Ilmiy tadqiqotlarda qiyosiy va tahlil korpuslarining o‘xshash jihatlari

Narzillo Raxmatilloyevich ALOYEV
tayanch doktorant
Alisher Navoiy nomidagi o‘zbek tili va adabiyoti universiteti
Toshkent O‘zbekiston
vip.alayev@gmail.com

Annotatsiya

Maqolada korpus tilshunosligiga bag‘ishlangan ikkita til korpusi tahlilga tortilgan.

Tayanch so‘zlar: korpus lingvistikasi, til korpusi, dispersiya, stemlash, lemmalash.

Сходства сравнительных и аналитических корпусов в научных исследованиях

Назрулло Рахматуллоевич АЛОЕВ
базовый докторант
Университет узбекского языка и литературы имени Алишера Навои
Ташкент, Узбекистан
vip.alayev@gmail.com

Аннотация

Статья посвящена корпусному языковедению, и в ней анализируются два языковых корпуса.

Ключевые слова: корпусная лингвистика, языковой корпус, дисперсия, стемминг, лемматизация.

Korpus lingvistikasi bilan yoki uning doirasida ishlaydigan tadqiqotchilarning fikrlarini o‘rganish uchun tegishli onlayn jurnallardan kichik tadqiqot maqolalari korpusi – Korpus Lingvistikasi Tadqiqotlarini Baholash Korpusi (CLRA-Corpus Lunguistics Research Articles) tuzilgan. Faqat kitob nashrlariga diqqatni qaratish o‘rniga, tadqiqot maqolalari kengroq tadqiqotchilar doirasini ifodalashi mumkinligi sababli tanlangan. Maqolalar turli onlayn nashriyotlardan, agar ularning kalit so‘zlarida “korpus lingvistikasi” atamasini o‘z ichiga olgan bo‘lsa (37 ta maqola) yoki ularning kalit so‘zlari “korpus lingvistikasi” atamasini o‘z ichiga olmagan bo‘lsa, sarlavhasida mavjud bo‘lsa (10 ta maqola) yuklab olingan. Ushbu turdagi tanlovning sababi, maqolalar korpusi o‘z navbatida, korpus tilshunosligiga bag‘ishlanganligini aniq tasavvur qila olishdir.

