PARALLEL KORPUS TUZISHDA STEMMING VA LEMMATIZATSIYANING AHAMIYATI

Xolmonova Iqbola Alisher qizi
iqbolabintualisher@gmail.com
ToshDO‘TAU 2-kurs magistranti

Annotatsiya. Ushbu maqolada korpus tuzish uchun stemming va lemmatizatsiyaning ahamiyati yoritilgan. Stemming va lemmatizatsiyaning bir biridan farqi ochib berilgan va misollar bilan isbotlangan. O‘zbek tilidagi so‘zlar lemmasini aniqlash uchun lemmatizatsiya algoritmi izohlangan. Shu bilan birga, mazkur maqolada lemmatizatsiyaning o‘zbek-turk parallel korpusini yaratishdagi ahamiyati ham ochib berilgan. O‘zbek va turk tilidagi lemmani aniqlash uchun modellar taklif qilingan.

Abstract. This article highlights the importance of stemming and lemmatization for corpus building. The difference between stemming and lemmatization is explained and illustrated with examples. The lemmatization algorithm for determining the lemma of words in the Uzbek language is explained. At the same time, this article also reveals the importance of lemmatization in creating the Uzbek-Turkish parallel corpus. Models for determining the lemma in Uzbek and Turkish languages are proposed.

Аннотация. В этой статье подчеркивается важность стемминга и лемматизации для построения корпуса. Объясняется и иллюстрируется примерами разница между стеммингом и лемматизацией. Объясняется алгоритм лемматизации для определения леммы слов узбекского языка. В то же время в данной статье также раскрывается значение лемматизации при создании узбекско-турецкого параллельного корпуса. Предложены модели определения леммы на узбекском и турецком языках.

Kalit so‘zlar: *lemma, lemmatizatsiya, stem, stemming, model, UzLemmatizator, token, tokenizatsiya*

Kirish

Lemmatizatsiya so‘zning lemma (leksema) shaklini aniqlash va uni flektiv/hosila shakllari o‘rnida ishlatishni o‘z ichiga oladi. Lemmatizatsiya jarayoni orqali til korpusi matnlari mazmunini tahlil qilish aniqligini oshirish mumkin. Shuningdek, lemmatizatsiya jarayoni orqali qidiruv tizimlari algoritmlari korpus matnlari mazmunini tushunishi, uni tartiblash jarayonini amalga oshirishga yordam beradi [Jabeen, 2014: 264].

Lemmatizatsiya odatda quyidagicha ishlaydi:

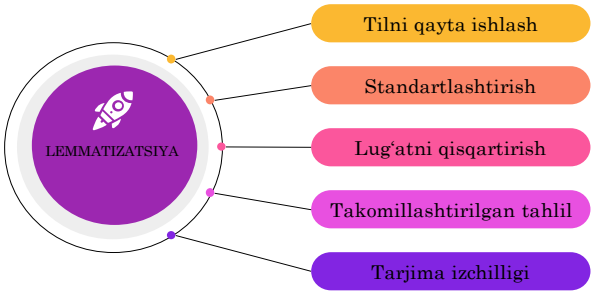
- Tokenizatsiya: Birinchi qadam matnni alohida so‘zlarga yoki tokenlarga bo‘lishdir. Buni turli usullar yordamida amalga oshirish mumkin, masalan, matnni bo‘shliqlar asosida ajratish.

- POS yorlig‘i: Nutq qismlarini teglash har bir tokenga grammatik toifani (masalan, ot, fe’l, sifat va boshqalar) belgilashni o‘z ichiga oladi. Lemmatizatsiya ko‘pincha bu ma’lumotlarga tayanadi, chunki so‘zning asosiy shakli uning gapdagi grammatik roliga bog‘liq bo‘lishi mumkin.
- Lemmatizatsiya: Har bir so‘z tokenizatsiya qilingan va nutq qismi tegi tayinlangandan so‘ng, lemmatizatsiya algoritmi har bir so‘zning lemmasini aniqlash uchun leksika yoki til qoidalaridan foydalanadi. Lemma so‘zning asosiy shakli bo‘lib, u so‘zning ildizi bilan bir xil bo‘lishi shart emas.
- Qoidalarni qo‘llash: Lemmatizatsiya algoritmlari ko‘pincha lingvistik qoidalarga tayanadi.
- Natija: Lemmatizatsiya natijasi so‘zlarning asosiy yoki lug‘at shaklidagi to‘plami bo‘lib, matnning asosiy ma’nosini tahlil qilish va tushunishni osonlashtiradi [Saumyab, 2024: 14].

