



Integration of electronic thesauri with search systems and methodology for selecting electronic thesaurus entries

Mukhammadjon NAJMIDDINOV¹

Kokand University

ARTICLE INFO

Article history:

Received December 2024

Received in revised form

15 December 2024

Accepted 25 January 2025

Available online

15 February 2025

Keywords:

electronic thesaurus,
search engine,
semantic analysis,
thesaurus integration,
database,
artificial intelligence,
linguistic resources.

ABSTRACT

This article examines the integration of electronic thesauri with search engines and the methodology for selecting electronic thesaurus entries. The use of thesauri in search engines enables semantic analysis, broadens the search scope, increases the relevance of results, and facilitates the creation of thematic categories. Additionally, the article discusses the processes of creating, analyzing, and synthesizing electronic thesauri, as well as the corresponding algorithms and technologies.

2181-3663/© 2024 in Science LLC.

DOI: <https://doi.org/10.47689/2181-3701-vol3-iss1-pp90-96>

This is an open-access article under the Attribution 4.0 International (CC BY 4.0) license (<https://creativecommons.org/licenses/by/4.0/deed.ru>)

Электрон тезаурусларнинг қидирув тизимлари билан интеграцияси ҳамда электрон тезаурус бирликларини танлаш методикаси

АННОТАЦИЯ

Калит сўзлар:

электрон тезаурус,
қидирув тизими,
семантик таҳлил,
тезаурус интеграцияси,
маълумотлар базаси,
сунъий интеллект,
лингвистик ресурслар.

Ушбу мақолада электрон тезаурусларнинг қидирув тизимлари билан интеграцияси ва электрон тезаурус бирликларини танлаш методикаси таҳлил қилинади. Тезаурусларнинг қидирув тизимларида қўлланилиши семантик таҳлил, қидирув доирасини кенгайтириш, натижалар долзарблигини ошириш ва мавзу тоифаларини яратиш имкониятларини беради. Шунингдек, мақолада электрон тезаурусларни яратиш, уларни таҳлил этиш ва синтез қилиш жараёнлари, шунингдек, тегишли алгоритмлар ва технологиялар ҳақида сўз юритилади.

¹ PhD, Associate Professor, Kokand University.

Интеграция электронного тезауруса с поисковыми системами и методология выбора единиц электронного тезауруса

АННОТАЦИЯ

Ключевые слова:

электронный тезаурус,
поисковая система,
семантический анализ,
интеграция тезауруса,
база данных,
искусственный интеллект,
лингвистические ресурсы.

В данной статье рассматривается интеграция электронных тезаурусов с поисковыми системами и методика выбора единиц электронного тезауруса. Использование тезаурусов в поисковых системах позволяет проводить семантический анализ, расширять область поиска, повышать актуальность результатов и создавать тематические категории. Также обсуждаются процессы создания, анализа и синтеза электронных тезаурусов, а также соответствующие алгоритмы и технологии.

Электрон тезаурусларнинг қидирув тизимлари билан интеграциялашуви маълумот қидириш соҳасининг ривожланишида муҳим қадамдир. Бу имконият қидирув тизимларига қидирув сўровлари “маъноси”ни аниқроқ тушуниш ва тегишли натижаларни тақдим этиш имконини беради. Электрон тезаурусларнинг қидирув тизимлари билан интеграциясининг аҳамияти қуйидагиларда кўринади:

Семантик таҳлил. Тезаурус қидирув тизимига сўровдаги сўзларнинг нафақат аниқ мослиги, балки уларнинг семантик маъносини ҳам таҳлил қилишга ёрдам беради. Масалан, фойдаланувчи “мева” ҳақида маълумот қидирса, тизим “резаворлар”, “ситрус” ва ҳоказо тушунчаларга тегишли натижаларни ҳам тақдим этиши мумкин, чунки бу атамалар тезаурусда семантик боғланган.

