

O'ZBEK TILIDAGI SO'ZLARNING LEMMASINI TOPISH

Sobirov Og'abek Ollayor o'g'li,

sobirov.o@urdu.uz,

UrDU, Kompyuter ilmlari kafedrası tayanch doktoranti

Annotatsiya. Lemma bu so'zning asosiy, ya'ni lug'atda uchraydigan shakli. Lemmani topish (lemmalash) – turli shakldagi so'zlarning lemmasini topish jarayoni. Lemmalash jarayoni o'zbek tilining agglyutinatív va morfologik boyligi jihatidan murakkab jarayondir. Ushbu maqolada o'zbek tili so'zlarining lemmasini topish haqida mulohazalar keltirilgan. Bunda o'zbek tili so'z turkumlarining tuzilishiga ko'ra turlari jadval shaklida keltirilib, barcha so'z turkumlarining qo'shimchalar orqali yasalgan holatidan lemmasiga qaytish qoidalari misollar bilan keltirilgan. Qo'shimchalar turlari, soni va so'zlarga qo'shilish qoidalari haqida ham fikrlar yuritilgan. Lemmalash tabiiy til jarayonlarida muhim birinchi qadamlardan biridir. Bu jarayon Ma'lumot olish va Mashina tarjimasi kabi sohalarda foydalaniladi.

Abstract. Lemma is the fundamental, dictionary form of a word. Lemmatization is the process of finding the lemma of words of different forms. The lemmatization is a complex process in terms of the agglutinative and morphological richness of the Uzbek language. This article contains comments on finding the lemma of Uzbek language words. In this, the types of Uzbek language Part-of-Speech according to their structure are presented in the form of a table, and the rules for returning to the lemma of all part-of-speeches from the state formed by affixes are presented with examples. There are also the types, number, and joining to word rules of affixes. Lemmatization is an important first step in Natural Language Processing. This process is used in fields such as Information Retrieval and Machine Translation.

Аннотация. Лемма – это основная словарная форма слова. Лемматизация – это процесс нахождения леммы слов различной формы. Лемматизация представляет собой сложный процесс с точки зрения агглютинативного и морфологического богатства узбекского языка. В данной статье содержатся комментарии по поиску лемм в словах узбекского языка. При этом в виде таблицы представлены типы частей речи узбекского языка по их структуре, а также правила возврата к лемме всех частей речи из состояния, образованного аффиксами, представлены с примерами. Также различаются типы, количество и правила присоединения к словам аффиксов. Лемматизация – важный первый шаг в обработке естественного языка. Этот процесс используется в таких областях, как поиск информации и машинный перевод.

Kalit so'zlar. *Tabiiy Til Jarayonlari (TTJ), lemma, lemmani topish, juft so'zlar, takror so'zlar, qo'shimchalar, o'zbek tili.*

Kompyuter va turli texnologiyalarning rivojlanishi bilan birgalikda, elektron matnlarning hajmi ham ortib bormoqda. Har kuni minglab, hattoki millionlab terabayt ma'lumotlar vujudga kelmoqda. Bunga misol qilib quyidagilarni aytish mumkin, kitoblarning elektron shaklini hosil qilish, yangiliklar saytidagi yangi postlar, ijtimoiy tarmoqlardagi yozishmalar, yana ko'plab holatlarni misol keltirish mumkin. Bularning barchasi kompyuterda matnlarni qayta ishlashga olib keladi, bu jarayon esa allaqachon soha sifatida mavjud bo'lib, ushbu soha Tabiiy Til Jarayonlari (TTJ) [inglizcha: Natural Language Processing (NLP)] deb ataladi.

