ISSN 2181-922X

TIL VA MADANIYAT

UZBEKISTAN

LANGUAGE & CULTURE

O‘ZBEKISTON

2024 Vol. 3

tsuull.uz
uzlc.tsuull.uz

ISSN 2181-922X

O‘ZBEKISTON:

TIL VA MADANIYAT

UZBEKISTAN:

LANGUAGE AND CULTURE

2024 Vol. 3

tsuull.uz

uzlc.tsuull.uz

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti

“O‘zbekiston: til va madaniyat” jurnali – lingvistika, tarix, adabiyot, tarjimashunoslik, san’at, etnografiya, falsafa, antropologiya va ijtimoiy tadqiqotlarni o‘rganish kabi sohalarni qamrab olgan akademik jurnal.

Jurnal bir yilda to‘rt marta chop etiladi.

Jurnalning maqsadi – ko‘rsatilgan sohalarga oid dolzarb mavzulardagi bahs-munozaraga undaydigan, yangi, innovatsion g'oyalarga boy, o‘z konsepsiyasiga ega bo‘lgan tadqiqotlarni nashr etishdir.

Ingliz, rus va o‘zbek tillaridagi, shuningdek, boshqa turkiy tillarda yozilgan maqolalar qabul qilinadi. Iqtisodiy tahlillar hamda siyosatga oid maqolalar e’lon qilinmaydi.

Jurnalda kitoblarga yozilgan taqrizlar, adabiyotlar sharhi, konferensiyalar hisobotlari va tadqiqot loyihalari natijalari ham e’lon qilinadi. Mualliflar fikri tahririyat nuqtayi nazaridan farq qilishi mumkin.

Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.

O‘zbekiston, Toshkent, Yakkasaroy tumani, Yusuf Xos Hojib ko‘chasi, 103.

Email: uzlangcult@gmail.com
Website: www.uzlc.tsuull.uz

"Uzbekistan: Language and Culture" is an academic journal that publishes works in the field of linguistics, history, literature, translation studies, arts, ethnography, philosophy, anthropology and social studies.

The journal is published four times a year.

The purpose of the journal is to publish the results of the latest research that are rich in new, innovative ideas and has its own concept, which stimulates debate on topical issues in these areas.

The language of articles can be English, Russian and Uzbek. Other Turkic languages are also welcome. We do not publish economic analyses or political articles.

In addition to research articles, the journal announces book and literary work reviews, conference reports and research project results.

The authors' ideas may differ from those of the editors'.

Alisher Navoʻi Tashkent State University of the Uzbek Language and Literature.

103, Yusuf Khos Hojib, Yakkasaray, Tashkent, Uzbekistan.

Email: uzlangcult@gmail.com
Website: www.uzlc.tsuull.uz

MUNDARIJA

Lingvistika

Adabiyotshunoslik

Fan. Ta'lim. Metodika

Tarix. Manbashunoslik

CONTENT

Linguistics

Literature

Science. Education. Methodology

History. Source studies

ISSN 2181-922X 6 — 21

LINGVISTIKA

LINGUISTICS

O‘zbek tilidagi mantlarni avtomatik morfologik tahlil qilishda lemmatizatsiya va stemming jarayoni

Manzura Abjalova[1]

Abstrakt

Tabiiy tilni qayta ishlash sohasida grafematik tahlil (tokenizatsiya), morfologik tahlil (lemmatizatsiya va stemming), sintaktik tahlil (parsing) va semantik tahlil bosqichlari NLPning deyarli barcha yo‘nalishlari uchun muhim hisoblanadi. Raqamli texnologiya uchun qayta ishlangan tabiiy tildan ko‘pgina dasturiy ta’minotlar yaratish mumkin. NLPda morfologik tahlilning lemmatizatsiya va stemming texnologiyalari barcha tillar uchun xos bo‘lib, ular so‘zshakllarni lug‘atdagi normal shaklini aniqlab beradi. Lemmatizatsiya va stemming vazifasi bir xil bo‘lsa-da, natijani chiqarish jihatidan ular farqlanadi. Tezkor jarayon sifatida stemming qiymatli bo‘lsa, aniq lingvistik natijani berishi jihatidan lemmatizatsiya muhim sanaladi. Fleksiyani aniqlash xususiyati bilan lemmatizatsiya flektiv tillar uchun mo‘ljallangan bo‘lsa-da, hozirda agglyutinativ tillar uchun ham qo‘llaniladi. O‘zbek tilini qayta ishlashda har ikki texnologiya ham muhim hisoblanadi. Mazkur maqolada lemmatizatsiya va stemmingning o‘xshash va farqli jihatlari, o‘zbek tilida har ikki texnologiyaning qo‘llanishi, “morfologik tahlil” terminining NLP va o‘zbek tilshunosligida farqlanishi yoritib berilgan.

Kalit so‘zlar: *o‘zbek tili, morfologik tahlil, tabiiy tilni qayta ishlash, NLP, lemmatizatsiya, stemming, ma’lumot olish texnologiyasi, tovush o‘zgarishlari, lemma, stem, normal shakl, lug‘at shakli, asos, qoidalarga asoslangan metod, lug‘atli metod, lug‘atsiz metod, stoxastik metod.*

Kirish

Hozirda kundalik turmushda va ko‘plab sohalarda NLP (Natural Language Processing) ishlanmalariga, ayniqsa, ijtimoiy

[1] *Abjalova Manzura* – filologiya fanlari doktori, professor v. b., Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti.
E-mail: abjalovamanzura@navoiy-uni.uz
ORCID ID: 0000-0002-1927-2669
Iqtibos uchun: Abjalova, M. A. 2024. “O‘zbek tilidagi mantlarni avtomatik morfologik tahlil qilishda lemmatizatsiya va stemming jarayoni”. *O‘zbekiston: til va madaniyat* 3: 6 – 21.

