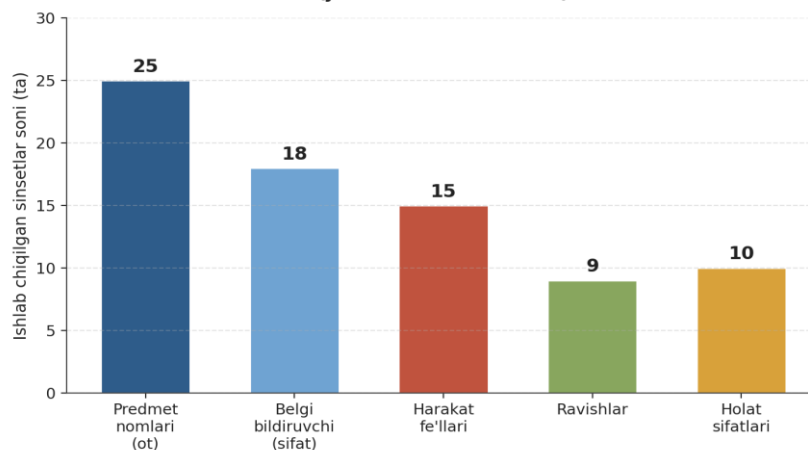


2-rasm. Elektron tezaurusning iyerarxik tuzilishi: deskriptorlar katalogi, sinsetlar tizimi va semantik munosabatlar qatlami

TAHLIL VA NATIJALAR

Olti bosqichli metodikani qo'llash natijasida 5 ta toifaga oid 77 ta leksik birlik uchun sinsetlar shakllantirildi: 25 ta predmet nomi (32,5%), 18 ta belgi bildiruvchi sifat (23,4%), 15 ta harakat/holat fe'li (19,5%), 10 ta holat sifati (13,0%) va 9 ta ravish (11,6%). Predmet nomlari va belgi bildiruvchi sifatlar barcha shakllantirilgan sinsetlarning yarmidan ko'pini tashkil etib, ular umumiy leksikaning ideografik tashkil etilishida markaziy o'rin egallashini ko'rsatadi.

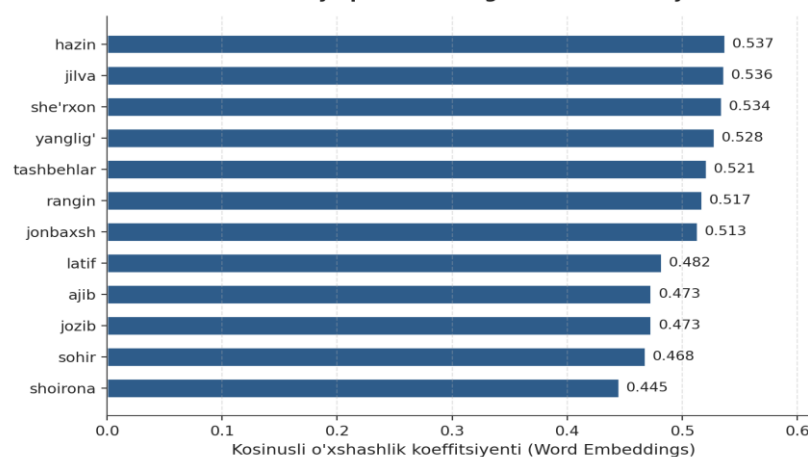
So'z turumlari bo'yicha ishlab chiqilgan sinsetlar taqsimoti (jami 77 ta leksik birlik)



3-rasm. So'z turumlari bo'yicha ishlab chiqilgan sinsetlar taqsimoti (jami 77 ta leksik birlik)

Metodikaning distributiv-semantik bosqichini (1-jadval, 4-bosqich) yoritish uchun inja sifatdoshi ARANEUM_UZBEKICUM korpusida kontekstual so'z vektorlariga asoslangan “o'xshash so'zlar” funksiyasi orqali tahlil qilindi. Natijada kosinusli o'xshashlik koeffitsiyenti 0,445 dan 0,537 gacha bo'lgan 12 ta semantik yaqin sifatdosh aniqlandi; eng yuqori ko'rsatkichga ega hazin, jilva, she'rxon va yanglig' so'zlari ekspert tekshiruvidan so'ng inja sinsetining a'zosi sifatida tasdiqlandi, past ko'rsatkichli nomzodlar esa qo'shni, ammo mustaqil ideografik quyi maydonlarga kiritildi.

«Inja» so'ziga ARANEUM_UZBEKICUM korpusida semantik yaqin so'zlarning o'xshashlik darajasi



4-rasm. “Inja” so'ziga ARANEUM_UZBEKICUM korpusida semantik yaqin so'zlarning kosinusli o'xshashlik ko'rsatkichlari

Ko'p ma'nolilik va omonimiya masalalari alohida ko'rib chiqildi. Ko'p ma'noli leksemalar uchun har bir alohida ma'no mustaqil sinsetga birlashtirildi: masalan, qo'ralay so'zi “hayvon nomi” ideografik maydonida bitta sinset, “o'z” maydonida esa ko'chma ma'noli ikkinchi sinsetni hosil qiladi. Omonim leksemalar tarkibiy jihatdan mustaqil birliklar sifatida qaraldi: bog' so'zi (“bog'-rog” / “bog'lash vositasi”) izohli lug'at ma'lumotlariga tayanib ikkita mustaqil sinsetga ajratildi, ular o'rtasida omonimlik munosabatini belgilovchi giperhavola orqali o'zaro bog'lanadi [1. 1974:B.112].

