**O‘ZBEKISTON MILLIY UNIVERSITETI XABARLARI, 2020, [1/2]**
**ISSN 2181-7324**

**FILOLOGIYA**
www.uzmuxabarlari.uz
**Social sciences**

***Манзура АБЖАЛОВА,***
*Навоий давлат кончилик институти*
*Илмий-услубий бўлими мудири,*
*филология фанлари бўйича фалсафа доктори (PhD)*
*E-mail: manzura_ok@mail.ru*
***Отабек ИСКАНДАРОВ,***
*дастурчи*
*E-mail: bek@mail.ru*

## MODULE OF MORPHOLOGICAL ANALYSIS OF TEXTS AND ITS ALGORITHM

Abstract

Today various text processing software and linguistic programs are being created. The automatic morphological analysis mainly checks the morphological features of a word and the spelling of words. for this purpose, morphological analysis technologies are created. Among them, we can mention the process of stemming and lemmatization, which today have become more effective technologies. Stemming works with the symbols in the word form that arise from one basis, that is, the process of finding the basics of words, and lemmatization takes into account the shape of the word forms, i.e. inflected endings are discarded, and returns to the basic or dictionary form of the word. As it turns out, the use of stemming technology in automatic analysis of texts in the Uzbek language, which has an agglutinative character, is considered appropriate. This article presents an algorithm of the morphological module for automatic editing and analysis of texts in Uzbek language.

**Keywords:** morphological analysis, module, technology, stemming, lemmatization, model, token.

## МОДУЛЬ МОРФОЛОГИЧЕСКОГО АНАЛИЗА ТЕКСТОВ И ЕГО АЛГОРИТМ

Аннотация

В настоящее время создаются различные программные и лингвистические приложения для обработки текстов. В автоматическом морфологическом анализе в основном проверяется морфологические признаки слова и правописания слов, для этого создан технологии морфологического анализа. Из них можно отметить процесс стемминга и лемматизации, которые сегодня стали более эффективными технологиями. Стемминг работает с символами в словоформе, которые возникают из одной основы, то есть процесс нахождения основы слова, а лемматизация учитывает форму словоформы, то есть отбрасываются флективные окончания и возвращается основная или словарная форма слова. Как выясняется, использование технологии стемминг в автоматическом анализе текстов на узбекском языке, имеющем агглютинативный характер, считается целесообразным. В данной статье приведен алгоритм морфологического модуля программы для автоматического редактирования и анализа текстов на узбекском языке.

**Ключевые слова:** морфологический анализ, модуль, технология, стемминг, лемматизация, модель, токен.

## МАТНЛАРНИ МОРФОЛОГИК ТАҲЛИЛ ҚИЛИШ МОДУЛИ ВА УНИНГ АЛГОРИТМИ

Аннотация

Бугунги кунда матнларга қайта ишлов бериш бўйича турли дастурий таъминотлар ва лингвистик дастурлар яратилмоқда. Автоматик морфологик таҳлилда, асосан, сўзшаклларининг имлосини текшириш кўзда тутилган бўлиб, бунинг учун морфологик таҳлил технологиялари яратилган. Шулардан бугунги кунда самарадор технологияларга айланган стемминг ва лемматизация жараёнини қайд этиш мумкин. Стемминг бир асосдан юзага келган сўз шакллардаги белгилар билан ишлайди, яъни сўзнинг асоси (ўзак)ни топиш жараёни, лемматизация эса бир лемманинг флектив (аффикс қўшилиши натижасида ўзгаришга учраган) шаклини эътиборга олади. Маълум бўладики, агглютинатив табиатга эга ўзбек тилидаги матнларни автоматик таҳлил қилишда стемминг технологиясидан унумли фойдаланиш мақсадга мувофиқ саналади. Мазкур мақолада ўзбек тилидаги матнларни автоматик таҳрир ва таҳлил қилиш дастурининг морфологик модули алгоритми берилди.

**Калит сўзлар:** морфологик таҳлил, модуль, технология, стемминг, лемматизация, модель, токен.

КИРИШ. Ахборот асрига келиб автоматик матн таҳрири ва таҳлили жаҳон тилшунослигининг шиддат билан ривожланаётган соҳасига, шундай вазифага эга дастурнинг ўзи эса беқиёс замонавий иш қуролига айланди. Улар тил материалини ахборот технологиялари дастурлари орқали таҳрир ва таҳлил қилувчи тезкор ва иқтисодий тежамкор усул бўлиш баробарида машина таржимаси сифатини оширишда ҳам муҳим омил саналади.