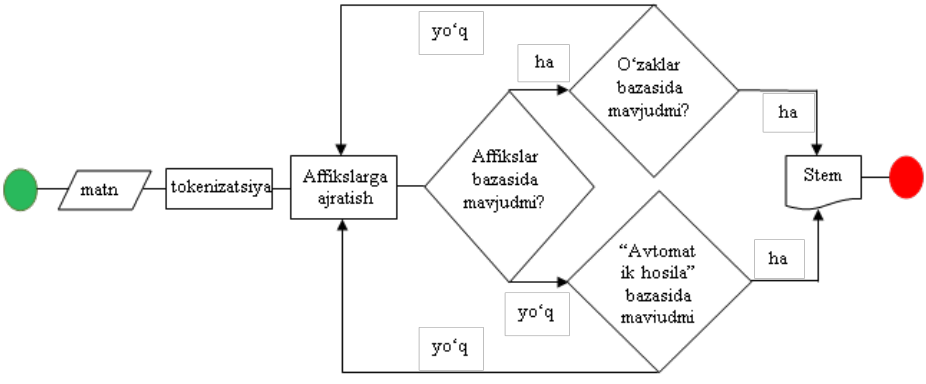
himoyada birinchi o‘rinda ko‘zi ojiz va boshqa jihatlardan imkoniyati cheklanganlar uchun nutqiy sintezator hamda nutqni tanish dasturiy ta’minotlariga ehtiyoj katta. Bunday dastur va tizimlar xaridorgirligini ta’minlash, foydalanuvchilar ishonchiga sazovor bo‘lish uchun ham tabiiy til to‘liq qayta ishlanishi kerak. NLP nomi bilan mashhur tabiiy tilni qayta ishlash sohasi ancha ommalashgan bo‘lib, sun’iy intellekt bilan birga talabgir yo‘nalishlariga ega. Tabiiy tilni qayta ishlash – sun’iy intellekt, til bilimlari va mashinali o‘qitish uchligi integratsiyasiga asoslangan, insoniyat foydalanadigan tabiiy muloqot tili bilan mashina (kompyuter texnologiyalari) o‘rtasida aloqani ta’minlashga qaratilgan tadqiqot sohasi hisoblanib, ilmiy hamda ijtimoiy hayotda faoliyatni ancha qulaylashtirib bermoqda. Jumladan, matnni ovozli matnga aylantirib beruvchi **nutqiy sintezator**lar (asosan, ko‘zi ojizlar uchun zarur imkoniyat), og‘zaki matnni yozma matnga aylantiruvchi **nutqni tanish** dasturlari (jismoniy imkoniyati cheklanganlar va sud, ilmiy seminar hamda majlislar uchun zaruriy vosita), tezkor **avtomatik tarjima** (tarjima qilish imkoniyatiga ega bo‘lmagan va tarjima jarayonida so‘z tanlashda vaqtni tejash uchun eng yaxshi imkoniyat), matndagi emotsiyani aniqlash tizimi – **sentiment tahlil** (asosan, marketing, targetologiya, nizoli jarayonlar uchun zarur vosita), **matnlarni qayta ishlash** (matn bilan ishlash jarayoni uchun), **ma’lumotlarni olish, qidiruv tizimi, savol-javob tizimlari** muhim dasturiy ta’minotlardan hisoblanadi. Ushbu maqolaning asosiy maqsadi NLPda o‘zbek tili lemmatizatsiyasi va stemminggiga xos xususiyatlarni yoritib berishdan iborat.

Manbalar tahlili

NLPda lemmatizatsiya va stemming ko‘p tadqiq etilgan masala bo‘lib, har bir tilning xususiyati bo‘yicha bazaning ishlanishi talab etiladi. Ammo bu ikki texnologiyaning umumiy vazifasi so‘zshakldagi affikslarni kesish hisoblanadi. Ingliz [Tomlinson 2003, 286-300], [Balakrishnan and Lloyd-Yemoh 2014], rus [Саввина, Саввин], nemis [Nicolai and Kondrak 2016], arab tili [Zeroual, Abdelhak 2016, 109–114], [Freihat and others 2018], ispan [Zenón Hernández-Figueroa and others 2013], hind tili [Kasthuri et al. 2014] va hatto, urdu tili [Jabbar, Abdul et al. 2016] lemmatizatsiyasi va stemminggi chuqur tadqiq etilgan. O‘zbek tili bo‘yicha ham bir qancha ishlar qilingan [Bakayev 2021], [Sharipov, Sobirov 2022], [Xusainova 2022].

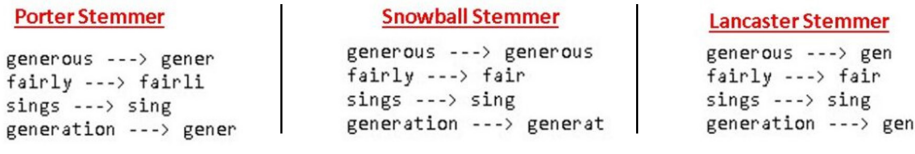
Stemming – morfologik tahlildagi oson va tezkor texnologiya:

affikslar o‘zakka qadar kesib olinadi, xolos. Algoritmi ham oddiy: so‘zshakldagi affiksni kesadi va uni affikslar bazasidan qidirib ko‘radi, “ha”, ya’ni kesilgan affiks mavjud bo‘lsa, uni eslab, qolgan qismni so‘zlar bazasi va “avtomatik hosila” bazasidan qidirib ko‘radi, “ha” bo‘lsa, u “stem” deb qabul qilinadi, “yo‘q” bo‘lsa, affiksni tekshirib, kesishda davom etadi, “ha” javobi chiqsa, shu grammatik ko‘rsatkichni kesib oladi. Bu jarayon so‘zlar bazasida yoki “avtomatik hosila”da leksik birlik topilguncha davom etadi. (1-shakl). Stemming jarayoni natijasida qolgan asos (root) stem deyiladi.



1-shakl. Stemming jarayoni arxitekturasi.