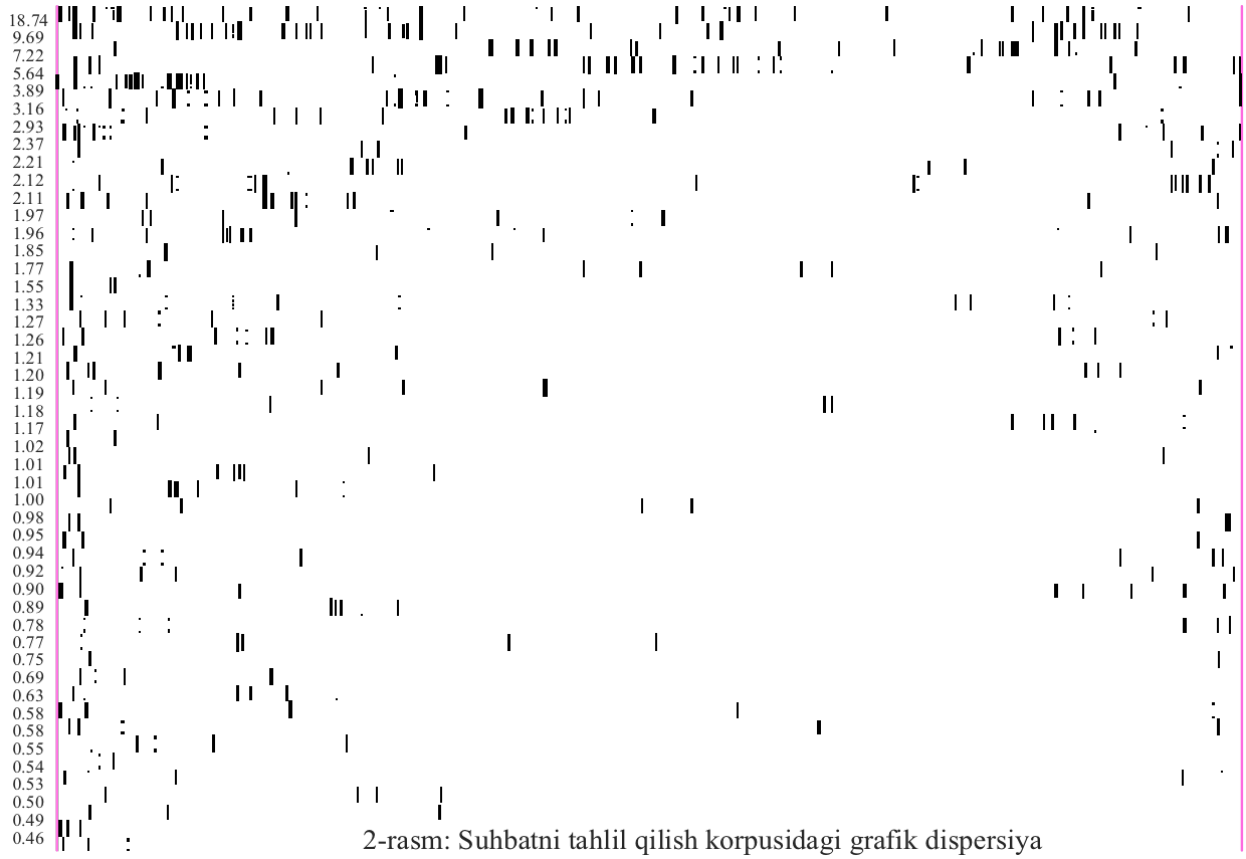
Mavjud barcha ma’lumotlarni o‘rganishdan oldin korpus lingvistikasiga ta’rif bermaslik uchun maqolalar oldindan tanlanmagan. 1996-yildan 2007-yilgacha 20 ta turli jurnaldan jami 47 ta maqola to‘plangan, natijada 463489 ta token tashkil etilgan. Korpusning o‘lchami shunchaki *Siena universitetida* obuna bo‘lgan jurnallar va ularni *WordSmith* vositalari yordamida tahlil qilish uchun zarur bo‘lgan .txt formatiga aylantirish mumkin yoki yo‘qligi bilan aniqlanadi. Matnlar *sarlavha, kalit so‘zlar, referat, havolalar* va *matnni aniqlash* uchun belgilangan. Asosiy qism kirishdan xulosagacha, izohlar bundan mustasno, deb belgilanfdi va asosiy qism ichida kirish hamda xulosa bo‘limlari ham belgilanadi.

Taqqoslash va kengaytirish uchun ikkita qo‘shimcha korpus tuzildi. Birinchisi, suhbat tahliliga oid tadqiqot maqolalaridan iborat bo‘lib, u korpus tilshunosligi kabi tilshunos-amaliyotchi jamiyatida alohida o‘rin egallaganligi sabab tanlab olingan. Suhbatni tahlil qilish korpusi CLRA korpusi kabi onlayn jurnallardan tanlangan 58 ta tadqiqot maqolasini (630049 ta token) o‘z ichiga oladi: agar ularning kalit so‘zlarida *“suhbat tahlili”* atamasi bo‘lsa va ularni .txt formatiga aylantirish mumkin bo‘lgan taqdirda, yuklab olingan. Ikkinchi qiyosiy korpus, CONF, Corpus Lingvistika-2005 konferensiyasida taqdim etilgan ma’ruzalardan iborat (2005-yil 14–17-iyul kunlari Birmingemda bo‘lib o‘tgan) va 69 ta maqolani (349942 token) o‘z ichiga oladi. “.txt” formatiga o‘zgartirilishi mumkin bo‘lgan barcha maqolalar *Proceedings for the Corpus Linguistics Series 2005* veb-saytidan yuklab olingan. Garchi kalit so‘zlar maqola formatining bir qismi bo‘lmasa-da, ular Korpus lingvistikasi-2005 konferentsiyasida taqdim etilganligi sababli asosan korpus lingvistikasiga tegishliligi ta’kidlab o‘tildi. Barcha korpuslar WordSmith Tools 3.0 yordamida tahlil qilindi.