Korpus yaratish uchun lemmatizatsiya jarayoni muhim hisoblanadi. Chunki lemmatizatsiya so‘zlarni asosiy shaklga standartlashtirishga yordam beradi, bu so‘zlarning turli xil flektiv shakllarini birlashtirib, korpusning murakkabligini kamaytiradi. Ushbu standartlashtirish matnlarni tahlil qilish va solishtirishni osonlashtiradi. So‘zlarni asosiy shakliga qisqartirish orqali lemmatizatsiya hissiyotlarni tahlil qilish, mavzuni modellashtirish va ma’lumotlarni qidirish kabi matn tahlili vazifalarining aniqligini oshirishga yordam beradi. Bu bir xil so‘zning o‘zgarishlari yagona obyekt sifatida ko‘rib chiqilishini ta’minlaydi. Lemmatizatsiya so‘zlarning flektiv shakllarini umumiy tayanch shaklga jamlash orqali umumiy lug‘at hajmini kamaytiradi. Tabiiy tilni qayta ishlash (NLP) ilovalarida lemmatizatsiya ko‘pincha matn ma’lumotlarini keyingi qayta ishlashga tayyorlash uchun ishlatiladi, masalan, xususiyatlarni ajratib olish, mashinani o‘rganish va tilni modellashtirish. Umuman olganda, samarali matn tahlili va NLP ilovalari uchun yanada standartlashtirilgan, izchil va boshqariladigan korpus yaratish uchun lemmatizatsiya zarur.

Lemmatizatsiya jarayoni parallel korpus yaratish uchun ham muhim ahamiyat kasb etadi. Jumladan, turli tillardagi tarjima matnlaridan iborat parallel korpusda lemmatizatsiya manba va maqsadli tillar o‘rtasida mos keladigan so‘zlarni yanada aniqroq moslashtirishga yordam beradi. Lemmatizatsiya so‘zlarni asosiy shakllariga qisqartirish orqali tillardagi ekvivalent so‘zlar va iboralarni yaxshiroq moslashtirishga yordam beradi. Lemmatizatsiya manba tildagi turli flektiv shakllardagi bir xil so‘zning maqsadli tildagi bir xil tayanch shaklga o‘tishini ta’minlaydi. Bu tarjimalarda izchillikni saqlashga va parallel korpus sifatini yaxshilashga yordam beradi. So‘zlarni asosiy shakllariga standartlashtirishda lemmatizatsiya noaniqlikni kamaytirish va ekvivalent so‘z va iboralarning tillar o‘rtasida to‘g‘ri mos kelishini ta’minlash orqali tarjima sifatini yaxshilashga yordam beradi. Bundan tashqari, lemmatizatsiya mashina tarjimasi, tillararo ma’lumotni qidirish va ko‘p tilli tabiiy tilni qayta ishlash kabi tahlil vazifalarining aniqligini oshirishi mumkin. Umuman olganda, lemmatizatsiya turli xil ko‘p tilli matnlarni

qayta ishlash va tarjima dasturlari uchun zarur bo‘lgan yuqori sifatli, izchil va moslashtirilgan parallel korpusni yaratishda hal qiluvchi rol o‘ynaydi.



1-rasm. Lemmatizatsiyaning parallel korpus uchun ahamiyati

Asosiy qism. Lemmatizatsiya jarayoni korpus funksionalligini va tahlil samaradorligini oshirish uchun foydali jarayon hisoblanadi [Xusainova, 2023: 2].

Lemmatizatsiyani korpusda to‘g‘ri qo‘llash uchun uni stemmingdan farqlash lozim. Lemmatizatsiya va stemming − tabiiy tilni qayta ishlash (NLP) da tahlil murakkabligini kamaytirish maqsadida ishlatiladigan ikkita usul. Bu ikkala usul ham so‘zlarni asosiy shakllariga qisqartirishdan iborat, ammo ular bu amalni bajarish maqsadi va unga erishish usullari bilan farqlanadi [Elov, 2023: 46].