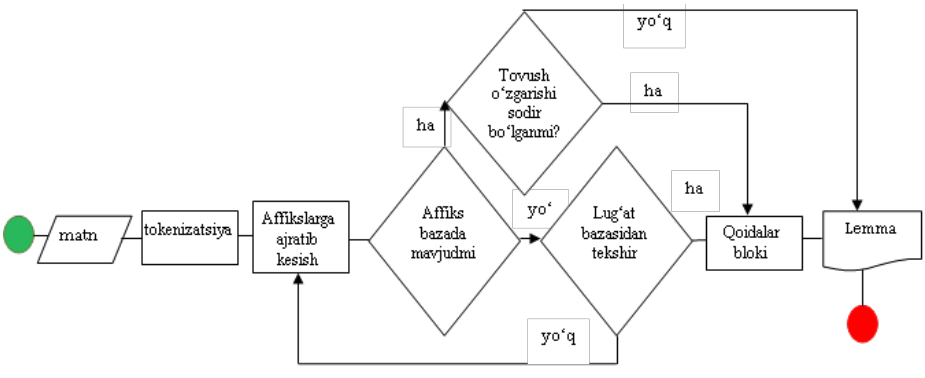
Ingliz tili uchun Porter, Snowball va Lancaster Stemmer dasturiy ta’minotlari ishlab chiqilgan. Ulardan Snowball Stemmer nisbatan aniqlikka ega (2-shakl). Ba’zi manbalarda ingliz tilidagi stemming jarayonida o‘zakkacha affikslar kesilishi ta’kidlanadi, linvistik termin mohiyatini e’tiborga olganda, o‘zbek tilidagi o‘zak aniqlangunga qadar kesish protsessi lemmatizatsiyaga xos hisoblanadi. Chunki stemmingda asosiy maqsad affikslarni tezkorlikda kesish hisoblansa, lemmatizatsiyada affikslarni kesib, o‘zakda ro‘y bergan o‘zgarishlarni asl holatiga keltirish hisoblanadi va bu natija o‘zakni topish deb ataladi.



2-shakl. Porter stemmer, Snowball stemmer va Lancaster stemmer qiyosi.

Lemmatizatsiya algoritmi esa sal murakkablikka ega: so‘zshakldan affiks kesiladi, qolgan qism asoslar bazasidan tekshiriladi, 90 % o‘xshashlikka ega asos chiqsa, u asos deb olinadi, qoidalarga asoslangan metod yordamida fleksiyaning bartaraf

etilishi natijasida qolgan 10 % aniqlik ham tiklanadi. Agar shartli blokda “yo‘q” javobi chiqsa, affikslar bazasidagi affiksga muvofiq qolgan affikslar ham kesiladi, bu jarayon shartli blokda “ha” javobi chiqqunga qadar davom etadi (3-shakl). Lemmatizatsiya jarayoni natijasida qolgan qism lemma deyiladi va u so‘zshaklning bazaviy qismi, ya’ni lug‘atda berilgan normal shakli hisoblanadi (o‘zak emas). Ingliz tili uchun Prinston universitetida yaratilgan WordNet lemmatayzer mavjud.



3-shakl. Lemmatizatsiya jarayoni arxitekturasi.

Lemmatizatsiya va stemming jarayonlari uchun umumiy vazifa bo‘lsa-da, har ikkisi ham afzallik va kamchiliklariga ega (1-jadval).

1-jadval. Lemmatizatsiya va stemming jarayonlarining afzalliklari va kamchiliklari.

	Lemmatizatsiya	**Stemming**
Afzalliklari	• Fleksiyani tiklaydi. • O‘zakni aniqlaydi	• Jarayon tezkor bajariladi. • Lug‘at talab qilinmaydi.
Kamchiliklari	• Stemingga nisbatan tezligi pastroq. • Algoritmi stemmingga nisbatan ko‘p bosqichli. • Lug‘at talab qilinadi. • Morfologik qoidalar bo‘lishi shart.	• Fleksiyani tiklamaydi
Umumiy xususiyat	so‘zshakldagi affikslarni kesish	

Munozara

Ma’lumki, morfologik tahlil jarayonida tilshunoslikning “morfologiya” bo‘limi birliklari tahlil qilinadi: so‘zshaklning turkumi aniqlanadi, uning morfologik xususiyatlari (otning soni, kelishigi, turlanish turi, fe’llarda zamon turi, sifat darajasi turi kabilar)

ochiladi, grammatik ko‘rsatkichlar turi (masalan, ko‘plik shakli, kelishik turi, zamon shakli, egalik shaklining turi va h.k.) aniqlanadi.

Jahon tilshunosligida morfologik tahlil termini ostida so‘zshakldagi ham yasovchi, ham grammatik affikslar yuzasidan tahlil qilish jarayoni tushuniladi. Masalan, *improvements → improve, changer → change.*

Umuman, morfemalar ikki xil bo‘ladi:

I. O‘zak morfema – so‘zning tub (atash) ma’nosini bildirib, mustaqil qo‘llanadi. Masalan: *savat, kitob, daftar, chizg‘ich* va h.k.

II. Affiksal morfema – o‘zakka (asosga) qo‘shilib, turli ma’nolarni ifodalaydi, so‘zlarni bog‘lash uchun xizmat qiladi.

Affiksal morfema (keyingi o‘rinlarda *affiks*)lar vazifasi va so‘zga qo‘shilib anglatadigan ma’nosiga ko‘ra ikki turga bo‘linadi:

1. So‘z yasovchi affikslar – asosga qo‘shilib, asosdan o‘sib chiquvchi yangi ma’noli so‘z yasaydi: *kitob – kitobxon, kuch – kuchli, ma’no – ma’nosiz, davlat – nodavlat, ek – ekin, iste’mol – iste’molchi* kabi.