Морфологик таҳлил (МТ) тарихи қадимги ҳинд тилшуноси Панин билан боғлиқ. У Aṣṭādhyāyī матнидан фойдаланиб санскрит тили морфология-

сининг 3959 та қоидасини яратади[1]. Юнон-рим грамматикаси анъанаси ҳам МТ билан шуғулланиш бўлган[2]. 1859 йилда А.Шлейхер тилшуносликдаги “морфология” терминини ўйлаб топади[3]. Ўтган асрнинг 60-70-йилларига келиб, машинали морфология қамровидаги барча тадқиқотлар машина луғатини яратиш билан бошланган[4].

Бугунги кунда дунё информацион технологиялари ва жаҳон тилшунослигида морфологик таҳлил технологиялари яратилган бўлиб, бугунги кунда уларнинг самарадор турларидан стемминг ва лемматизация жараёнини қайд этиш мумкин. Стемминг бир асосдан юзага келган сўзшакллардаги белгилар билан ишлайди, лемматизация эса бир лемманинг флектив (аффикс қўшилиши натижасида ўзгаришга учраган) шаклини эътиборга олади. Маълум бўладики, агглютинатив табиатга эга ўзбек тилидаги матнларни автоматик таҳлил қилишда стемминг технологиясидан унумли фойдаланиш мақсадга мувофиқ.

АСОСИЙ ҚИСМ. Маълумки, эгалик ва келишик аффикслари билан ўзгариш хусусиятига Ai орқали аффикслар қуйидаги ҳолатда ифодаланади:

| Ai | Белги изоҳи | ID | Лемма туркуми |
|---|---|---|---|
| A1 | кўплик аффикси | k_a | Исм асосли шакллар |
| A2 | эгалик аффикси | e_a | |
| A3 | келишик аффикси | ke_a | |
| A4 | ўрин-жой от аффикси | u_j | |
| A5 | қарашлилик аффикси | q_a | |
| A6 | чегаралаш аффикси | ch_a | |
| A7 | сифат даражаси аффикслари | Adj_a | |
| A8 | аффиксли юкламалар | aff_part | |
| A9 | инкор шаклни ҳосил қилувчи афф. | ink_a | |
| A10 | ажратиш (-гина) | aj_a | |
| A11 | тегишлилик (-лиги) | teg_a | |
| A12 | ўхшатиш, солиштириш (-чалик, -дай, -дек) | us_a | |

```
SELECT sg_form.name
FROM  sg_entry, sg_form, coord_pairs
WHERE sg_entry.name='давлат'
AND sg_form.id_entry=sg_entry.id
AND coord_pairs.id=sg_form.id_dims
AND coord_pairs.str_pairs
LIKE '%эгалик+келишик%'
```

– эгалик ва келишикдаги сўзшакл.

List of derivate: давлатимни / давлатингга / давлатидан / давлатингизнинг / давлатларида / давлатимизни / давлатингизда / давлатнинг ...

Дастурнинг ЛТ wordform_set_coord майдонига эса аффикслар тартиби киритилди. Жумладан, wordform_set_coord(lemma+ˣs_ya +sh_ya + sin_ya). Бу ерда lemma – асос, s_ya – сўз ясовчи аффикс, sh_ya – шакл ҳосил қилувчи аффикс, sin_ya – синтактик шакл ҳосил қилувчи аффикс. “ˣ” белгиси аффикснинг бирикиш тартиби қатъий талаб қилинмайдиган ҳолатни англатади. Шундан сўнг сўзшаклнинг матнда учрайдиган ҳолати намоён бўлади: wordform_refresh( wrd ). Айнан морфологик таҳлилнинг шу жараёни машина таржимасида муҳим ўринга эга.

Маълумки, ҳар қандай дастур алгоритмлар асосида ишлайди, алгоритмлар эса маълумотлар эга сўзлар исмлар атамаси остида бирлаштирилади. Улар таркибига от, сифат, сон, олмош, тақлид сўз, феълнинг сифатдош ва ҳаракат номи киради. Шуни назарда тутиб, тилшуносликдаги турланиш ва бошқа грамматик категорияларнинг сўзларга бирикиб келиш ҳодисаси исмларга мансуб туркумларда кузатилгани боис дастурнинг ЛТни яратишда морфологик синтез исмлар доирасида амалга оширилди. Бунинг учун луғавий шакл ясовчи ва синтактик шакл ҳосил қилувчи аффикслар базаси Accessда яратилди (gr_form). Кўплик шаклига эга ёхуд семантик кўпликка эга сўзлар (асосан, саналмайдиган отлар, мавҳум отлар) sol_CorrNounNumber гуруҳига бирлаштирилди.