Stemming asosan so‘zlardan affikslarni ajratib, faqat asosiy shaklni qoldiradi [Manning, 1999: 134]. Stemming va lemmatizatsiya ikkala usul ham so‘zlarni asosiy yoki ildiz shakliga qisqartirish uchun tabiiy tilni qayta ishlashda qo‘llaniladi. Biroq, ular o‘zlarining yondashuvlari va ular ta’minlaydigan normalizatsiya darajasida farqlanadi. Quyidagi jadvallarda stemming va lemmatizatsiya farqlarini ko‘rib chiqamiz:

STEMMING	LEMMATIZATSIYA
Stemming - bu qo‘shimcha va old qo‘shimchalarni olib tashlash orqali so‘zlarni ildiz shakliga keltirish jarayonidir. Bu oddiy qoidalar va evristika yordamida, so‘zning kontekstini hisobga olmagan holda amalga oshiriladi	Lemmatizatsiya, aksincha, so‘z kontekstini ko‘rib chiqishda so‘zlarni ularning asosiga yoki lug‘at shakliga (lemma) qisqartirishni o‘z ichiga oladi.
Stemmingda so‘zlar har doim ham haqiqiy so‘zlar bo‘lavermaydi, chunki jarayon so‘z o‘zagini olish uchun affikslarni olib tashlashga qaratilgan.	Lemmatizatsiya so‘zning asos yoki lug‘at shaklini to‘g‘ri qaytarish uchun lug‘at tahlili va morfologik tahlildan foydalanadi.

Agglyutinativ tillarda ko‘pincha ildizga qo‘shila oladigan qo‘shimcha va old qo‘shimchalar ko‘p bo‘lganligi sababli, stemming o‘zaklarning haqiqiy so‘z bo‘lmasligi va so‘zning asosiy ma’nosini to‘liq anglatmasligi mumkin.	Agglyutinativ tillarda lemmatizatsiya so‘zning asos shaklini to‘g‘ri qaytarish uchun turli affikslar va ularning ma’nolarini hisobga oladi. Bu lemmalashtirilgan shaklning haqiqiy so‘z bo‘lishini va mo‘ljallangan ma’noni saqlab qolishini ta’minlaydi.

1-jadval. Stemming va lemmatizatsiyaning farqlari

STEMMING	LEMMATIZATSIYA
taroq → taro	taroq → tara
maktabim → maktab	maktabim → maktab
mening → me	mening → men
shunday → shun	shunday → shu
burni → bur	burni → burun
oltovlon → olt	oltovlon → olti

2-jadval. Stemming va lemmatizatsiyaga misollar

Umuman olganda, stemming oddiy va tezkor usul bo‘lib, har doim ham haqiqiy so‘zlarni keltirib chiqarmasligi mumkin, lemmatizatsiya esa so‘z kontekstini hisobga olgan holda haqiqiy so‘zlarni qaytaradigan murakkabroq jarayondir. Agglyutinativ tillarda ya’ni so‘zlar ildizga affikslar qo‘shilishi bilan hosil bo‘lgan tillarda o‘zak va lemmatizatsiya o‘rtasidagi farqlar alohida ahamiyatga ega bo‘ladi. Bunda affikslarning ko‘pligi sababli so‘z shakllarining to‘liq murakkabligini anglash qiyin bo‘lishi mumkin, lemmatizatsiya esa so‘z morfologiyasining nozik tomonlarini ko‘rib chiqish va kontekstda to‘g‘ri asos shakllarini ta’minlash uchun ko‘proq mos keladi.

So‘zlarning tarkibiy qismlari bilan shug‘ullanadigan morfologik tahlil NLPning asosiy yo‘nalishlaridan biri hisoblanadi. Bugungi kunda morfologik tahlilni yanada samaraliroq amalga oshirish uchun turli usullar va algortimlar ishlab chiqilgan hamda joriy qilingan [Sharipov, 2022: 4]. O‘zbek tilidagi so‘zlar lemmasini aniqlash uchun turli xil algoritmlar taklif qilingan [Elov, 2022: 16].