2. Shakl yasovchi affikslar. Vazifasi va ma’nolariga ko‘ra ikki guruhga bo‘linadi:

a) lug‘aviy shakl yasovchi affikslar asosga qo‘shilib, ma’noni bir oz o‘zgartiradi. Ular sirasiga quyidagi affikslar kiradi:

– qarashlilik affiksi: *-niki;*

– o‘rin-joy oti affiksi: *-dagi;*

– chegaralash affiksi: *-gacha;*

– ko‘plik affiksi: *-lar;*

– sifat darajasi affikslari: *-roq; -ish, -sh; -imtir, -mtir;*

– fe’lning vazifadosh shakllari affikslari, nisbat va mayl affikslari: *-v, -uv, -moq, -mak; -sh, -ish; -r, ar, -gan, -kan, -qan; -adigan, ayotgan; -b, -ib, -a, -y, -gach, -kach, -qach, -guncha, -kuncha, -quncha, -gancha, -kancha, -qancha; -gani, -kani, -qani*;

– affiksli yuklamalar: *-mi, -chi, -gina, -kina, -qina, -dir, -u, -yu, -da, -a, -ya*, *-mish, -ov, -yov, -mikan;*

– inkor shaklni yasovchi affiks: *-mas, -siz;*

– o‘xshatish-solishtirish affiksi: *-chalik, dek, -day;*

b) sintaktik shakl yasovchi affikslar so‘zlarni bir-biriga bog‘lashda xizmat qiladi. Bularga kelishik, egalik, shaxs-son affikslari kiradi [Abjalova 2020].

O‘zbek tilshunosligida so‘z turkumlari va ularning grammatik kategoriyalarini o‘rganuvchi bo‘lim “Morfologiya”, morfemalarni o‘rganuvchi bo‘lim “Mofemika”, so‘z yasash vazifasiga ega affikslarni o‘rganuvchi bo‘lim nomi “So‘z yasalishi” deb nomlanadi va har

birining o‘z o‘rganish obyekti mavjud. Shu bois ham “morfologik tahlil”da morfologiya bo‘limi asosida faqat so‘z turkumi va uning grammatik ma’nosi, grammatik ma’noni yuzaga keltirgan affiksal birliklar aniqlanadi, xolos. Jahon tilshunoslogidan farqli o‘laroq o‘zbek tili lemmatizatsiyasida lemma sifatida lug‘atdagi normal shakl olinadi (unda yasalish hodisasi tekshirilmaydi), stemmingda esa o‘zak aniqlanadi, ya’ni affikslar bazasi yordamida affiks kesiladi (qanday vazifani bajarishidan qat’i nazar) va o‘zak aniqlanadi. Stemmingdagi bitta kamchilik: o‘zakda ro‘y bergan tovush o‘zgarishi bartaraf etilmaydi. Masalan, “*sanoq*” yasalmasidagi *-q* affiksi kesiladi va “*sano*” (“*sana*” emas) qolgan qism stem hisoblanadi. Shuning uchun o‘zbek tili uchun “stem = o‘zak” qonuniyati to‘g‘ri kelmaydi. Lemmatizatsiyada esa faqat grammatik ko‘rsatkichlar kesilishi e’tiborga olinganda, unda ham “lemma=o‘zak” qonuniyati to‘g‘ri bo‘lmaydi.

LEMMA ≠O‘ZAK≠STEM

Aniq bo‘lganidek, lemma ham, stem ham o‘zbek tilidagi o‘zak terminiga aynan muvofiq bo‘lmaydi (2-jadval), muqobil tarjimasiga ega bo‘lmagani bois mazkur terminlar kalkalab olingani maqsadga muvofiq.

2-jadval. So‘zshakllarning morfologik tahlili (lemmatizatsiya va stemming).

So‘z shakllar	**Stem (affiks kesib tashnalanadi, xolos)**	**Lemma (tovush o‘zgarishi bartaraf etiladi)**
ko‘ylag|**ing**	ko‘ylag	ko‘ylak
og‘z|**i**	og‘z	og‘iz
yurag|**im**	yurag	yurak
so‘rog‘|**i**	so‘rog‘	so‘roq
tarog‘|**ing**	tarog‘	taroq
taroq|**lari**	taroq	taroq
shahr|**im**	shahr	shahar
qishlog‘|**imiz**	qishlog‘	qishloq
bulog‘|**i**	bulog‘	buloq
tuprog‘|**i**	tuprog‘	tuproq
qovog‘|**i**	qovog‘	qovoq
uzog‘|**i**	uzog‘	uzoq
kapalag|b	kapalag	kapalak
singl|**ing**	singl	singil

Metodlar

O‘zbek tili ingliz tili morfologiyasiga nisbatan qiyoslanganda morfologiyasi boy til hisoblansa, rus va arab tillari morfologiyasiga nisbatan esa morfologik qonuniyatlari o‘rta qiymatli sanaladi. Lemmatizatsiya va stemming jarayonlari NLPda morfologik tahlil bosqichining eng zarur texnologiyasi hisoblanadi. Har ikki jarayonda to‘rt metodga asoslaniladi: 1) qoidalarga asoslangan; 2) lug‘atli metod; 3) lug‘atsiz metod; 4) stoxastik metod.

4.1. O‘zbek tilidagi orfografik qoidalar. Qoidalarga asoslangan metod bo‘yicha o‘zbek tilining elektron formatini shakllantirishda uch xil tovush o‘zgarishi algoritmlashtirildi.

4.1.1. Tovush tushishi. O‘zbek tilida, asosan, ikkinchi bo‘g‘indagi *a, i, u* tovushlari tushadi. Bu hodisa so‘zlarga egalik qo‘shimchasi qo‘shilganida sodir bo‘ladi.

1) 4.1.1.1. A asosning L-1 – harfi, ya’ni S[L-1] harf o‘chiriladi va oxirgi L – harfi o‘chirilgan L-1 – harf o‘rniga qo‘yiladi. Natijada L uzunlikdagi asos L-1 uzunlakka ega bo‘ladi, egalik affiksi qo‘shilgandan keyin N natijaviy so‘zshakl hosil bo‘ladi. Misol: *shahar И* fonetik usul: urg‘u o‘rni almashtirilib yoki fonema o‘zgartirilishi natijasida yangi so‘z hosil qilinadi: yózma (fe’l) – yozmá (sifat), yangí (sifat) – yángi (ravish);

2) affiksatsiya / morfologik usul (leksema+affiks): *gulchi, kitobxon, tilchi, paxtakor* kabi;

3) kompozitsiya / sintaktik usul (leksema+leksema): *asalari, havo rang, olib kelmoq, rahmdil* kabi;

4) konversiya / semantik usul (boshqa turkumga ko‘chish) – $[W_1] \rightarrow [W_2]$: *ko‘k* (osmon) – *ko‘k* (rang);

5) abbreviatsiya (so‘zni qisqartirish): *ToshDO‘TAU, O‘zR, MDH* kabi;

6) takrorlash yoki juftlash usuli: *bipbip* (avtobus), *xola-xola* (o‘yin nomi), *bor-kel* (qatnamoq) kabi.

