JOURNAL OF ADVANCED LINGUISTIC STUDIES
VOL. 11, NO. 1, JAN-JUN 2024 (ISSN 2231-4075)

Linguistic Database of Inflectional and Derivational Morpheme Models of POS in Uzbek Corpus

SAODAT BOYSARIYEVA
Termez State University Termez, Uzbekistan

ABSTRACT

The present paper investigates the major morphotactic and tagging underlying the formation of suffixed words in Uzbek, with the purpose of tackling the issue of their formalization. After having identified the models of non-finite of the verbs for Uzbek corpus. Corpus is as main resource for develop NLP morpheme analysis is considered significant in the stage of text analysis. Consequently, inflectional and derivational morphemes in linguistic database are implemented to identify stemming stage accordingly. By obtained statistical results of morphotactic combination non-finite forms of the Uzbek and general models of derivational models developed tagging system in Uzbek corpus.

Keywords: Corpus, inflectional and derivational morpheme models, FST, POS tagging

INTRODUCTION

There have been a number of investigations implemented in Uzbek Computational linguistics in the scope of computational morphology [1, 2, 3, 4, 5]. Morphological richness and complexity of the Uzbek language trigger some challenges for analysis like other rest of Turkic languages. In our article, we focus on the issues morphotactic modeling and tagging based on FST for non-finite forms, namely Verbal noun, Converb Participle, Verbal Adjective for Uzbek verbs. Modeling and tagging of language units (Part of speech) is the most significant process in Natural Language Processing (NLP). Like all Turkic languages, the Uzbek language rules follow to the law of agglutination. Conversely, the Uzbek language has a relatively simple morphotactic structure which far less harmonies than other Turkic languages. In the Uzbek language, the verb is a major type of paradigms of part of speech. Common grammatical forms of the verb are tense, person-number, mood, adverbial, adjective, relative, participle-infinitive adverbs. Manifestation of incomprehensible forms in the morphotactic state was determined on the basis of theories developed by scientists. However, the work on morphotactic methods of the Uzbek language morpho-analysis of forms that do not match the rules to a certain extent and their tagging for morpho-analysis in Uzbek corpus [13] did not come to an end. After having completed morphological modeling by using FST for Uzbek corpus, we will have to tackle some problems of exceptions and other rest of rules.

Related Works

Crucially, there has been a great deal of works on morpheme, analyzing system for agglutinative languages. For example, POSTAG for the Korean language is based a statistical/rule-based hybrid POS tagging system which comprises a combination of a morpheme pattern dictionary which encodes general lexical patterns focused to generalization unfamiliar morphemes with a posteriori syllable tri-gram estimation to predict relatively [6]. Moreover, morphological segmentation system like CHIPMUNK applied for six languages including three stages of analysis like morphological segmentation, stemming and morphological tag classification by the approach of the framework of labeled morphological segmentation (LMS) allocates a fine-grained morphotactic tag to each segment [7]. While R. Kenneth truly evaluated challenges of models of the morphemes: Morphotactic and Phonological/Orthographical, he indicated FST morphology grammatical rules expressed by regular expressions. There are operations lexc, twolc, replace are used for in different positions of morphemes modules [8]. If we look through other works agglutinative languages, we can see FST is more active implemented for inflectional models of Part of Speech. For example, Xerox Finite State Tools is used for the Uyghur Verbs of Finite state two-level morphological analyzer by using [9].

VERB MORPHOTACTIC DESCRIPTION

Parts of Speech (POS) are divided two groups: nominal (Noun, Adjective, Numeral, and Pronoun) and verbal (Verb) according to similar morphemes adjoining consequently. This approach can help to simplify the models, occurring repeating by number, hence inner paradigmatic categories are used the same. Our morphotactics description is carried out using a special language lexc via finite state transducer which is done by a special compiler lexc (Lexicon Compiler). However, we have to consider the types of non-finite forms of the verbs what general models of affixes. Using Tagset of Universal dependencies is identified each grammatical feature of the words.