Ўзбек тилидаги исмларнинг грамматик шаклланиши учун махсус дастурий қисм (sol_GenIsmForm) яратилди. Бунда 1) грамматик шакллар номи жадвалга махсус белги бериб киритилди (id_entry,iform); 2) id_dims майдонига эга аффикслар ID_Wordra боғланади, натижада List of derivate майдонида сўзшаклни юзага келтиради.

манбаига таянади. Матнларни қайта ишловчи лингвистик дастурлар ҳам миллий тилнинг лингвистик қоидаларига таяниб тузилган алгоритмлар асосида ўз вазифасини бажаради. МТ алгоритмида қуйидаги белгилардан фойдаланилади:

Us – лингвистик таъминотдаги сўзлар базаси, Us="SELECT * FROM 'Us'";

Ys – ўзбек тилидаги барча ясовчи аффикслар базаси, Ys="SELECT * FROM 'Ys'";

Sq – ўзбек тилидаги грамматик категориялар базаси, Sq="SELECT * FROM 'Sq'";

$S_i$ – W матндан ажратиб олинган сўзшакллар, $1 \le i \le n$, n – W матндаги сўзшакллар сони;

$Sq_j$ – Sq базадаги аффикслар, $1 \le j \le m$, m – Sq базадаги аффикслар сони.

Tz – сўзнинг қусурли ёзилганлигини визуал кўрсатувчи ва ёзилган хато сўзга мақбул сўзшакл вариантларини берувчи махсус функция.

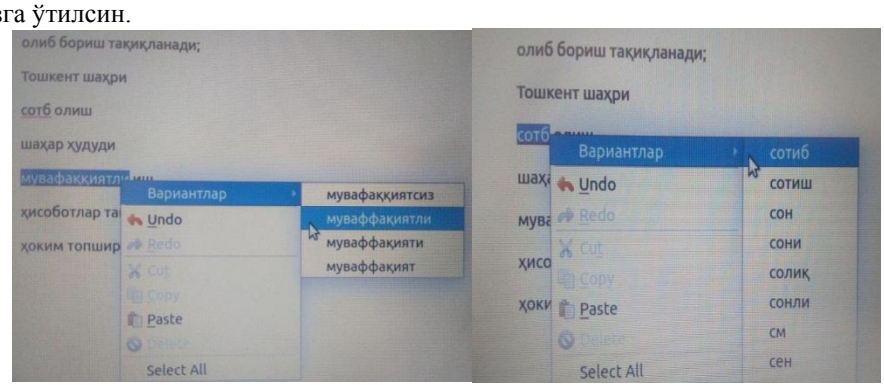
Қуйида гапларни токенларга ажратган ҳолда, лемма бўйича ўзбек тилидаги сўзлар базасидан изланади, топилмаса ўзбек тилидаги барча ясама сўзлар базасидан қидирилади. Асос ёки ясалма базадан топилгач, унга бирикиш эҳтимолидаги аффикслар IDси бўйича ўзбек тилидаги барча

аффикслар базасидан олинади. Демак, таҳлил алгоритми қуйидагича бўлади:

1. $S_i$ даги ҳар бир сўз, Us базасидан излансин. Топилса, кейинги қадамга, акс ҳолда 5-қадамга ўтсин.
2. $S_i$ сўзнинг Us базадаги ID (тартиб рақами) олинсин.
3. $S_i$ сўзнинг ID рақамига тўғри келадиган аффикс $Sq_j$ – Sq базадан излансин.
4. $S_i$+$Sq_j$ тўғри бўлса 10-қадамга ўтилсин, акс ҳолда кейинга қадамга ўтилсин.
5. $S_i$ даги ҳар бир сўз, Ys базасидан излансин. Топилса, кейинги қадамга, акс ҳолда 3-қадамга ўтилсин.
6. $S_i$ сўзнингYs базадаги ID (тартиб рақами) олинсин.
7. $S_i$ сўзнинг ID рақамига тўғри келадиган аффикс $Sq_j$ – Sq базадан излансин.
8. $S_i$+$Sq_j$ тўғри бўлса, 10-қадамга ўтилсин, акс ҳолда кейинги қадамга ўтилсин.
9. Tz фукция амалга оширилсин ва 10-қадамга ўтилсин.
10. Сўз тўғри ёзилганлиги тасдиқлансин ва кейинги $S_{i+1}$ сўзга ўтилсин.