Lemmatizatsiya va stemming uchun affiksatsiya hamda takrorlash yoki juftlash usullari muhim hisoblanadi. Aynan lemmatizatsiyada tovush o‘zgarishlarini aniqlab, lemmani tiklash uchun aynan normal shakl va affikslar bazasi zarur. Natijada stemming va lemmatizatsiya jarayonida lug‘atli usul qo‘l mehnatiga asoslangani uchun ancha ishonchli hisoblanib qoladi. Juft va takror qo‘llanib yasalgan so‘zlar borki, ularni yangi ma’no hosil qilmaydigan oddiy juft va takror so‘zlardan farqlash semantik tahlil mukammalligini ta’minlaydi. Masalan, *katta-kichik* (hamma), *yo‘l-yo‘l* (ola-bula rang), *es-es* (zo‘rg‘a) so‘zlari juft va takror qo‘llanib, yangi

so‘z hosil qilganligi bois ular shundayligicha lemma hisoblanadi. Shu bois aynan shu usulda hosil qilingan so‘zlar bazasi ham muhim.

4.3. Lug‘atsiz metod stemming jarayoniga xos. Stemmingda qo‘shimchalar kesib olinadi va o‘zakning o‘zi so‘zshakldagi holaticha qoladi. Masalan, ***shahrim*** → *shahr|im* => *shahr+-im,* ***singling*** → *singl|ing* => *singl+-ing,* ***burni*** → burn/*i* => burn+-i; ***sanash*** → *san|a|sh* => san (stem – san; lemmasi “son” bo‘ladi) + -a + (yasovchi affiks) + -sh (shakl o‘zgartiruvchi affiks). Misollardan ko‘rinib turganidek, lug‘atsiz metod stemmizatsiya uchun qulay keladi. Ammo bugungi kunda Machine Learning yo‘nalishida rekkurent neyron tarmoqlar (RNN), ya’ni takrorlashga asoslangan usulda lug‘atsiz metod qo‘llaniladi.

4.4. Stoxastik metod. Ushbu metod modifikatsiya qilingan bo‘lib, unda statistik metod va ehtimollikka asoslangan metod qo‘llanadi. Stoxastik metod til korpusi asosida ishlaydi [Abjalova and others 2024]. U so‘zning turli shakllari va uning lemmasi yoki stemining qo‘llanish chastotasini aniqlash uchun katta hajmdagi korpus materiallarini tahlil qilishga asoslangan. Statistik algoritmlardan foydalanib, har bir so‘zshakl uchun kontekstdan kelib chiqqan holda uning eng ehtimoliy lemmasi (stemi)ni aniqlaydi. Masalan, “*ko‘ragi*” so‘zshaklida lemmasi korpusda qo‘llanish chastotasi yuqori bo‘lgan “*ko‘rak*” so‘zi uning lemmasi ekanligi ehtimol qilib aniqlanadi, **-i** egalik affiksi ekanligi aniq bo‘ladi va bundan “*ko‘rag*” qismi stem ekanligi ma’lum bo‘ladi.

Ushbu to‘rt metod qo‘l mehnati yoki avtomatik jarayonga asoslanadi (4-shakl).

4-shakl. Metodlarning qo‘llanilishi.

Shuningdek, metodlarni qo‘llashda har bir metodning afzalligi va har biridagi kamchiliklar ularni modifikatsiya tarzida qo‘llashga ham sabab bo‘ladi (3-jadval).

3-jadval. Morfologik tahlilda qo‘llanuvchi metodlarning afzalligi va kamchiliklari.

Metod turlari	Afzalliklari	Kamchiliklari
Qoidalarga asoslangan	Natijalar aniq chiqadi, grammatik xususiyatlar hisobga olinadi.	Tabiiy tildagi barcha lingvistik qoidalar qo‘lda qayta ishlanishi kerak.
Lug‘atli metod	Natijalar aniq chiqadi.	Ma’lumotlarni saqlash uchun ko‘p joy zarur.
Lug‘atsiz metod	Bazaga kirmagan leksik birliklar ham tahlil qilinadi.	Kammahsul yoki maxsus toifalar ustida ishlashda noto'g'ri natijalarni beradi.
Stoxastik metod	Qo‘llashda moslashuvchan va universal	Yangi yoki kam qo‘llanilgan birliklar yuzasidan aniq natija chiqmaydi.

Tajriba va natijalar

Yig‘ilgan baza Uztextanalysis matnlarni avtomatik tahlil qilish dasturida sinovdan o‘tkazildi. Uztextanalyzer dasturiy ta’minoti uchun 84 ming lemma jamlandi va 11 ta shakl yasovchi affiks, 83 ta sintaktik vazifador affikslar, 337 ta so‘z yasovchi affiks bazasi shakllantirildi. Har ikki jarayon uchun lug‘aviy shakl yasovchi affikslardan otga xos 17 ta, sifatga xos 8 ta, fe’lga xos 61 ta, songa xos 10 ta, yuklamaga xos 14 ta affiks aniqlandi; 337 ta so‘z yasovchi affikslardan 114 ta ot yasovchi, 117 ta sifat yasovchi, 58 ta fe’l yasovchi, 48 ta ravish yasovchi affikslar aniqlandi va ularning bazasi shakllantirildi (4-jadval).

4-jadval. O‘zbek tilidagi affikslar miqdori.