Quyida Maksud Sharipov va Og‘abek Sobirov tomonidan ishlab chiqilgan o‘zbek tilidagi so‘zning lemmasini aniqlash uchun lemmatizatsiya algoritmi taqdim etilgan (1-chizma).

Lemmatizatsiya algoritmini amalda qo‘llash uchun qadamlar ketma-ketligi:

1-qadam. Dasturga matn kiritiladi;

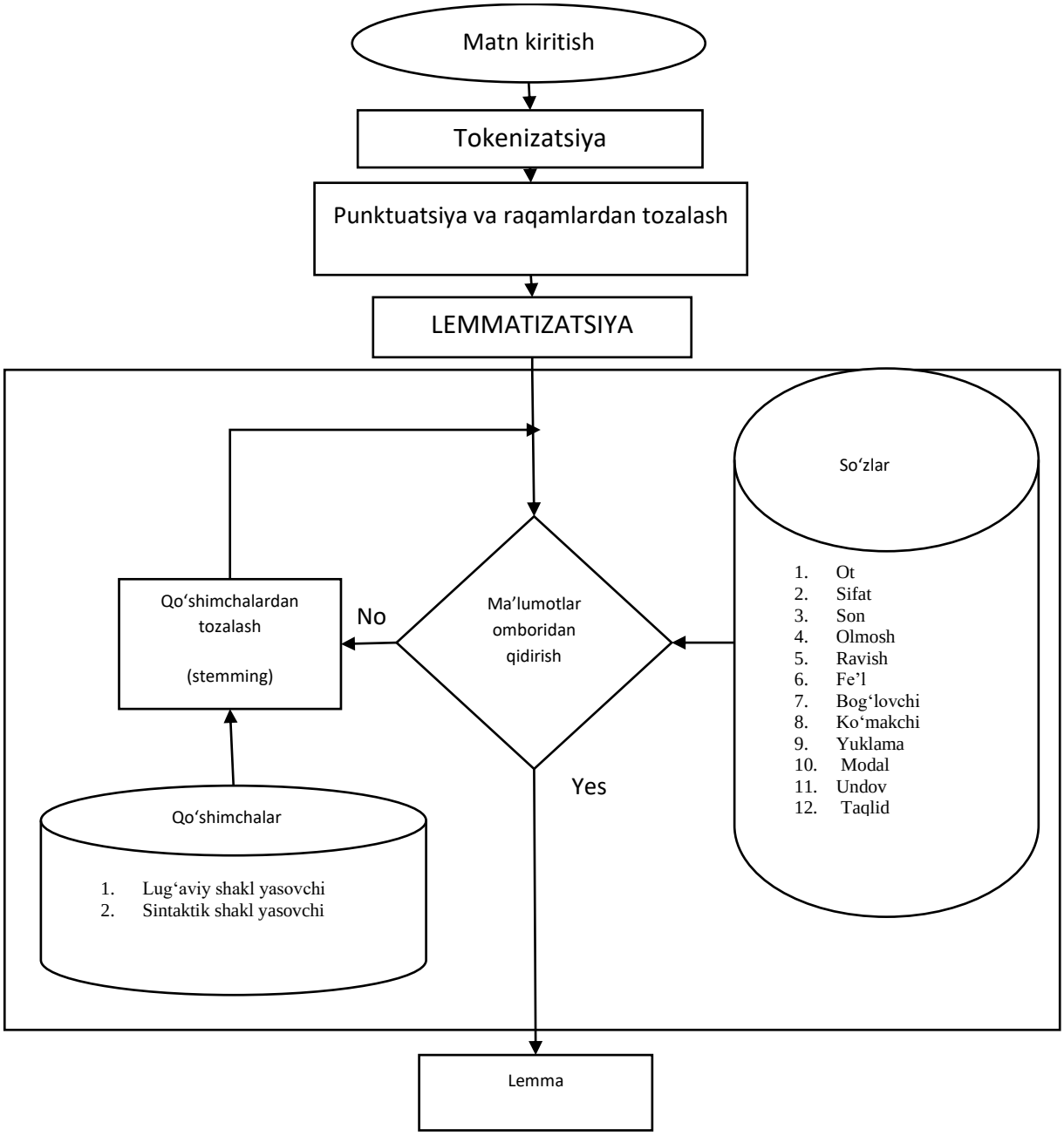
2-qadam. Matn tokenizatsiya yordamida tokenlarga ajratiladi;

3-qadam. Agar matnda tinish belgisi yoki raqamlar bo‘lsa, u olib tashlanadi.