Қидирув доирасини кенгайтириш. Тезаурус туфайли қидирув тизими маълумотни нафақат сўровдаги аниқ сўз(лар), балки синоним, антоним ва бошқа тегишли бирликлар бўйича ҳам топади. Бу қидирув имкониятларини сезиларли даражада кенгайтиради; фойдаланувчи сўровни шакллантиришда ҳисобга олмаган маълумотларни топиш имконини беради.

Натижаларнинг долзарблигини ошириш. Тезаурус қидирув тизимига қидирув натижаларини сўровга мослигига қараб тартиблашда ёрдам беради. Калит сўзлар ва тегишли тезаурус бирликларини қамраб олган натижалар рўйхатда юқорироқ кўринади.

Мавзу тоифаларини яратиш. Тезаурус мавзу тоифаларини яратиш ва улар бўйича маълумотларни таснифлаш имконини беради. Бу фойдаланувчига маълум мавзу(лар) бўйича маълумот қидиришни осонлаштиради, уларга керакли материал(лар)ни тез ва аниқ топиш имконини беради.

Электрон тезурус тузиш муаммолари доирасида энг долзарб масалалар сифатида қуйидагиларга ечим топиш талаб этилади:

- 1) электрон тезаурусни ишлаб чиқиш алгоритлари;
- 2) тезаурусни тақдим этиш стандартлари;
- 3) тезаурусни қидирув тизимларига интеграциялаш алгоритмлари;
- 4) тезаурусларни бирлаштириш муаммолари ва ечими.

Шулардан келиб чиқиб айтиш керакки, электрон тезурус тузиш жараёнида шу муаммоларга эътибор қаратиш талаб этилади.

Электрон тезаурусларни анализ ва синтез қилиш. Электрон тезаурусларни анализ ва синтез қилиш қидирув тизимлари (Google, Yandex ва ҳ.), машина таржимаси тизимлари (Google translate, QTranslate 4.1, Lingoes 2.8.1, Microsoft Bing, Yandex Translate, Baidu Translate va DeepL) ва сунъий интеллект тизимлари (GPT-4, Gemini) каби интеллектуал тизимларни яратиш ва ривожлантиришдаги асосий жараёнларга ёрдам беради.

Электрон тезаурусларни таҳлил қилиш. Электрон тезаурусларни таҳлил қилиш қуйидаги амалларни қамраб олади:

1. *Семантик муносабатларни аниқлаш.* Тезаурус таҳлили терминлар орасидаги семантик алоқалар (синонимия, антонимия, гиперонимия, гипонимия ва б.) ҳақида маълумот олиш имконини беради. Бу матнларнинг маъносини тушуниш ва семантик тармоқ(лар) яратишда зарур.

2. *Тезаурус сифатини баҳолаш.* Таҳлил тезауруснинг мукамаллик даражаси, аниқлиги ва изчиллигини баҳолаш имконини беради. Бу унинг муайян вазифаларга мувофиқлигини аниқлашда муҳим.

3. *Хато ва номувофиқликни аниқлаш.* Мазкур таҳлил тезаурус таркибидаги тўсиқ, зиддият ва нотўлиқ таъриф/изоҳ/теглар каби хатоларни аниқлашга ёрдам беради.

4. *Тезаурусларни таққослаш.* Мазкур таҳлил турли тезаурусларни солиштириш, уларнинг ўхшашлик ва фарқларини аниқлаш имконини беради. Бу таҳлил турли тезаурусларни ягона тизимга бирлаштиришда аҳамиятли.

Электрон тезауруслар синтези. Электрон тезаурусларни синтез қилиш қуйидаги амалларни қамраб олади:

1. *Янги тезаурус яратиш.* Синтез мавжуд маҳсулот асосида ёки янги тезаурус яратишга имкон беради. Бу янги мавзулар ёки мавжуд тезаурусларни кенгайтиришда керак бўлади.

2. *Мавжуд тезаурусларни янгилаш.* Синтез янги бирлик, атама ва муносабатларни қўшиш, эскирганларини олиб ташлаш орқали мавжуд тезаурусларни янгилаш имконини беради.

3. *Тезаурус интеграцияси.* Синтез турли тезаурусларни ягона тизимга бирлаштиришга имкон беради, бу эса билим(маълумот)ларнинг янада умумлаштиради ва тўлдиради.