CLRA korpusida korpus lingvistika atamasining chastotasi va tarqalishi juda qiziq holatda namoyon etildi. Dispersiya grafigidan (quyidagi rasm) ko‘rinib turibdiki, korpus lingvistikasi atamasi tadqiqot maqolalari orasida ham, ular ichida ham notekis taqsimlangan bo‘lib chiqdi. O‘nta maqolada korpus lingvistikasiga oid faqat bitta eslatma mavjud bo‘lib, u korpusning tuzilmasini hisobga olgan holda,

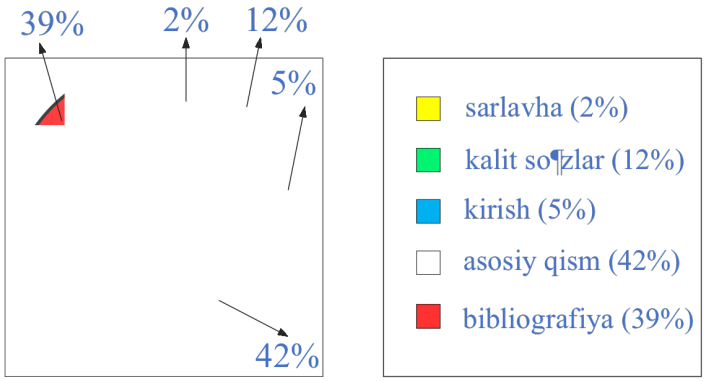
sarlavhada, kalit so‘zlarda yoki abstraktda bo‘lgan. Ajablanarlisi shundaki, 47 ta tadqiqot maqolaning taxminan yarmi – 23 tasi matnning asosiy qismida korpus lingvistikasi haqida umuman eslatib o‘tilmagan. Ushbu 23 tadan 20 tasi o‘z kalit so‘zlarida “korpus lingvistika” atamasini o‘z ichiga olganligini hisobga olsak, bu kalit so‘zlarning haqiqiy funksiyasi yoki “korpus lingvistikasi” atamasining qo‘llanilishi haqida turli xil savollar tug‘dirishi mumkin.

1-rasm: CLRA korpusidagi dispersiya grafigi



2-rasm: Suhbatni tahlil qilish korpusidagi grafik dispersiya

Taqqoslash qulayligi uchun 2-rasmdagi dispersiya grafiklari ketma-ket joylashtirilgan va qidiruv so‘zlarining paydo bo‘lish chastotasidagi farqlari aniq namoyon etilgan. Suhbatni tahlil qilish korpusida “suhbat tahlili” 388 marta (0,6 ptw) va uning qisqartmasi “ST” 434 marta (0,7 ptw) sodir bo‘lgan. Taqqoslash uchun, “korpus lingvistikasi” atamasi CLRAda atigi 0,7 ptw marta ko‘rinadi. Ajratma diagrammasi ikkala korpusning maqolalarida ham xuddi shunday tendensiyalarni ko‘rsatdi, chunki hodisalar boshi va oxirida jamlangan. Biroq, ko‘rinib turibdiki, “Suhbat Tahlili” korpusida “CLRA” korpusidan farqli o‘laroq, bitta yozuvli maqolalar mavjud emas edi. Quyidagi rasmda korpus lingvistikasining tadqiqot maqolalarida taqsimlanishi batafsilroq ko‘rsatilgan. Sarlavha, kalit so‘z va abstrakt holatlar umumiy hajmning 19 foizini tashkil etgan bo‘lsa, 38 foizi bibliografiyalarda, atigi 43 foizi maqolaning asosiy matnida topilgan. Bundan tashqari, asosiy matndagi 9 ta holat kichik sarlavhalarda bo‘lgan, bu korpusda matnni korpus lingvistikasiga tegishli deb belgilanishiga umumiy moyillikni aks ettiradi.