4-qadam. Token “so‘zlar” ma’lumotlar bazasidan qidiriladi, agar bazada mavjud bo‘lsa o‘tkazib yuboriladi. Mavjud bo‘lmasa 5-qadam bajariladi.

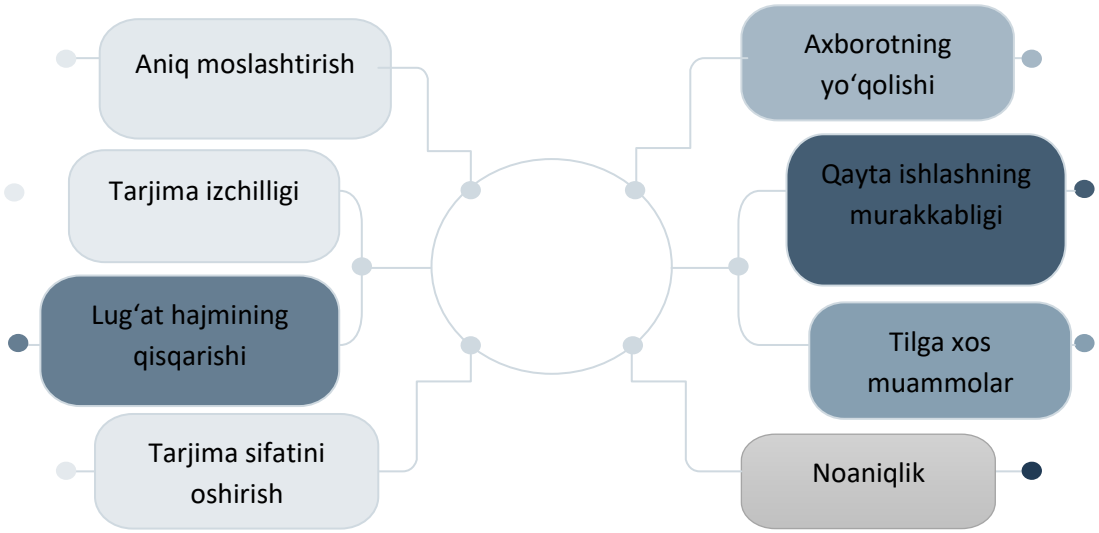
5-qadam. Token qo‘shimchalardan tozalanadi, ya’ni stemming amalga oshiriladi. Bunda qo‘shimchalar bazasidan foydalaniladi.

Parallel korpuslar yaratish uchun stemming va lemmatizatsiya bir necha sabablarga ko‘ra muhimdir: stemming va lemmatizatsiya so‘zlarni parallel korpusda turli tillarda ifodalashda izchillikka erishishga yordam beradi. So‘zlarni ildiz shakllariga qisqartirish orqali, stemming va lemmatizatsiya mashina tarjimasi tizimlarining aniqligini oshirishi mumkin. So‘zlar o‘zlarining asosiy shakllariga tushirilganda, bu boshqa tildagi mos keladigan so‘zlarni aniqroq topishga yordam beradi. Bu tarjima modellarini yaratish va mashinali tarjima tizimlarini o‘qitish uchun parallel korpuslarni moslashtirish uchun zarurdir. Umuman olganda, stemming va lemmatizatsiya parallel korpuslarni mashina tarjimasi va boshqa tabiiy tilni qayta ishlash vazifalariga tayyorlashda hal qiluvchi rol o‘ynaydi.