Several studies suggest that affix ordering follows three types of factors to build morphemes [10]:

1. Grammatical factors: a) Syntax and semantics; b) Phonology
2. Arbitrary, stipulated via language-specific position classes
3. Extra-grammatical factors (frequency, parsability, productivity)

RESULT

During our work, the tagging system used for Turkic languages was implemented for morphotactic rules. In the process of tagging, we created tagset for both languages English and Uzbek. In the Uzbek language, there are some forms of the verb that require the use of human validation for morphotactic formation. While morphemes are divided two groups according to formal-semantic aspect: derivation morphemes and inflectional morphemes, all morphemes are categorized as linguistic database in regarding to for the process stemming of the lexemes. Hence, expert evaluation is necessary in the active and relative forms of the verb which obtained in our previous work by having done several methods of automatic generation of the morphological model of Uzbek verbs using FST technology for POS. As next furtherance of the above work, we tried to clarify the morphotactic rules of the functional forms of the verb (noun of the action, adjective, adverb), as well as the morphological features of relative adverbs. Since the adverbial form is attached to the verb and expresses the characteristics of the action expressed in it, it does not have the nominal feature, that is, it cannot take the forms of the noun after it. The noun of action and the verb in adjective form have the characteristic of accepting grammatical forms typical of nouns after it: agreement, possessive, plural. Based on the above grammatical rules, we used the following tags when we modeled verbs and adjectives morphotactically using FST technology:

Non finite form of the verb	**National Tag**	**English name**	**English tag**	**Types**	**Affixes**
Harakat nomi	HK	Verbal noun	Vnoun		-ish, -sh, -v, -uv, -moq, -mak, -maslik
Sifatdosh	SIF	Participle, verbal adjective	Part	Part+Past	-gan (kan, -qan)
				Part+Press	-ayotgan, -yotgan
				Part+Fut	-ydigan, adigan, -(a)r
Ravishdosh	RAV	Converb	Conv		-b, -ib, -y, -a, -gancha (-kuncha, -quncha), -guncha (-kuncha, -quncha), -gach (-kach, -qach), -gani (-kani, -qani), -gali

And we combined those affixes with Voice of the verb (Causative, Passive, Reflexive, Reciprocal voice)In particular, the morphotactic model of person-number forms is as follows:

```
ayt+VERB+AFF+Part+Past+PP1+PSG:aytganim
ayt+VERB+AFF+Part+Past+PP2+PSG:aytganing
ayt+VERB+AFF+Part+Past+PP3+PSG:aytgani
ayt+VERB+AFF+Part+Past+PP1+PPL:aytganimiz
ayt+VERB+AFF+Part+Past+PP2+PPL:aytganingiz
ayt+VERB+AFF+Part+Past+PP3+PPL:aytganlari
  ayt+VERB+AFF+Part+Past+PP1+PSG+Dat:aytganimga
  ayt+VERB+AFF+Part+Past+PP2+PSG+Dat:aytganingga
  ayt+VERB+AFF+Part+Past+PP3+PSG+Dat:aytganiga
  ayt+VERB+AFF+Part+Past+PP1+PPL+Dat:aytganimizga
  ayt+VERB+AFF+Part+Past+PP2+PPL+Dat:aytganingizga
  ayt+VERB+AFF+Part+Past+PP3+PPL+Dat:aytganlariga
    ayt+VERB+AFF+Vnoun+PP1+PSG:aytishim
    ayt+VERB+AFF+Vnoun+PP2+PSG:aytishing
    ayt+VERB+AFF+Vnoun+PP3+PSG:aytishi
    ayt+VERB+AFF+Vnoun+PP1+PPL:aytishimiz
    ayt+VERB+AFF+Vnoun+PP1+PPL:aytishingiz
    ayt+VERB+AFF+Vnoun+PP1+PPL:aytishlari
    ayt+VERB+AFF+Vnoun+PP1+PSG+Gen:aytishimning
    ayt+VERB+AFF+Vnoun+PP2+PSG+Gen:aytishingning
    ayt+VERB+AFF+Vnoun+PP3+PSG+Gen:aytishining
    ayt+VERB+AFF+Vnoun+PP1+PPL+Gen:aytishimizning
    ayt+VERB+AFF+Vnoun+PP1+PPL+Gen:aytishingizning
    ayt+VERB+AFF+Vnoun+PP1+PPL+Gen:aytishlarining
```

Voice is one of the active forms appearing in the verb. Although the Uzbek language has 5 types of relative forms according to the degree of participation of the subject in the process of action and state, 4 of them have special grammatical forms. We tagged in English and Uzbek as follows.