3-rasm: Korpus lingvistikasining ilmiy maqolalardagi taqsimoti

CONF korpusida juda o‘xshash natijalar olingan. Korpus lingvistika atamasi 70 ta maqoladan atigi 37 tasida mavjud bo‘lib, 107 ta holatdan bibliografik havolalar taxminan 60 foizni tashkil etgan. Ushbu bibliografik ma’lumotnomalarni hisobga olmaganda, 70 ta maqoladan faqat 16 tasida korpus lingvistika atamasi mavjud bo‘lib chiqdi. 2-rasmda ko‘rsatilganidek, tadqiqot maqolalarining boshida va oxirida kuzatilgan diqqatlikka qaytadigan bo‘lsak, batafsilroq tahlil shuni ko‘rsatdiki, bizning qidiruv so‘zimiz kirish qismida 34 marta (16 maqolada) va xulosalarda 12 marta (9 ta maqolada) sodir bo‘lgan, bu umumiy miqdorning mos ravishda 27 va 10 foizini tashkil qiladi. Ushbu mutanosiblik CONF korpusida yuqoriroq edi: asosiy korpusda 38 ta holat mavjud bo‘lib, ulardan 16 tasi (42%) kirish qismida va 7 tasi (18%) xulosalarda topilgan.

“Korpus lingvistikasi” atamasi kutilganidan kamroq tarqalganligi haqida bahslashish mumkin, chunki CLRA korpusidan “korpus” so‘zining R1 birikmalaridan ko‘rinib turganidek, ko‘plab muqobil variantlar mavjud (1-jadvalga qarang). Masalan, asoslangan va boshqariladigan birikmalar korpus lingvistikasining turini belgilovchi qo‘shimcha qatlamini qo‘shish sifatida ko‘rish mumkin. Biroq, atamalar korpus + asosli/boshqariladigan + tilshunoslik shaklida kamdan-kam uchraydi. Ushbu korpusdagi eng tez-tez uchraydigan klasterlar korpusga asoslangan tadqiqotlar/tadqiqotlar va korpusga asoslangan tadqiqotlar edi. Bu “korpus lingvistikasi” atamasi yoki boshqa atamalar qo‘llanilayotganligi haqidagi xavotirni ko‘rsatishi mumkin, chunki kalit so‘zlarda ishlatilgan korpus lingvistikasi tadqiqot faoliyati emas, balki tadqiqot sohasi yoki intizomini va

akademik yo‘nalishini bildiradi. Bu o‘z navbatida, masalan, korpus tahlili va korpus tadqiqoti birikmalarida tasvirlangan:

	Umumiy	R1	Maqolalar sanog¶i
Asos	196	154	27
Lingvistika	138	124	24
Ma¶lumotlar	136	78	16
Tahlillar	74	28	14
Nazoratli	37	24	5
Lingvistik	37	23	10
Dalil	33	21	11
Yondashuv	50	17	5
Izlanmalar	59	17	10
Tadqiqot	77	16	13

1-jadval. Tanlab olingan R1 CLRAdagi maqolalarning asosiy matnidan foydalanib korpusni yig¶adi.

Bo‘limning asosiy qismida (CLRA-Y) korpus lingvistikasini o‘z ichiga olgan tadqiqot maqolalarini (CLRA-N) bo‘lmaganlar bilan taqqoslash.