O'zbek tili uchun ushbu sohada, ya'ni TTJ sohasida bir qancha ishlar amalga oshirilgan va muhim natijalar olingan, hozirda ham davom etmoqda. Quyida shu ishlar bilan tanishib chiqilgan. Ushbu maqolada lemma topish jarayonining qoidaga asoslangan usuli chekli holat mashinasi yordamida ishlab chiqilgan, bunda qo'shimchalar bazasi ham shakllantirilgan [Sharipov, Sobirov, 2022]. Bu maqolada asos topish jarayonining qoidaga asoslangan usuli ishlab chiqilgan, bunda qo'shimchalar bazasi shakllantirilmagan, aksincha qo'shimchalarning 15 ta guruhi shakllantirilib chekli holatlar mashinasi yordamida bajarilgan [Sharipov, Yuldashov, 2022]. Ushbu maqolada o'zbek tili agglyutativ til ekanligi eslatilib, ot va olmosh so'z turkumlarini matnlarda aniqlash haqida so'z yuritilgan, bunda tub ot va olmosh so'zlar bazasi yaratilgan [Sharipov, Mavlanova, Sobirov, 2023]. Keyingi ishlarda matn tahlili va ma'lumot olish kabi TTJ vazifalari uchun o'zbek tilidagi matnlarda nomuhim so'zlarni aniqlash va avtomatik tahlil qilish borasidagi muhim ishlar va ularning natijalari bayon qilingan [Madatov, Bekchanov, 2023]. Tadqiqotlarda matnlarga dastlabki lingvistik ishlov berish jarayonlari yaxshi yoritilgan bo'lib, o'zbek, turk va uyg'ur tillari uchun asos topish, lemma topish va matnlarni belgilash ishlari amalga oshirilib, muhim natijalar bayon qilingan [Elov va boshqalar, 2023]. Ushbu maqolada matnlarni belgilash va atoqli otlarni tanish uchun o'zbek tili korpusi yordamida Yashirin Markov Modelidan foydalanib belgilash algoritmi ishlab chiqilgan, tahlil qilingan va muhim natijalar e'lon qilingan [Elov va boshqalar, 2023]. Matnlarni belgilash borasida yana bir muhim ishlardan birini ushbu maqolada ko'rishimiz mumkin, matnlarni avtomatik tahlil qilish uchun matnlarni belgilash uslubi haqida so'z borgan [Abjalova va boshqalar, 2021]. So'zlarning lemmasini topish, ushbu jarayonning TTJda muhimligi, qo'llanilish sohasi va o'zbek tili uchun bu jarayon bir muncha murakkabligi, turli misollar bilan yoritib berilgan [Xusainova, 2023].

O'zbek tili uchun TTJ sohasining dastlabki muhim lingvistik jarayoni bo'lmish lemmani topish (lemmalash) jarayoni uchun ushbu maqolada muhim ma'lumotlar va ularga misollar keltirilgan. Lemmani topish (lemmalash) [inglizcha: lemmatization] - turli shakldagi so'zlarning lemmasini topish jarayoni. Lemma bu so'zning asosiy ya'ni lug'atda uchraydigan shakli.

O'zbek tilida so'z turkumlarini morfologiya bo'limi o'rganadi. So'z turkumlari jami 12 ta, mustaqil so'z turkumlari: *ot, fe'l, sifat, ravish, son, olmosh*; yordamchi so'z turkumlari: *bog'lovchi, ko'makchi, yuklama*; oraliq yoki alohida

olingan so‘z turkumlari: *undov, taqlid, modal*. Bulardan mustail so‘z turkumlari tuzilishiga ko‘ra 4 xil *sodda, qo‘shma, juft* va *takror* bo‘ladi [1-rasm]. Boshqa so‘z turkumlari bunday tuzilish turlariga ega emas, quyidagi ba’zilaridan tashqari. Undov (qo‘shma: *bor bo‘ling*, juft: *eh-he*, takror: *hay-hay*); Taqlid (juft: *taqur-turuq*, takror: *yalt-yalt*); Modal (birikma: *mening fikrimcha*).