4. *Автоматик тезаурус қуриш.* Синтез катта матн корпусига асосланган тезаурус яратиш жараёнини автоматлаштириш имконини беради.

Электрон тезаурусларни анализ ва синтез қилиш қуйидаги соҳаларда қўлланилади (Қаранг: 1-чизма).

Электрон тезаурусларни анализ ва синтез қилиш усуллари ҳам мавжуд бўлиб, улар орасида қуйидагилар оптимал саналади (Қаранг: 2-чизма).

Хулоса қилиб шуни айтиш мумкинки, электрон тезаурусларнинг анализ ва синтези замонавий ахборот технологияларини ривожлантиришда муҳим ўрин тутди. Улар табиий тилни чуқур тушунадиган, қайта ишлай оладиган ақлли тизимларни яратишга имкон беради.

Қидирув механизмлари:

- Тезаурус таҳлили нафақат сўзларнинг аниқ мос келиши, балки уларнинг семантик муносабатларини ҳам ҳисобга олган ҳолда қидирув натижаларининг долзарблигини оширади.

Машина таржимаси тизимлари

- Тезаурус синтези юқори сифатли машина таржимаси учун зарур бўлган кўп тилли тезаурусларни яратишга ёрдам беради.

Сунъий интеллект тизимлари:

- Тезаурус таҳлили ва синтези кўплаб сунъий интеллект тизимларининг асоси бўлган онтологияларни яратишда ишлатилади.

Ахборот тизимлари:

- Тезаурус турли ахборот тизимларидаги маълумотларни таснифлаш ва тартибга солиш учун ишлатилади.

1-чизма. Электрон тезаурусларни анализ ва синтез қилиш соҳалари



Статистик усуллар: сўз частотасини таҳлил қилиш



Машинани ўрганиш: семантик муносабатларни автоматик равишда ажратиб олиш учун катта матнли корпусда моделларни ўргатиш.



Лингвистик усуллар: грамматик ва лексик қоидалардан фойдаланиш.



Эксперт тизимлари: тезаури яратиш ва янгилаш учун эксперт билимларидан фойдаланиш.

2-чизма. Электрон тезаурусларни анализ ва синтез қилиш усуллари

Электрон тезаурус маълумотлар базаси ҳақида. Электрон тезаурус яратиш – бу маълумотларни синчковлик билан танлашни ва қайта ишлашни талаб қиладиган мураккаб жараён. Олинган тезауруснинг сифати кўп жиҳатдан фойдаланиладиган ахборот манбаларининг ишончлилиги ва долзарблигига боғлиқ.

Электрон тезаурус яратиш учун маълумотлар манбасининг қуйидаги асосий турлари келтирилади:

I. Лингвистик ресурслар: *Луғат.* Тезаурус яратишда изоҳли, этимологик, синоним, антоним ва бошқа турдаги луғатлар асосий лексик бирликлар ва уларнинг маънолари ҳақида маълумот беради.

Грамматика. Маълум бир тилнинг электрон тезаурусини яратиш учун сўзнинг грамматик категорияларини, уларнинг бирикиши ва синтактик функцияларини тасвирлашда тилнинг грамматикаси мазкур маълумотларни тақдим этиши билан тезауруснинг маълумотлар базаси вазифасини бажаради.

Матн корпуслари муайян тилдаги матнларнинг тўплами сўзлар частотаси, уларнинг қўлланилиш контексти ва семантик муносабатларни аниқлаш имконини беради.

II. Онтологиялар: *Мавжуд тезауруслар* янги тезаурус яратиш ёки мавжуд бўлган тезаурусни кенгайтириш учун асос сифатида фойдаланилиши мумкин.

Онтологиялар тушунча, хоссалар ва уларнинг ўртасидаги муносабатларни ўз ичига олган предмет соҳаси ҳақидаги билимларни тақдим этади.

Экспертлар тизимлари муайян соҳадаги экспертларнинг билими ушбу соҳанинг ўзига хос хусусиятларини акс эттирувчи тезауруслар яратишда ишлатилиши мумкин.