1-chizma. O‘zbek tilidagi so‘zlar lemmasini aniqlash uchun lemmatizatsiya algoritmi

Quyidagi sxemada parallel korpus uchun lemmatizatsiyaning afzalliklari va kamchiliklarini ko‘rib chiqish mumkin:



2-chizma. Lemmatizatsiyaning afzalliklari va kamchiliklari

Mazkur 2-chizma quyidagi 3-jadvalda izohlangan:

Parallel korpus uchun lemmatizatsiyaning afzalliklari	**Parallel korpus uchun lemmatizatsiyaning kamchiliklari**
Lemmatizatsiya manba va maqsadli tillar o‘rtasida mos keladigan so‘zlarni yanada aniqroq moslashtirishga yordam beradi, bu esa yanada sifatli parallel korpusga olib keladi.	Lemmatizatsiya fleksiyali shakllarda mavjud bo‘lgan o‘ziga xos morfologik yoki sintaktik ma’lumotlarning yo‘qolishiga olib kelishi mumkin, bu tarjima matnlarining boyligi va nuanslariga ta’sir qiladi.
Lemmatizatsiya manba tildagi turli flektiv shakllardagi bir xil so‘zning tarjimalarda izchillikni saqlab, maqsadli tildagi bir xil tayanch shaklga o‘tishini ta’minlaydi.	Lemmatizatsiya jarayoni, ayniqsa, keng ko‘lamli parallel korpus qurish va texnik xizmat ko‘rsatish uchun hisoblash murakkabligi va vaqtni oshishiga sabab bo‘ladi.
Lemmatizatsiya so‘zlarning flektiv shakllarini umumiy tayanch shaklga birlashtirib, umumiy lug‘at hajmini qisqartiradi, korpus bilan ishlash va tahlil qilishni osonlashtiradi.	Lemmatizatsiya algoritmlari barcha tillar uchun bir xil darajada samarali bo‘lmasligi mumkin, chunki ular flektiv yoki agglyutinativ tillar bilan kurash olib borishi mumkin, bu esa moslashtirish va tarjimada noaniqliklarga olib keladi.
So‘zlarni asosiy shakllariga standartlashtirish noaniqlikni kamaytiradi va tillar o‘rtasida to‘g‘ri	Lemmatizatsiya ba’zi so‘zlar yoki iboralardagi noaniqliklarni to‘liq bartaraf etmasligi mumkin, bu parallel

moslashishni ta’minlaydi va tarjima sifatini yaxshilashga yordam beradi.	korpusda moslashtirish va tarjimaning to‘g‘riligiga ta’sir qilishi mumkin.

3-jadval. Lemmatizatsiyaning afzalliklari va kamchiliklari

Umuman olganda, lemmatizatsiya parallel korpus qurish uchun muhim afzalliklarni taqdim etsa-da, u potensial ma’lumot yo‘qotilishi, hisoblash murakkabligi, tilga xos muammolar va noaniqliklarni hal qilish bilan bog‘liq muammolarni ham keltirib chiqaradi. Parallel korpuslarni yaratish va tahlil qilishda lemmatizatsiyani qo‘llashda ushbu omillarni diqqat bilan ko‘rib chiqish kerak.

Bizning ishimizda asosiy maqsad o‘zbek-turk parallel korpusini yaratishdir. Shunung uchun, o‘zbek va turk tilidagi so‘zlarning lemmasini parallel ravishda aniqlash tadqiqotning asosiy maqsadidir. Quyidagi jadvalda o‘zbek va turk tilidagi so‘zlarning lemmasi parallel ravishda aniqlanganligiga misollarni ko‘rishingiz mumkin:

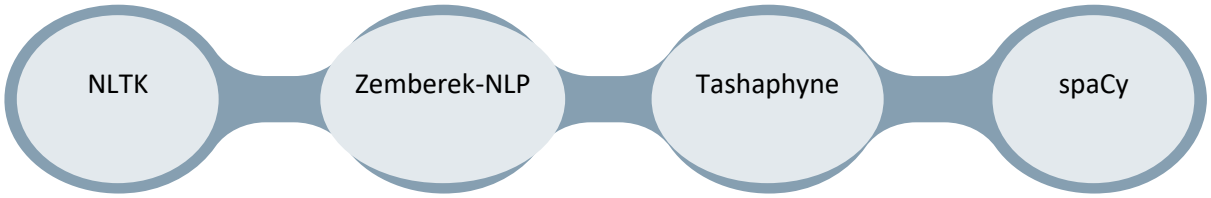
O‘zbekcha so‘zlar	**Turkcha so‘zlar**	**O‘zbekcha va Turkcha so‘zlar lemmasi**
derazadan → deraza o‘yladim → o‘yla nafasimni → nafas issiqligidan → issiq ko‘ryapman → ko‘r meniki → men	pencereden → pencere düşündüm → düşün nefesimi → nefes sıcaklığından → sıcak görüyorum → gör benim → ben	deraza – pencere o‘ylamoq – düşünmek nafas – nefes issiq – sıcak ko‘rmoq − görmek men − ben