Tuzilish turi		Ot	Fe‘l	Sifat	Ravish	Son	Olmosh
Sodda	Tub	<i>kitob</i>	<i>o‘qi</i>	<i>yashil</i>	<i>tez</i>	<i>besht</i>	<i>men</i>
	Yasama	<i>guldona</i>	<i>ishla</i>	<i>aqlli</i>	<i>eskicha</i>	-	<i>kimdir</i>
Qo‘shma		<i>qorbobo</i>	<i>kashf etdi</i>	<i>mehnatsevar</i>	<i>biroz</i>	<i>o‘n bir</i>	<i>mana bu</i>
Juft		<i>aka-uka</i>	<i>aytdi-qo‘ydi</i>	<i>katta-kichik</i>	<i>qishin-yozin</i>	<i>ikki-uch</i>	<i>men-sen</i>
Takror		<i>qon-qoniga</i>	<i>o‘qib-o‘qib</i>	<i>uzun-uzun</i>	<i>oz-oz</i>	<i>uchta-uchta</i>	<i>o‘sha-o‘sha</i>
		Qisqartma <i>BMT</i>	KFSQ <i>aytib yuribdi</i>	Darajalash <i>juda, sal, eng, tim, nim, liq</i>			
			To‘liqsiz fe‘l <i>edi, emish, ekan, emas</i>				

1-jadval. O‘zbek tili mustaqil so‘z turkumlarining tuzilishiga ko‘ra turlari

Matnda so‘zlar turli shakllarda uchraydi, uning lug‘atda mavjud variantini topish esa aynan lemmalash jarayoni hisoblanadi. Yuqoridagi jadvaldan foydalanib, ona-tilimiz qonun qoidalariga rioya qilgan holda matnda so‘zlarning turli shakldan uning lemmasini topish uchun qoidalarni ko‘rib chiqib, misollar keltiramiz.

1. Sodda so‘z shakllari.

Sodda so‘z shakllari tub va yasama holatda uchraydi, bunda sodda tub so‘zlar hech qanday qo‘shimcha olmaydi va matnlarda shunday holatda uchrashi mumkin (yoki shakl yasovchi qo‘shimchalar qo‘shilgan holatda uchraydi), sodda yasama so‘zlar esa tarkibida so‘z yasovchi qo‘shimchalar olgan holda uchraydi, lemmalash jarayonida esa ushbu so‘z yasovchi qo‘shimchalar olib tashlanmaydi.

Matn	Soʻz	Lemma
<i>Kitobsiz uy – qalbsiz tanaga oʻxshaydi.</i>	<i>Kitobsiz</i>	<i>kitob</i>
<i>Siz oshpaz boʻlib ishlagan boʻlarmidingiz.</i>	<i>ishlagan</i>	<i>ishlamoq</i>
<i>Koʻzingni yumginda uchgacha sana.</i>	<i>uchgacha</i>	<i>uch</i>

2-jadval. Sodda soʻz shakllarini lemmalash jarayoniga misollar

2. Qoʻshma soʻz shakllari.

Qoʻshma soʻz shakllarida *feʼl*, *son* va *olmosh* har doim ajratib yoziladi, *sifat* esa har doim qoʻshib yoziladi, *ot* va *ravish* esa 2 xil holatda: ajratib va qoʻshib ham yoziladi.

Matn	Soʻz	Lemma
<i>Ishxonamiz yaqinda yangicha qiyofaga kiradi.</i>	<i>Ishxonamiz</i>	<i>ishxona</i>
<i>Men sizlarga yaxshi yangilik olib keldim.</i>	<i>olib keldim</i>	<i>olib kelmoq</i>
<i>Gultojixoʻroz urugʻini arzon narxda sotib oling.</i>	<i>Gultojixoʻroz</i>	<i>gultojixoʻroz</i>

3-jadval. Qoʻshma soʻz shakllarini lemmalash jarayoniga misollar

3. Juft soʻz shakllari.

Juft soʻzlar har doim chiziqli bilan qoʻshib yoziladi. Chiziqli va soʻzlar orasida boʻsh joy tashlanmay qoʻshiladi. Ular oʻzaro yaqin munosabatdagi soʻzlardir. Juft soʻzdan qoʻshimcha bilan yasalgan soʻzlar ham chiziqli bilan yoziladi.