III. Интернет ресурслари: *Википедия* турли тушунча ва атамалар ҳақида кенг маълумотларни ўз ичига олади.

Ихтисослаштирилган веб-сайтлар муайян предмет соҳаларида тезаурус яратиш учун манба вазифасини ўтайди.

Ижтимоий тармоқлар фойдаланувчилар контентини таҳлил қилиш янги атамалар ва уларнинг қўлланилишини аниқлашга имкон беради.

Демак, электрон тезауруслар яратишда лингвистик ресурслар, онтологиялар, интернет ресурслари лингвистик таъминот вазифасини бажаради. Тезаурус маълумотлар базасини ишлаб чиқишда мазкур манбалардан маълумот олиш мақсадаг мувофиқ.

Маълумотлар манбаини танлаш мезонлари. Маълумотлар манбаларини танлашда қуйидаги **мезонларга** амал қилиш мақсадга мувофиқ:

Долзарблик: танланаётган ва маълумот олинadиган манба тезауруснинг фан соҳасига тегишли маълумотларни ўз ичига олиши керак.

Ишончлилик: танланаётган ва маълумот олинadиган манба ишончли муаллиф ёки ташкилотлар томонидан яратилган бўлиши мақсадга мувофиқ.

Доимийлик: танланаётган, маълумот олинadиган манбадаги маълумот(лар) долзарб бўлиши, билимларнинг замонавий ҳолатини акс эттириши талаб этилади.

Манбанинг очиқ эканлиги: танланаётган ва маълумот олинadиган манба тезаурус яратиш жараёнида фойдаланиш учун очиқ эканлиги муҳим, ёпиқ манбалардан фойдаланиш имконсиз. Бу ерда open source, API масаласи кун тартибига чиқади.

Маълумотни ажратиб олиш усуллари. Маълумотлар манбаларидан маълумотларни олишнинг бир қанча усуллари мавжуд бўлиб, уларнинг энг мақбуллари қуйида келтирамиз:

Қўлда таҳлил қилиш усули. Мутахассис матн(лар)ни қўлда таҳлил қилади, ундан атама ва уларнинг муносабатларини аниқлайди.

Матнни автоматик қайта ишлаш усули. Морфологик, синтактик ва семантик таҳлил каби компьютер лингвистикаси усулларида фойдаланган ҳолда таърифланаётган атама (тезаурусдаги бирлик) ва улар орасидаги лексик-семантик муносабатлар автоматик равишда аниқланади. Бунга word embeddings функцияси мисол бўлади (Бу функциядан фойдаланиш ҳақида ишнинг кейинги қисмларида батафсил тўхталамиз).

Аралаш (бирлашган) усуллар. Қўлда ва автоматик таҳлилнинг комбинацияси олинган маълумот(лар)нинг аниқлиги ва мукамаллигини ошириш имконини беради.

Электрон тезаурус яратиш учун воситалар. Электрон тезаурус яратиш учун махсус воситалар талаб қилинади. Тезауруслар “0”дан яратилмайди, балки маълум бир маълумотлар базасидан маълумотлар йиғилади ва тартибланади. Шу каби воситалар сирасида қуйидагиларни келтириш мумкин:

Тил корпуслари. Катта матнларни таҳлил қилиш асосида сўзлараро семантик муносабатларни аниқлаш учун тил корпуслари матнлари таҳлил қилинади, натижада маъноси яқин сўзлар – ўхшаш сўзлар аниқланади. Бу word embedding функцияси учун талаб этилади. Мазкур функцияни ишга тушириш учун тил корпуси талаб этилади. Намуна учун ўзбек тили морфологик анализаторидан word embedding функциясидан намуна келтирамиз. *Мева* гуруҳига кирувчи меваларни топиш учун ўзбек тили корпуси матнларига ишлов бериш асосида шакллантирилган word embedding функцияси асосида мева номлари тўпланди (Қаранг: 1-жадвал). Жадвалдан кўринадики, йиғилган сўзларнинг ҳаммаси ҳам мева номлари эмас, шунчаки ўсимлик номлари ҳам бу тўпламга кириб қолган: *ловия, бодринг, редиска, отқулоқ, карам* ва ҳ. сўзлар мева номи эмас. Лекин статистик усул асосида йиғилгани сабабли бундай сўзлар ҳам *мева* уясига мансуб деб хулоса қилинган. Тезаурус тузиш жараёнида ҳосил қилинган рўйхат шундай таҳлил қилиниб, натижанинг аниқлиги текширилади ва гуруҳ таркиби тўғри шакллантирилади.