4-jadval. O‘zbek va turk tilidagi so‘zlarning lemmasi parallel ravishda aniqlanganligiga misollar

O‘zbek va turk tillarining parallel korpusini yaratish uchun bir nechta lemmatizatsiya modellari variantlari ishlab chiqilgan. Masalan, o‘zbek va turk tillari uchun oldindan tayyorlangan lemmatizatsiya modellari mavjud. Ushbu modellar katta korpuslarda o‘qitilgan va so‘zlarni o‘z tillarida aniq lemmatizatsiya qila oladi. O‘zbek tili uchun o‘zbekcha matn bo‘yicha o‘rgatilgan modellarni, turkcha uchun esa turkcha lemmatizatsiya uchun maxsus ishlab chiqilgan modellarni topish mumkin. Jumladan:

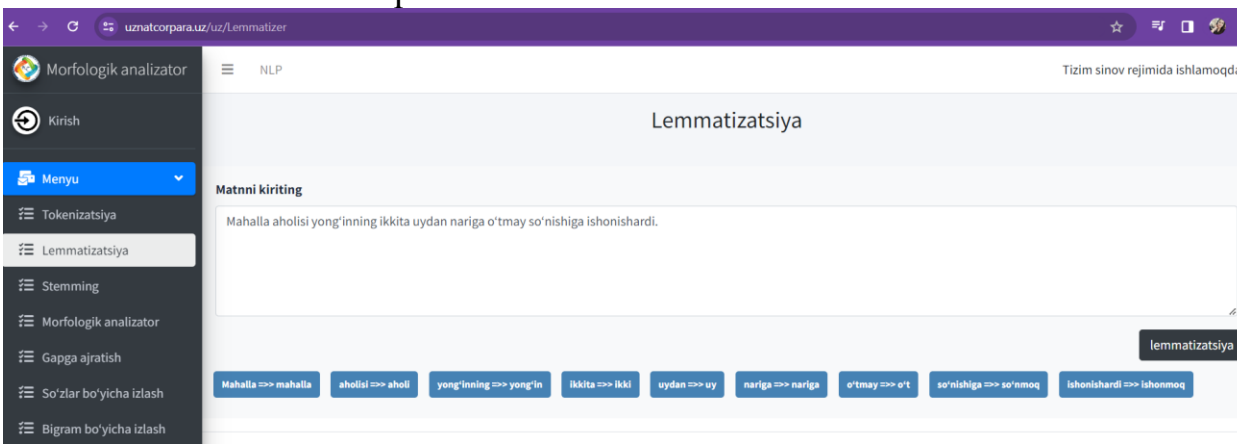
- NLTK (Natural Language Toolkit): NLTK o‘zining nltk.stem.snowball moduli orqali turk tilini lemmatizatsiya qilishni ta’minlaydi.
- spaCy: spaCy – Turk va o‘zbek tillari uchun lemmatizatsiyani qo‘llab-quvvatlaydigan mashhur tabiiy tillarni qayta ishlash kutubxonasi.
- Tashaphyne: Tashaphyne turkiy tillarda, jumladan turk va o‘zbek tillarida tabiiy tillarni qayta ishlash vazifalari uchun maxsus ishlab chiqilgan Python kutubxonasidir. Bu tillar uchun lemmatizatsiya funksiyasini taqdim etadi.

➢ Zemberek-NLP: Zemberek-NLP Java-ga asoslangan tabiiy tillarni qayta ishlash kutubxonasi bo‘lib, turk tili uchun lemmatizatsiyani qo‘llab-quvvatlaydi.