Matn	Soʻz	Lemma
<i>Keyin mehr-shafqatdan soʻz ochdi.</i>	<i>mehr- shafqatdan</i>	<i>mehr-shafqat</i>
<i>Imtihon uchun kecha-yu kunduz tayyorlandi.</i>	<i>kecha-yu kunduz</i>	<i>kecha-kunduz</i>
<i>Har kimga oʻn-oʻn beshtadan tegadi.</i>	<i>oʻn-oʻn beshtadan</i>	<i>oʻn-oʻn besh</i>

4-jadval. Juft soʻz shakllarini lemmalash jarayoniga misollar

4. Takror soʻz shakllari.

Takror soʻz shakllari ham juft soʻzlar kabi har doim chiziqli bilan yoziladi. Takror soʻzlar maʼnoni kuchaytirish uchun bitta soʻzni takrorlashdan hosil qilinadi. Lugʻatlarda soʻzlar har doim ham takroriy shakllarda uchramaydi shuning uchun

ularning lemmasi faqat bitta soʻzdan yoki ularning takrorlanishidan iborat boʻlishi mumkin.

Matn	Soʻz	Lemma
<i>Uzoq-uzoqlarga ketib qolgan ekan.</i>	<i>Uzoq-uzoqlarga</i>	<i>uzoq</i>
<i>Eshik oʻzidan-oʻzi ochilib ketdi.</i>	<i>oʻzidan-oʻz</i>	<i>oʻz</i>
<i>Har kimga bitta-bittadan beramiz.</i>	<i>bitta-bittadan</i>	<i>bitta-bitta</i>

5-jadval. Takror soʻz shakllarini lemmalash jarayoniga misollar

Qolgan shakllarda ham oʻzbek tilining tilshunoslik qoidalari asosida lemmalash jarayoni amalga oshiriladi. Bunday shakllar quyidagilardir:

- Otlarda qisqartmalar: *BTMning* – *BMT*
- Feʼllarda KFSQ: *aytib yuribdi* – *aytmoq*
- Toʻliqsiz feʼllar: *borgan ekan* – *bormoq*
- Sifatlarda daraja shakllari: *tim qora* – *qora*

Foydalanilgan adabiyotlar:

1. Sharipov M, Sobirov O. Development of a Rule-Based Lemmatization Algorithm Through Finite State Machine for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 154 – 159.
2. Sharipov M, Yuldashov O. UzbekStemmer: Development of a Rule-Based Stemming Algorithm for Uzbek Language. CEUR Workshop Proceedings, CEUR-WS 2022, 137 – 144.
3. Sharipov M, Mavlonova M, Sobirov O. Oʻzbek tilidagi matnlarda ot va olmosh soʻz turkumlariga tegishli soʻzlarni aniqlovchi algoritmi yaratish. Computer Linguistics: Problems, Solutions, Prospects 2023; 1.
4. Madatov K, Bekchanov S, Vičič J. Automatic Detection of Stop Words for Texts in the Uzbek Language. Informatica (Slovenia) 2023; 47: 143 – 150.
5. Boltayevich EB, Adali E, Mirdjanovna KS, Xolmoʻminovna AO, Yuldashevna XZ, Nizomaddin Uktamboy oʻgʻli X. The Problem of Pos Tagging and Stemming for Agglutinative Languages (Turkish, Uyghur, Uzbek Languages). UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, 2023, 57 – 62.
6. Boltayevich EB, Samariddinovich SS, Mirdjanovna KS, Adali E, Yuldashevna XZ. Pos Tagging of Uzbek Text Using Hidden Markov Model. UBMK 2023 - Proceedings: 8th International Conference on Computer Science and Engineering, 2023, 63 – 68.
7. Abdurashetona AM, Ismailovich IO. Methods of Tagging Part of Speech of Uzbek Language. Proceedings - 6th International Conference on Computer Science and Engineering, UBMK 2021, 2021, 82 – 85.
8. Xusainova Z. Methods of Lemmatizing Various Structured Words in Uzbek Language. CAJLPC 2023; 4: 67–72.