O‘zbek tilidagi so‘z turkumlari	so‘z yasovchi affikslar (1)	Shakl yasovchi affikslar (2)	Sintaktik vazifador affikslar (3)
Ot (N)	114	17 (I)	18 (II)
Fe’l (V)	58	61	63
Sifat (Adj)	117	8	II
Ravish (Adv)	48	–	II
Son (Num)	–	10	II
Olmosh (Pron)	–	I	II
Bog‘lovchi	–	–	2+II
Ko‘makchi	–	I	II
Yuklama	–	14	–
Modal (Mod)	–		II
	–		II
	–		II
Jami:	**337**	**110**	**83**

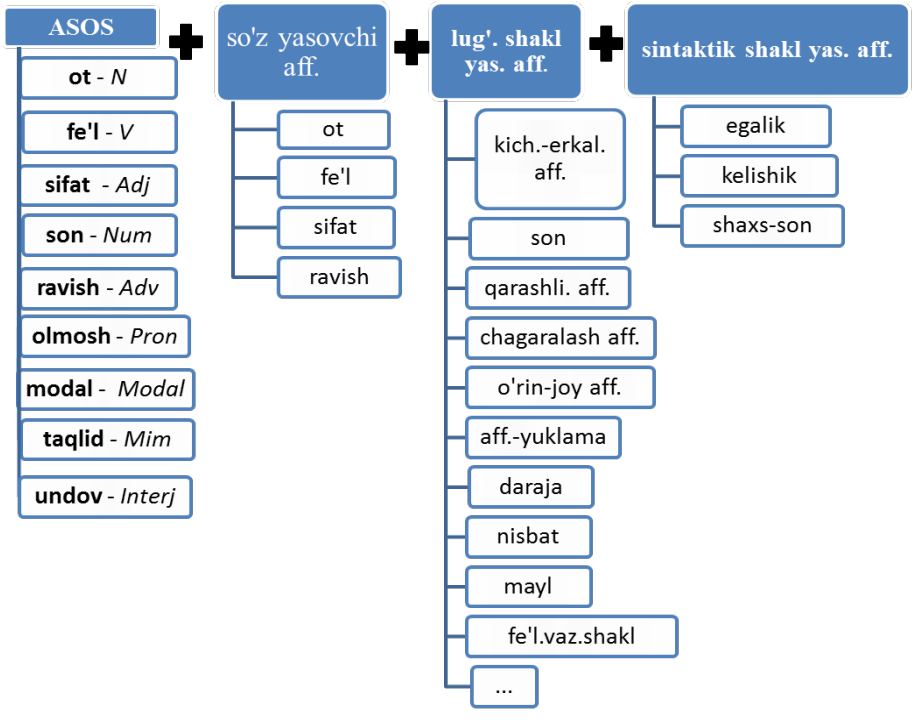
Matnni ATT qiluvchi dasturning algoritmini tuzishda yuqorida keltirilgan shakl yasovchi affikslarning har biriga ma’lum

belgi qo‘yish talab qilinadi. Ramzga ega affikslar Ai va Ei orqali ifodalandi (5-jadval):

5-jadval. Shakl yasovchi affikslar teglanishi.

Ai	Belgi izohi	Ramzi	Lemma turkumi
A1	ko‘plik affiksi	k_a	Ism asosli shakllar
A2	egalik affiksi	e_a	
A3	kelishik affiksi	ke_a	
A5	o‘rin-joy ot affiksi	õ_j	
A6	qarashlilik affiksi	q_a	
A7	chegaralash affiksi	ch_a	
A8	sifat darajasi affikslari	Adj_a	
A9	affiksli yuklamalar	aff.Part	
A10	inkor shaklni hosil qiluvchi aff.	[	
A11	ajratish (-gina)	aj_a	
A12	tegishlilik (-ligi)	teg_a	
A13	o‘xshatish, solishtirish (-chalik, -day, -dek)	õx_a	
E1	fe’lning vazifadosh shakllari affikslari	V_i	Fe’l asosli shakllar
E2	nisbat affikslari	x	
E3	shaxs-son affikslari	y	
E4	zamon affikslari	z	
E5	mayl affikslari	m	

Dasturning LT wordform_set_coord maydonida so‘z turkumlarining grammatik shakllanish tartibi berildi (5-shakl):



Ushbu affikslar quyidagi ketma-ketlikda kombinatsiya hosil qiladi:

Asos – $Lemma_{ism}$ = L_{ism}
S – so‘z yasovchi affiks
L_{ism}= kitob

L_{ism}+ A1	derivation>
L_{ism}+ A2	derivation>
L_{ism}+ A3	derivation>
L_{ism}+ A5	derivation>
L_{ism}+ S+A6	derivation>
L_{ism}+ A7	derivation>
L_{ism}+ A8	derivation>
L_{ism}+ A9	derivation>
L_{ism}+ A10	derivation>
L_{ism}+ A11	derivation>
L_{ism}+ A12	derivation>
L_{ism}+ A13	derivation>
L_{ism}+ A1+ A2	derivation>
L_{ism}+ A1+ A2+A3	derivation>
L_{ism}+ A1+ A2+A9	derivation>
L_{ism}+ A1+ A2+A3+A9	derivation>
L_{ism}+ A1+ A2+A3+A10	derivation>
L_{ism}+ A1+ A2+A3+A10+A9	derivation>
L_{ism}+ A5+ A1+A2+A3+A10+A9	derivation>
L_{ism}+ S+A5+ A1+A2+A10+A9	derivation>