3-chizma. Lemmatizatsiya modellari

Bundan tashqari, o‘zbek tili uchun O‘zbek tili morfologik analizatoridagi UzLemmatizatordan foydalanish mumkin. Alisher Navoiy nomidagi Toshkent davlat o‘zbek tili va adabiyoti universiteti kompyuter lingvistikasi va raqamli texnologiyalar kafedrasi jamoasining tashabbuskorligi natijasida yaratilgan *uznatcorpara.uz*[10] sayti orqali lemmatizatsiya amallarini bajarish imkoni yaratilgan bo‘lib, bu o‘zbek-turk tillari parallel korpusini yaratishda muhim va kerakli unsurlardan bo‘lib xizmat qiladi.



2-rasm. UzLemmatizatordan qidiruv natijasi

Oldindan tayyorlangan modellarning mavjudligi va sifati har xil bo‘lishi mumkinligini yodda tutgan holda, ularni maxsus talablar va foydalanish holati asosida baholash muhimdir. Agar oldindan o‘rgatilgan modellar mavjud bo‘lmasa yoki qoniqarli ishlashni ta’minlamasa, maxsus lemmatizatsiya modellari mashinani o‘rganish texnikasi yordamida ishlab chiqilishi mumkin. Bu har ikki tildagi so‘zlarning asosiy shakllarini to‘g‘ri bashorat qilish uchun o‘zbek va turk matnlarning katta korpusida lemmatizatsiya modelini o‘rgatishdan iborat. O‘zbek va turk tillarining parallel korpusini yaratishda har bir til uchun maxsus ishlab chiqilgan yoki bu tillarda yaxshi ishlashi ko‘rsatilgan lemmatizatsiya modellarini tanlash

[10] https://uznatcorpara.uz/uz/Lemmatizer

muhim ahamiyatga ega. Bundan tashqari, ikkala tildagi ma’lumotlarning kichik namunasi bo‘yicha lemmatizatsiya modellarining to‘g‘riligini baholash parallel korpus sifatini ta’minlash uchun juda muhimdir.

Xulosa qilib shuni aytish mumkinki, lemmatizatsiya ayniqsa turkiy tillar kabi agglyutinativ tillardagi parallel korpuslar uchun bu tillarning murakkab morfologik tuzilishi tufayli muhim ahamiyatga ega. Agglyutinativ tillarda so‘zlar o‘zak yoki o‘zakga old qo‘shimcha, qo‘shimcha va infiks qo‘shish orqali hosil bo‘ladi, natijada har bir so‘z uchun turli xil flektiv shakllar paydo bo‘ladi. Umuman olganda, lemmatizatsiya parallel korpusdagi agglyutinativ tillarning boy morfologiyasini boshqarishda hal qiluvchi rol o‘ynaydi, bu aniqroq moslashtirish, tarjima sifatini yaxshilash, so‘z boyligini oshirish, hisoblash samaradorligi va lingvistik tahlilni yaxshilash imkonini beradi.

Foydalanilgan adabiyotlar:

1. Jabeen, Balakrishnan, Ethel, Khyani. Stemming and Lemmatization: A Comparison of Retrieval Performances. January 2014. Lecture Notes on Software Engineering 2(3):262-267 DOI:10.7763/LNSE.2014.V2.134
2. Saumyab. Stemming vs Lemmatization in NLP: Must-Know Differences. 15 Jan, 2024
3. Xusainova Z. O‘zbek tili milliy korpusi qidiruv tizimini optimallashtirishda lemmatizatsiyadan foydalanish. 2023 Vol. 2 (6) compling.tsuull.uz
4. Elov B.B, Hamroyeva sh.M., Abdullayeva O.X., Xusainova Z.Y., Xudayberganov N.U., 2023: 46
5. Christopher Manning, Hinrich Schütze. Foundations of Statistical Natural Language Processing, MIT Press, 1999
6. Sharipov, Sobirov. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language Urgench State University, Department of Information Technologies, 14, Kh.Alimdjon str, Urgench city, 220100, Uzbekistan
7. Elov B. Hamroyeva Sh. Axmedova X. Methods for creating a morphological analyzer// 14th International Conference on Intellegent human Computer Interaction, IhCI Tashkent.2022, 19-23 October. https://link.springer.com/chapter/10.1007/978-3-031-27199-1_4