Алгоритм иш тартибини мисол орқали таҳлил қиламиз: истеъмолчиларнинг сўзшакли стемминг қилинади, яъни wordform_set_coord майдонида берилган аффикслар тартиби бўйича кесиб чиқилади: истеъмол/чи/лар/нинг. Шундан сўнг истеъмол лемма имлоси List of Words базасидан текширилади, -чи, -лар, -нинг аффикслари $Sq_j$ – Sq базасидан қидирилади, бунда -лар, -нинг шакл ясовчи ва синтактик муносабат шакли эканлиги аниқланади, -чи ясовчи аффиксни топиш учун Ys базасига мурожаат қилинади. Лемма ва аффикслар-нинг ЛТда мавжудлиги топилгач, сўзнинг тўғри ёзилганлиги тасдиқланади ва кейинги сўзга ўтилади.

Мисолни дастуримиздаги таҳлил жараёни билан бойитамиз. Ubuntu 16.4 (Linux) операцион системаси Python 3 дастурлаш тилида яратилган автоматик таҳрир ва таҳлил дастурининг мор-фологик модули имкониятини қуйидаги тасвирда кўриш мумкин. Тасвирда хато ёзилган ясама сўз ва равишдош билан шаклланган хато сўзнинг дастур томонидан таҳлил қилиниши натижасида уларга тўғри вариантлар таклиф этилиши фрагменти кўрсатилган.




9-расм. Морфологик таҳлил жараёни.

ХУЛОСА. Ўзбекча матнларни таҳрир ва таҳлил қилиш дастурининг морфологик модули ва унинг алгоритмини яратишда қуйидаги хулосаларга келинди:

Автоматик таҳрир ва таҳлил дастурининг лингвистик таъминотини яратиш лингвистик меъёрлар ва алгоритмга эга қоидаларнинг ишлаб чиқилиши, лексикографик манбаларнинг таъминот-га киритилиши, сўзшакллар ва сўзларнинг ўзаро боғланиш моделлари тузилиши билан белгиланади.

Лингвистик процессорни яратишда лингвис-тик модулнинг ўрни ва аҳамияти ўта муҳим. Комп-ьютер лингвистикасида модуль термини дастурий таъминотнинг муайян лингвистик жараёнини қамраган мустақил таркибий қисми сифатида қўлланилади. Шу маънода ўзбек тилидаги матнларни таҳрир ва таҳлил қилувчи дастурнинг морфологик модулида сўзшакллар анализи (сўз шаклдан лексемага қадар таҳлил) ва синтези (лексеманинг грамматик шаклланиши таҳлили жараёни) амалга оширилади.

Ўзбек тилидаги ясовчи аффиксларни статис-тик жиҳатдан ҳисоблаш ва уларнинг лексик базаси ва лемма+ясовчи аффикс моделини яратиш лингвистик таъминотнинг муҳим таркибий қисми ҳисобланади. Компьютер учун аффиксларни сўз ясовчи ва шакл ясовчи аффикслар тарзида алоҳида базаларга ажратиш, моделларини аниқлаш таҳлил жараёнида муҳим аҳамият касб этади.

Лингво-анализ дастури нафақат таҳрир ва таҳлил жараёнини тезлаштирувчи восита, балки ўзбекча матнларни бекусур ёзиш кўникмасини шакллантирувчи инновацион система сифатида муҳим аҳамиятга эга ҳисобланади.

## АДАБИЁТЛАР

1. Leonard Bloomfield *(1927).* On some rules of Pāṇini. Journal of the American Oriental Society. *47. American Oriental Society. – P. 61-70.*
2. https://en.wikipedia.org/wiki/Morphology_(linguistics)
3. Schleicher, August. Zur Morphologie der Sprache. Mémoires de l'Académie Impériale des Sciences de St.-Pétersbourg. VII°. I, N.7. St. Petersburg. 1859. – P. 35.
4. Ножов И.М.. Морфологическая и синтаксическая обработка текста (модели и программы): Дис. канд. филол.наук. – Москва, 2003. – С. 54.