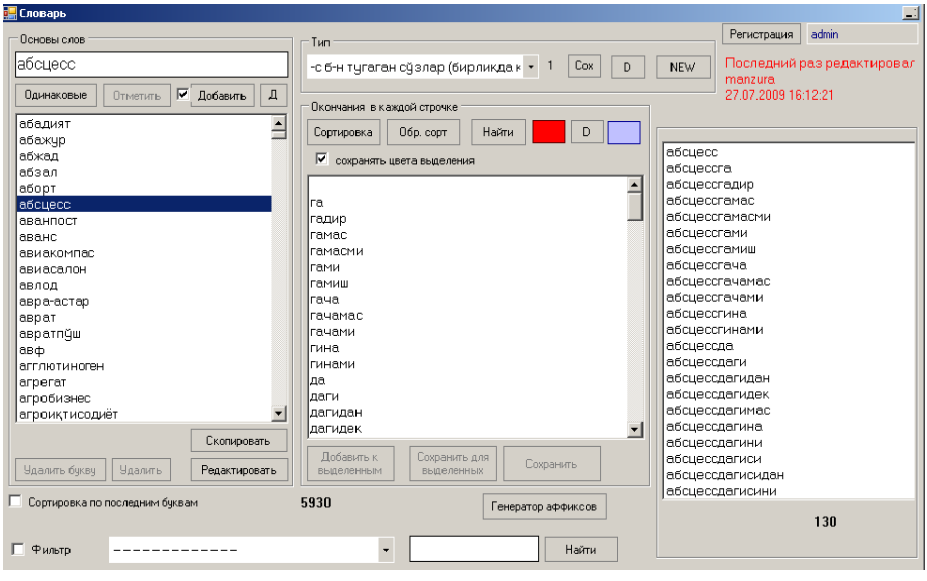
Algoritmlar ketma-ketligi shu tartibda davom ettiriladi.

Uztextanalysis dasturida so‘zshakllar quyidagicha sintez qilindi (6-shakl):

Tanlab olingan leksema: *abssess.*

Affikslar kombinatsiyasi: 129 ta.

Dasturga kiritilish holati: 130 ta (129+ 1(asos/lemma).



6-shakl. Lingvistik bazani tahrirlash oynasi.

Xulosa

O‘zbek tilini qayta ishlashda lemmatizatsiya jarayoni uchun chuqur tilshunoslik bilimlari asosida maxsus lingvistik baza va lug‘at bazasini yaratish muhim. Lemmatizatsiya matn generatsiyasi, matn annotatsiyasi, matnni xulosalash, mashina tarjimasi, nutqni tanish, qidiruv tizimi uchun muhim texnologiya hisoblanadi, stemmingda esa protsedura tezligi yuqori bo‘lgani uchun, asosan, qidiruv tizimi va ma’lumotlarni olish tizimlari uchun muhimdir. O‘zbek tili agglyutinativ til hisoblangani bois unda o‘zak va grammema chegarasi aniq bo‘ladi, so‘zshaklda faqatgina tovush o‘zgarishlari sodir bo‘ladi. Lemmatizatsiya va stemming uchun o‘zbek tilidagi tovush tushishi (1), tovush orttirilishi (2) va tovush almashish (3) hodisalari algoritmlari tuzib olindi. Ushbu tadqiqot natijasida 84000 lemmadan iborat baza, grammatik kategoriya-larga mansub 193 ta affiks bazasi aniqlandi. O‘zbek tilshunosligi va NLPda morfologik tahlil jarayoni farqli ekanligi yoritib berildi va o‘zak bilan lemma va stem terminlari aynan emasligi asoslandi.

Adabiyotlar

Tomlinson, S. 2003. *Lexical and algorithmic stemming compared for 9 European languages with Hummingbird Searchserver at CLEF.* In Proc. Cross-Language Evaluation Forum, 286 – 300.

Balakrishnan, V. and Lloyd-Yemoh E., 2014. *Stemming and lemmatization: a comparison of retrieval performances.* Lect. Notes Softw. Eng. 2 (3): 262.

Саввина, Г.В., Саввин И.В. 2016. “Лемматизация слов русского языка в применении к распознаванию слитной речи”. *Труды СПИИРАН* 1(12). 63 – 73.

Nicolai, G. and Kondrak G., 2016. *Leveraging inflection tables for stemming and lemmatization.*

Zeroual, I., Abdelhak L. 2016. “Arabic Information Retrieval: Stemming or Lemmatization?” 18th IEEE. ACIS International Conference on, 109 – 114.

Freihat, A.A., Abbas, M., Bella, G., & Giunchiglia, F. 2018. “Towards an Optimal Solution to Lemmatization in Arabic”. *International Conference on Arabic Computational Linguistics.*

Zenón Hernández-Figueroa and others. 2013. “Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix’s prominence issue”. *Expert Systems with Applications* 40: 17, 7122-7131, https://doi.org/10.1016/j.eswa.2013.06.056.

Kasthuri, Magesh et al. 2014. *A comprehensive analyze of stemming algorithms for Indian and non-indian languages.*

Jabbar, Abdul et al. 2016. “A survey on Urdu and Urdu like language stemmers and stemming techniques”. *Artificial Intelligence Review* 49: 339 – 373.

Bakayev, I. 2021. “Development of a stemming algorithm based on a linguistic approach for words of the uzbek language”. *International Conference on Scientific, Educational & Humanitarian Advancements.* 195 – 202. Bukhara.

Sharipov, M., Sobirov O. 2022. “Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language”. *The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing*. 1 – 6. Koper, Slovenia.

Xusainova, Z.Y. 2022. “NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash”. *“O‘zbek amaliy filologiyasi istiqbollari” mavzusidagi respublika ilmiy-amaliy konferensiyasi,* 1: 154 – 163. Toshkent.

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Abjalova, M.A. 2020. *Tahrir va tahlil dasturlarining lingvistik modullari*. Toshkent: Nodirabegim.

Abjalova, M., Tukeyev, U., Adilova M., Abduraxmanova M. 2024. *Development and Realization of Bigram Models for Recognizing Homonyms in the Uzbek Language.* ACIIDS 2024, Part II, CCIS 2145 (Communications in Computer and Information Science).

Abjalova, M., Iskandarov O. 2021. “Methods of Tagging Part of Speech of Uzbek Language”. *IEEE – UBMK – 2021: 6th International Conference on Computer Science and Engineering,* 82-85. Ankara – Turkey. DOI: 10.1109/UBMK52708.2021.9558900.

Abjalova, M., Iskandarov O. 2021. “Lingvistik dasturning morfologik tahlil moduli (morfoanalizator)”. *FAN va JAMIYAT Ilmiy-uslubiy jurnal* 2: 49 – 82. Nukus.