1-жадвал.

Ўзбек тили морфологик анализаторидан word embedding функцияси олма сўзи “ўхшашлари” қидирув натижаси

1. banan 0,7350973	21. ananas 0,6686466
2. olma 0,7338421	22. turshak 0,6665593
3. olmaning 0,726738	23. olmani 0,6659229
4. mango 0,7162268	24. do'lana 0,6644421
5. nok 0,7115675	25. loviya 0,662054
6. mandarin 0,71104	26. shaftoli 0,6616899
7. apelsin 0,7086598	27. qulupnay 0,6562068
8. kivi 0,7054288	28. o'rik 0,6547231
9. bodom 0,7006688	29. lolahovuz 0,6546866
10. nok 0,7000052	30. mandarinlar 0,6544503
11. o'rik 0,6992868	31. bodring 0,6543946
12. shaftoli 0,6978537	32. na'matak 0,6510405
13. rediska 0,696831	33. xurmo 0,650313
14. meva 0,6949912	34. olxo'ri 0,6500984

15. olmalar 0,6939033	35. anjir 0,6489938
16. limon 0,6892098	36. otquloq 0,6479493
17. turp 0,6860158	37. mevalar 0,6448801
18. ulfatan 0,682215	38. ko'chasi-tup 0,6443496
19. karam 0,6729299	39. nokni 0,6423482
20. olcha 0,6715319	40. avokado 0,6423462

Онтология муҳаррирлари. Онтологияларни яратиш ва таҳрирлаш учун мавжуд онтологияларга эҳтиёж сезилади.

Семантик тармоқ визуализация дастури. Олинган натижаларни тақдим этиш ва таҳлил қилиш учун семантик тармоқ натижаларини визуализация дастурлари талаб этилади. Намуна учун ARANEUM корпусидаги визуализациядан намуна келтирамыз (Қаранг: XX-чизма).

Хулоса сифатида айтиш керакки, электрон тезаурус яратиш учун тил корпуси, онтология, семантик тармоқ визуализация дастури каби махсус воситаларга эҳтиёж сезилади.

Фойдаланилган адабиётлар рўйхати:

1. Mamatov, A. E. (2019). Zamonaviy lingvistika. Toshkent: Noshir, 135.
2. Suyunov, B. (2021). Tezaurus va kompyuter texnologiyalariga doir. Computer Linguistics: Problems, Solutions, Prospects, 1(1).
3. Abdullayeva, N. (2023, April). Tezaurus Tushunchasi, Etimologiyasi, Va Til O'rganishda Qo'llanilishi. In Conference on Applied and Practical Sciences (pp. 117-119).
4. Abjalova, M., & Sharipov, E. (2021). O'ZBEK TEZAURUS LUGATI UCHUN SIFAT TURKUMI BAZASINI YARATISH MASALASI. COMPUTER LINGUISTICS: PROBLEMS, SOLUTIONS, PROSPECTS, 1(1).
5. Ermatov, I. R. (2012). Morfem va so'z yasash satxlari terminologik tizimi. Toshkent: O'zbekiston Milliy Universiteti Nashriyoti.
6. Kuzmenko, N. V. (2004). Meronymy in Modern English in the Light of Grammaticalization Theory. Moscow State University.
7. Ermatov, I. R. (2012). Morfem va so'z yasash satxlari terminologik tizimi. Toshkent: O'zbekiston Milliy Universiteti Nashriyoti.
8. Kuzmenko, N. V. (2004). Meronymy in Modern English in the Light of Grammaticalization Theory. Moscow State University.