Garchi bu yerda ishlatilgan korpus har qanday qat’iy umumlashtiruvchi xulosa uchun juda kichik bo‘lsa-da, WordSmith KeyWords taqqoslash natijalari shuni ko‘rsatadiki, korpus ikkita bir-biridan farq qiladigan nutq jamiyatini ifodalaydi. Masalan, CLRA-Y subkorpusi uchun 3 so‘zli klasterlarga Britaniya Milliy Korpusi (Keynes 31.4) va Bank of English (28.4) kiradi. Shunisi e’tiborga loyiqki, CLRA-Y subkorpusi uchun asosiy 2 so‘zli klasterlar “korpusga asoslangan” (39.4) va “korpusga asoslangan” (35.8) ni o‘z ichiga oladi, bu esa asosiy matnda korpus lingvistikasidan foydalanmaydigan maqolalarda korpus lingvistikasining oddiygina ushbu atamalar bilan almashtirilishi ehtimolini istisno qiladi. CLRA-Y uchun boshqa kalit 2 so‘zli klasterlarga korpus (35.8) va muvofiqlik satrlari (46.7), shuningdek semantik prosodiya (132.3), matn (56.8), matn tahlili (46.3), ma’no (31.3) va soha (24.5) kiradi. CLRA-Y uchun kalit so‘zlar quyidagi odamlarga tegishli: Sinclair, Stubbs, Lowe va Hunston va CLRA-N

uchun kalit so‘zlar ro‘yxatidagi yagona muallif Lakoff edi. Ikki nutq jamoalari o‘rtasidagi farqni ko‘rsatadigan boshqa kalit so‘zlar 2-jadvalda keltirilgan:

CLRA -Y		CLRA -N	
tahlillar	qidiruvlar	iqtibos	modal
tahlil	semantik	kontseptual	protsessual
uyushmalar	semantiklar	domen	bog¶liq
kombinatsiya	to¶plam	ekspres	dolzarb
kompyuter	ijtimoiy	ifoda	qayta
korpus	dasturiy ta¶minit	ifodalar	standart
empirik	tuzilma	ajratib olingan	statistik
til	texnologiya	shablon	tokenlar
lingvistik	matn	leksika	transkriptsiya
ma¶no	matnlar	ehtimollik	bayonot

2-jadval. CLRA-Y kalit so¶zlari CLRA-N kalit so¶zlari bilan solishtirilishi

Shunisi qiziqki, CLRA va CONFda ham “korpus lingvistikasi” atamasi ishlatilmagan maqolalarda 3-jadvalda ko‘rsatilganidek, “korpus lingvistikasi” oldingi modifikatoridan foydalanish yoki korpus tilshunoslariga murojaat qilish ehtimoli kamroq bo‘lgan:

	CLRA -Y		**CLRA -N**		**CONF-Y**		**CONF-N**	
	freq	Ptw	freq	Ptw	freq	Ptw	freq	Ptw
corpus linguistic	3	0.015	3	0.015	3	0.014	1	0.005
corpus linguist/s	2	0.010	2	0.010	15	0.072	1	0.005

3-jadval. CLRA-Y kalit so¶zlari CLRA-N kalit so¶zlari bilan solishtirilishi

Tadqiqot ma’lumotlar bazasini kengaytirish maqsadida, WebCorp10 korpus+lingvistikasi+bu ketma-ketlik tarmog‘idan misollar to‘plash uchun ishlatiladi. Quyidagi mos satrlarda noaniq artikldan keyin misollar ko‘rsatilgan. Ko‘rinib turibdiki, yuqorida tavsiflangan talqinlarning xilma-xilligi tematik tarzda tashkil etilgan natijalarda o‘z aksini topadi.