Abjalova, M., Iskandarov O. 2023. “Python dasturlash tilida tabiiy tilni qayta ishlash (nlp) tizimlari”. *O‘zbekiston Milliy universiteti (O‘zMU) xabarlari* 1(4/1): 245 – 248.

Abjalova, M. 2015. “Matnlarni avtomatik tahrir va tahlil qilish dasturining lingvistik ta’minoti manbalari”. *Ilm sarchashmalari* 6: 36 – 39. Urganch.

Lemmatization and Stemming Processes in Automatic Morphological Analysis of Uzbek Texts

Manzura Abjalova[1]

Abstract

In the field of natural language processing, the stages of graphematic analysis (tokenization), morphological analysis (lemmatization and stemming), syntactic analysis (parsing), and semantic analysis are important for almost all areas of NLP. Many software programs can be created from natural language that has been remade for digital technology. In NLP, the lemmatization and stemming technologies of morphological analysis are common to all languages, and they determine the normal form of word forms in the dictionary. Although the task of lemmatization and stemming is the same, they differ in terms of output. While stemming is valuable as a quick process, lemmatization is important because it provides a precise linguistic result. Although inflectional lemmatization was originally intended for inflectional languages, it is now also used for agglutinative languages. Both technologies are important in processing the Uzbek language. This article describes the similarities and differences between lemmatization and stemming, the use of both technologies in the Uzbek language, and the difference between the term "morphological analysis" in NLP and Uzbek linguistics.

Key words: *Uzbek language, morphological analysis, natural language processing, NLP, lemmatization, stemming, Information Retrieval techniques, sound changes, lemma, stem, normal form, dictionary form, basis, rule-based method, dictionary method, dictionary-free method, stochastic*

[1] *Manzura Abjalova* – Doctor of Philological Sciences (DSc), Professor, Tashkent State University of Uzbek Language and Literature named after Alisher Navo’i.
E-mail: abjalovamanzura@navoiy-uni.uz
ORCID ID: 0000-0002-1927-2669
For citation: Abjalova, M. 2024. “Lemmatization and stemming processes in automatic morphological analysis of Uzbek texts”. *Uzbekistan: Language and Culture* 3: 6 – 21.

method.

References

Tomlinson, S. 2003. *Lexical and algorithmic stemming compared for 9 European languages with Hummingbird Searchserver at CLEF. In Proc. Cross-Language Evaluation Forum,* 286 – 300.

Balakrishnan, V. and Lloyd-Yemoh E., 2014. *Stemming and lemmatization: a comparison of retrieval performances.* Lect. Notes Softw. Eng. 2 (3): 262.

Savvina, G.V., Savvin I.V. 2016. "Lemmatizatsiya slov russkogo yazika v primenenii k raspoznavaniyu slitnoy rechi". Trudi SPIIRAN. 1(12). 63 – 73.

Nicolai, G. and Kondrak G., 2016*. Leveraging inflection tables for stemming and lemmatization.*

Zeroual, I., Abdelhak L. 2016. "Arabic Information Retrieval: Stemming or Lemmatization?" *18th IEEE. ACIS International Conference on*, 109 – 114.

Freihat, A.A., Abbas, M., Bella, G., & Giunchiglia, F. 2018. "Towards an Optimal Solution to Lemmatization in Arabic". *International Conference on Arabic Computational Linguistics.*

Zenón Hernández-Figueroa and others. 2013. "Automatic syllabification for Spanish using lemmatization and derivation to solve the prefix's prominence issue". *Expert Systems with Applications* 40: 17, 7122-7131, https://doi.org/10.1016/j.eswa.2013.06.056.

Kasthuri, Magesh et al. 2014. A comprehensive analyze of stemming algorithms for Indian and non-indian languages.

Jabbar, Abdul et al. 2016. "A survey on Urdu and Urdu like language stemmers and stemming techniques". *Artificial Intelligence Review* 49: 339 – 373.

Bakayev, I. 2021. "Development of a stemming algorithm based on a linguistic approach for words of the uzbek language". *International Conference on Scientific, Educational & Humanitarian Advancements.* 195 – 202. Bukhara.

Sharipov, M., Sobirov O. 2022. "Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language". *The International Conference and Workshop on Agglutinative Language Technologies as a challenge of Natural Language Processing.* 1 – 6. Koper, Slovenia.

Xusainova, Z.Y. 2022. "NLP: tokenizatsiya, stemming, lemmatizatsiya va nutq qismlarini teglash". *"O'zbek amaliy filologiyasi istiqbollari" mavzusidagi respublika ilmiy-amaliy konferensiyas*i, 1: 154 – 163. Toshkent.

https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

Abjalova, M.A. 2020. *Tahrir va tahlil dasturlarining lingvistik modullari*. Toshkent: Nodirabegim.

Abjalova, M., Tukeyev U., Adilova M., Abduraxmanova M. 2024.

Development and Realization of Bigram Models for Recognizing Homonyms in the Uzbek Language. ACIIDS 2024, Part II, CCIS 2145 (Communications in Computer and Information Science).

Abjalova, M., Iskandarov O. 2021. “Methods of Tagging Part of Speech of Uzbek Language”. I*EEE – UBMK – 2021: 6th International Conference on Computer Science and Engineering,* 82-85. Ankara – Turkey. DOI: 10.1109/UBMK52708.2021.9558900.

Abjalova, M., Iskandarov O. 2021. “Lingvistik dasturning morfologik tahlil moduli (morfoanalizator)”. *FAN va JAMIYAT Ilmiy-uslubiy jurnal* 2: 49 – 82. Nukus.

Abjalova, M., Iskandarov O. 2023. “Python dasturlash tilida tabiiy tilni qayta ishlash (nlp) tizimlari”. *O‘zbekiston Milliy universiteti (O‘zMU) xabarlari* 1(4/1): 245 – 248.

Abjalova, M. 2015. “Matnlarni avtomatik tahrir va tahlil qilish dasturining lingvistik ta’minoti manbalari”. *Ilm sarchashmalari* 6: 36 – 39. Urganch.