Sinonimik qatorlarning korpusdagi chastota tahlili sezilarli tarqoqlikni ko'rsatdi: “O'zbek tilining ta'limiy korpusi” bo'yicha eng past ko'rsatkich 132 marta (jozib), eng yuqori ko'rsatkich esa 7152 marta (teran) qayd etildi. Ushbu chastota nomutanosibli sof chastota mezonining o'zi sinsetga a'zolikni belgilash

uchun yetarli emasligini, uni semantik o'xshashlik ko'rsatkichi va ekspert bahosi bilan uyg'unlashtirish zarurligini tasdiqlaydi.

Tadqiqot natijalari elektron tezaurusning axborot-qidiruv, mashina tarjimasi va tabiiy tilni qayta ishlash tizimlariga integratsiyalashuvi uchun amaliy ahamiyatga ega. Xususan, so'rovni kengaytirish (query expansion) mexanizmlarida sinonimlarni faqat xom chastota bo'yicha emas, balki semantik o'xshashlik ko'rsatkichini ham hisobga olgan holda tortishlantirish tavsiya etiladi, aks holda past chastotali, ammo semantik jihatdan markaziy birliklar qidiruv natijalarida yetarli darajada ifodalanmay qolishi mumkin.

Xulosa. O'tkazilgan tadqiqot elektron tezaurusning an'anaviy lug'atdan nafaqat hajm, balki sifat jihatidan ham farqlanishini ko'rsatdi: uning belgilovchi xususiyati semantik munosabatlarning izchil, mashina tomonidan qayta ishlanadigan tarzda kodlashtirilishidir. Tezaurus birliklarini tanlash va sinsetlarni shakllantirishning olti bosqichli metodikasi — domenni tahlil qilish, korpus asosida ekstraksiya qilish, semantik bog'lanishlarni aniqlash, BERT asosida vektorli tahlil, ekspert tekshiruvi va sinset shakllantirish — o'zbek tilining 77 ta leksik birligiga nisbatan qo'llanildi va ko'p ma'nolilik hamda omonimiyani sinsetlarga ajratish va o'zaro bog'lash orqali formal ifodalash mumkinligini ko'rsatdi.

Ishlab chiqilgan metodika o'zbek tilida elektron tezauruslar yaratish, ularni qidiruv tizimlari va tabiiy tilni qayta ishlash tizimlariga integratsiya qilish uchun lingvistik zamin bo'lib xizmat qiladi. Kelgusi tadqiqotlarda metodikani boshqa leksik-semantik maydonlarga tatbiq etish, o'xshashlik chegarasini kengroq etalon to'plam asosida sinovdan o'tkazish hamda natijaviy tezaurusni boshqa tillarning WordNet tipidagi resurslari bilan ko'p tilli moslashtirish istiqbolli yo'nalish bo'lib qoladi.

Foydalanilgan adabiyotlar ro'yxati

1. Апресян Ю.Д. Лексическая семантика: синонимические средства языка. – Москва: Наука, 1974. – 367 б.
2. Asadov T. Leksik-semantik guruhlarining tasnifi. – Toshkent: Fan, 2015. – 214 b.
3. Berners-Lee T., Hendler J., Lassila O. The Semantic Web // Scientific American. – 2001. – Vol. 284, No. 5. – P. 34–43.
4. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of NAACL-HLT. – 2019. – P. 4171–4186.
5. Fellbaum C. (ed.) WordNet: An Electronic Lexical Database. – Cambridge, MA: MIT Press, 1998. – 423 p.
6. Gruber T.R. A Translation Approach to Portable Ontology Specifications // Knowledge Acquisition. – 1993. – Vol. 5, No. 2. – P. 199–220.
7. Karaulov Yu.N. Общая и русская идеография. – Москва: Наука, 1976. – 356 б.
8. Kasares J. Diccionario ideológico de la lengua española. – Barcelona: Gustavo Gili, 1942. – 887 p.
9. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // arXiv:1301.3781. – 2013.
10. Miller G.A. WordNet: A Lexical Database for English // Communications of the ACM. – 1995. – Vol. 38, No. 11. – P. 39–41.
11. Najmiddinov M.G'. Tezauruslar yaratishning lingvistik asoslari: Filologiya fanlari doktori (DSc) dissertatsiyasi avtoreferati. – Andijon, 2026. – 69 b.
12. “O'zbek tilining ta'limiy korpusi”. Elektron resurs. Manzil: <https://uzbekcorpus.uz>
13. Vaswani A. et al. Attention Is All You Need // Advances in Neural Information Processing Systems. – 2017. – Vol. 30. – P. 5998–6008.