Ta’riflar tilshunoslik sohasidan tortib, vositaga qadar farq qiladi. Satrlaridan ko‘rinib turibdiki, “ilmiy bo‘lmagan” yondashuv eng noaniq ta’riflardan biridir, chunki u korpus lingvistikasining fan sifatida tavsiflari bilan birga va metodologiya sifatida korpus lingvistikasining ta’riflari bilan birga paydo bo‘ladi (muvofiqlik chiziqlari yulduzcha bilan belgilangan). Ta’riflarda “empirik” so‘zidan foydalanish chastotasi CLRA-N bilan solishtirganda CLRA-Y subkorpusi uchun kalit so‘zlardan biri edi, shuningdek, BNC Amaliy fanlarni yozish bo‘limi bilan solishtirganda CLRA korpusi uchun kalit so‘z bo‘lib, bu korpusni qurishning asosiy xususiyati ekanligini ko‘rsatadi. Keyin ta’riflarni qidirish CLRA korpusida takrorlandi, u faqat nashr etilgan maqolalarni o‘z ichiga oladi va shuning uchun korpus lingvistikasida ishlaydigan tadqiqotchilarning tasavvurlarini yanada aniqroq aks ettiradi ta’rif bersa bo‘ladi. Quyidagi mos qatorlar bir xil korpus + tilshunosligi + bu WebCorp dan olingan ketma-ketlik natijalarini ko‘rsatadi. Ko‘rinib turibdiki, bir nechta aniq ta’riflar mavjud:

Korpus lingvistikasini tadqiqot sifatida tavsiflovchi oxirgi satrdagi shablon WebCorpning muvofiqlik satrlarida ham 13 marta topilgan. Ushbu misollarda undan keyin til, inson tili, tilshunoslik, lingvistik ma’lumotlar va elektron matnlar jamlanmasi keltirilib, yana bir-biridan farq qiluvchi talqinlar keltirilgan. Ushbu shablon korpus lingvistikasini shu soha bilan yoki uning doirasida ishlaydigan tadqiqotchilar nima qilayotganiga asoslanib belgilash tendentsiyasiga ishora qiladi.

Ushbu kichik tadqiqotlardan kelib chiqadigan narsa shundan iboratki, korpus lingvistikasi atamasi bilan o‘ziga yarasha “qiyinchiliklar” mavjud, chunki bu, birinchi navbatda, kam uchraydi. CLRA korpusidan olingan natijalarda maqolalar tanlab olingan, chunki kalit so‘zlarda “korpus lingvistikasi” atamasi qo‘llangan, agar kalit so‘zlar yo‘q bo‘lsa, sarlavhada ishlatilgan bo‘lib chiqadi. Quyida matn bilan python dasturlash tili yordamida til korpusiga misol keltirilgan:

Matndan so‘zlarni ajratib olish ko‘proq vaqt talab etadi va bu jarayon matndagi berib o‘tilgan mavzulardan ajratib olishdan ko‘ra ancha murakkabroqdir. Masalan, har bir 1000 ta matndan va har bir matnda o‘rtacha 500 tadan so‘z mavjud

til korpusini tahlil qilamiz. Demak, ushbu korpusni qayta ishlash uchun 500*1000=500000 ta amalni bajarish kerak bo‘ladi. Shunday qilib, ma’lum mavzularni o‘z ichiga olgan matnni tahlil qilishda, agar o‘rtacha 5 ta mavzudan iborat bo‘lsa, ishlov berish 5*500 so‘z = 2500 ta *amal (thread)ni* tashkil qiladi. Bu butun bir hujjatni qayta ishlashdan ko‘ra soddaroq ko‘rinadi, shuning uchun tematik modellashtirish shu kabi muammolarni hal qilishda qo‘l keladi va jarayonni vizualizatsiya qilishni osonlashtiradi.

NLPda matnni qayta ishlashni osonlashtiradigan matnga boshlang‘ich ishlov berish bosqichlarini amalga oshirish lozim:

– *Nomuhim so‘zlar va tinish belgilarini olib tashlash;*

– *Stemming;*

– *Lemmatizatsiya;*

– *Countvectorizer yoki Tf-Idf usuli yordamida statistik qiymatlarni shakllantirish.*

Stemming va Lemmatizatsiya

O‘zbek tili korpusi matnlarini stemlash va lemmalash bo‘yicha B. Elov va Z. Xusainovalar bir qator ilmiy tadqiqotlarni olib borganlar. O‘zbek tilidagi so‘zlarning stemini aniqlashda o‘zak va qo‘shimchaning bitta o‘zak bilan omonim bo‘lishi, so‘zlarga qo‘shimcha qo‘shilganda tovush o‘zgarishiga uchrashi va neologizm hamda NERlarni stemmlash kabi muammolar yuzaga kelishi mumkinligini qayd etishgan va bu muammoni hal etish modellarini keltirishgan. Stemni aniqlashdagi tovush o‘zgarishiga uchrashi muammosini hal qilish uchun birinchi bosqichda o‘zak va qo‘shimchalarning chegaralari aniqlanib, ikkinchi bosqichda esa lemmatizatsiya amalga oshirilgan. Lemmatizatsiya natijasida xato hosil qilingan stemlar lug‘atda mavjud o‘zak (root)ga o‘zgartiriladi. Maqolada keltirilgan usullar B. Elov, R. Alayev, Sh. Hamroyeva va Z. Xusainova tomonidan ishlab chiqilgan o‘zbek tili morfologik analizatoriga tatbiq etilgan.

Agglyutinativ tillarda o‘zak va qo‘shimchalarni birlashtirib so‘z hosil qilinadi. O‘zakka qo‘shimchalar qo‘shilganda fonetik uyg‘unlik va disgarmoniya yuzaga kelgani uchun ham fonetik, ham morfologik o‘zgarishlarni tahlil qilish

zarur. Koʻpgina NLP vazifalarni hal qilishda soʻz shakllarini ularning oʻzakkacha qisqartirish (stemlash)ga toʻgʻri keladi. Soʻzdan barcha flektiv affikslarni olib tashlash va soʻzning qolgan qismini lemmatizatsiya qilish tabiiy tilni qayta ishlash (NLP)ning muhim vazifalaridan biri hisoblanib, ushbu jarayon *stemming* deb yuritiladi. Stemming jarayoni axborot qidirish (IR, Information Retrieval) tizimlarida muhim ahamiyat kasb etadi. Agglyutinativ tillar uchun stemmingni amalga oshirish uchun baʼzi terminlar izohini keltiramiz:

Lemma (leksema) – faqat oʻzakdan yoki oʻzak+soʻz yasovchi qoʻshimcha shaklidan iborat boʻladi. ***L****eksema* (yun. lexis − soʻz, ifoda) – til qurilishining leksik maʼno anglatuvchi lugʻaviy birligi. Leksema bildiradigan maʼno soʻzning material qismi: maʼlum tovush kompleksini maʼlum obyektiv voqelikka bogʻlash bilan kishi ongida yuzaga keladigan mazmun-mundarija.

Stem – soʻz shaklning qoʻshimchalarini kesib tashlashdan hosil boʻluvchi qism boʻlib, baʼzi hollarda maʼno anglatmasligi mumkin. Shuningdek, stem soʻzning morfologik oʻzagi bilan aynan mos boʻlmasligi yoki mos tushishi mumkin.

Turk va uygʻur tillarida stemming jarayoniga quyidagicha taʼrif berilgan:

Stemming – bu soʻzga qoʻshilgan *derivatsion* va *flektiv* qoʻshimchalarni olib tashlash orqali uning oʻzagigacha qisqartirish vazifasidir.

Lemmatizatsiya – lemma (leksema)ning maʼnoli shaklini aniqlash va uni flektiv/hosila shakllari oʻrnida ishlatishni oʻz ichiga olgan tabiiy tilni qayta ishlash usuli.

Lemmatizatsiya va stemming – tabiiy tilni qayta ishlash (NLP)da tahlil murakkabligini kamaytirish maqsadida ishlatiladigan ikkita usul. Bu ikkala usul ham soʻzlarni asosiy shakllariga qisqartirishdan iborat, ammo ular bu amalni bajarish maqsadi va unga erishish usullari bilan farqlanadi.

Lemmatizatsiya – soʻzning asosiy shakli boʻlgan *lemmasi(leksema)ga* qisqarish jarayoni. Bu asosiy shakl koʻpincha soʻzning *lugʻatdagi shakli* deb ham ataladi. Masalan, *“bajargan”* soʻzining lemmasi *“bajarmoq”*, *“sening”* soʻzining

lemmasi esa **“***sen***”** hisoblanadi. Lemmatizatsiya so‘zning *kontekstini*, jumladan, *POS tegini, zamon* va *turkumi*ni hisobga oladi.

Stemming – kontekstni hisobga olmagan holda so‘zning asosiy shakli bo‘lgan o‘zagiga qisqartirish jarayoni. Stemlar ko‘pincha so‘zning o‘zak shakli deb ataladi. Masalan, *“yugurish”* so‘zining o‘zagi *“yugur”*, *“sening”* so‘zining o‘zagi esa *“sen”, sening* so‘zinig stemi esa *“se”*dir. *Stemming* – bu NLP masalalarini yechishdagi murakkabliklarni kamaytirishning asosiy omili bo‘lib, so‘zning kontekstini hisobga olmaydi. Bu esa so‘z ma’nosini ifodalashda ko‘plab xatolarga olib kelishi mumkin.

Lemmatizatsiya – qidiruv tizimi samaradorligini oshirish uchun so‘zlarni lemma (asosiy) shakliga qisqartirish jarayoni. Ushbu jarayon, ayniqsa, qidiruv natijalarining aniqligini va korpus matni kontentini tushunish uchun foydali bo‘lishi mumkin. Lemmatizatsiyaning asosiy afzalliklaridan biri shundaki, u *qidiruv tizimlariga matn konteksti va “ma’nosi”ni yaxshiroq tushunishga yordam beradi*.

FOYDALANILGAN ADABIYOTLAR:

1. Elov B., Alayev R., Aloyev N. (2024). Tematik modellashtirishning zamonaviy usullari. *Digital transformation and artificial intelligence*, *2*(1), 8–16. Retrieved from https://dtai.tsue.uz/index.php/dtai/article/view/v2i12
2. Elov.B., Alayev N. Matnlarini tematik modellashtirish va tasniflash usullari. barqarorlik va yetakchi tadqiqotlar onlayn ilmiy-amaliy jurnali. Vol. 3 No. 12 (2023). 263-276
3. Elov B., Aloyev N., Yuldashev A. (2023). SVD va NMF metodlari orqali tematik modellashtirish. O`zbekiston: til va madaniyat (Kompyuter lingvistikasi), 2023, 2(6). 55-66
4. Alghamdi, R. and Alfalqi, K. (2015). A Survey of Topic Modeling in Text Mining. *International Journal of Advanced Computer Science and Applications*, *6*(1). https://doi.org/10.14569/ijacsa.2015.060121

5. Tao, R., Wei, Y. and Yang, T. (2021). Metaphor Analysis Method Based on Latent Semantic Analysis. *Journal of Donghua University (English Edition)*, *38*(1). https://doi.org/10.19884/j.1672-5220.202010087

6. Darmalaksana, W., Slamet, C., Zulfikar, W. B., Fadillah, I. F., Maylawati, D. S. adillah and Ali, H. (2020). Latent semantic analysis and cosine similarity for hadith search engine. *Telkomnika (Telecommunication Computing Electronics and Control)*, *18*(1). https://doi.org/10.12928/TELKOMNIKA.V18I1.14874

7. Ke, Z. T. and Wang, M. (2022). Using SVD for Topic Modeling. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2022.2123813

8. Churchill, R. and Singh, L. (2022). The Evolution of Topic Modeling. *ACM Computing Surveys*, *54*(10). https://doi.org/10.1145/3507900

9. Kherwa, P. and Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, *7*(24). https://doi.org/10.4108/eai.13-7-2018.159623