Voice_type_Uzb	Tag_uzb	Voice_type_eng	Tag_eng	Morpheme
Aniq nisbat	ANN	Active or actor-focus voice	Act	-
O‘zlik nisbat	UZN	Reflexive voice	Rfl	-n, -in, -l, -il
Majhul nisbat	MJN	Passive or patient-focus voice	Pass	-n, -in, -l, -il
Orttirma nisbat	ORN	Causative voice	Cau	-t, -dir, -tir, -gaz, -kaz, -qaz, -giz, -kiz, -qiz, -g‘iz, -ir, -ar, -iz, -sat
Birgalik nisbat	BRN	Reciprocal voice	Rcp	-sh, -ish

The expression of Voice forms is presented in our model below. There are 15402 participle models that come with verbs, verbal noun, participle and converbs. 15402 of the 16052 generated models involved Voice forms.

```
ayt+VERB+AFF+Cau+Part+Past+PP1+PSG:ayttirganim
ayt+VERB+AFF+Cau+Part+Past+PP2+PSG:ayttirganing
ayt+VERB+AFF+Cau+Part+Past+PP3+PSG:ayttirgani
ayt+VERB+AFF+Cau+Part+Past+PP1+PPL:ayttirganimiz
ayt+VERB+AFF+Cau+Part+Past+PP2+PPL:ayttirganingiz
ayt+VERB+AFF+Cau+Part+Past+PP3+PPL:ayttirganlari
 ko‘r+Verb+Cau+Vnoun+NEG+PP1+PSG+Dat:ko‘rsatmasligimga
 ko‘r+Verb+Cau+Vnoun+NEG+PP2+PSG+Dat:ko‘rsatmasligingga
 ko‘r+Verb+Cau+Vnoun+NEG+PP3+PSG+Dat:ko‘rsatmasligiga
 ko‘r+Verb+Cau+Vnoun+NEG+PP1+PPL+Dat:ko‘rsatmasligimizga
 ko‘r+Verb+Cau+Vnoun+NEG+PP2+PPL+Dat:ko‘rsatmasligingizga
 ko‘r+Verb+Cau+Vnoun+NEG+PP3+PPL+Dat:ko‘rsatmasliklariga
```

More than one voice form can be added to the same verb stem to derive different grammatical meanings.

```
Collected morphotactic results:
[VERB+Cau+Cau+Part]=>
[VERB+Cau+Pass+Part]=>
o‘qi+VERB+Cau+Cau+Part+Past+PP1+PSG:o‘qittirganim
o‘qi+VERB+Cau+Cau+Part+Past+PP2+PSG:o‘qittirganing
o‘qi+VERB+Cau+Cau+Part+Past+PP3+PSG:o‘qittirgani
o‘qi+VERB+Cau+Cau+Part+Past+PP1+PPL:o‘qittirganimiz
o‘qi+VERB+Cau+Cau+Part+Past+PP2+PPL:o‘qittirganingiz
o‘qi+VERB+Cau+Cau+Part+Past+PP3+PPL:o‘qittirganlari
o‘t+VERB+Cau+Pass+Part+Past+PP1+PSG+Abl:o‘tkazilganimdan
o‘t+VERB+Cau+Pass+Part+Past+PP2+PSG+Abl:o‘tkazilganingdan
o‘t+VERB+Cau+Pass+Part+Past+PP3+PSG+Abl:o‘tkazilganidan
o‘t+VERB+Cau+Pass+Part+Past+PP1+PPL+Abl:o‘tkazilganimizdan
o‘t+VERB+Cau+Pass+Part+Past+PP2+PPL+Abl:o‘tkazilganingizdan
o‘t+VERB+Cau+Pass+Part+Past+PP3+PPL+Abl:o‘tkazilganlaridan
```

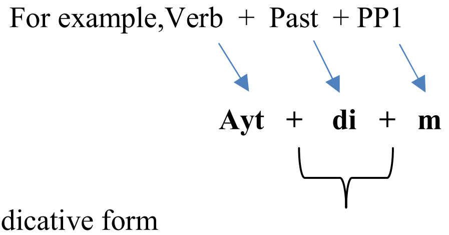
There were 16,052 models of the morphotactic state of the verbal noun, participle, converbs, subjunctive, and voice forms of the verb in the Uzbek language.

Grammatical category	Affix	Tags
Person I Possesive Singular	-m,-im	**PP1+PSG**
Person II Possesive Singular	-ng,-ing	**PP2+PSG**
Person III Possesive Singular	-i,-si	**PP3+PSG**
Person I Possesive Plural	-miz,-imiz	**PP1+PPL**
Person II Possesive Plural	-ngiz,-ingiz	**PP2+PPL**
Person III Possesive Plural	-lari	**PP3+PPL**
Genitive	-ning	**Gen**
Accusative	-ni	**Acc**
Locative	-da	**Loc**
Ablative	-dan	**Abl**
Past	-gan,-di	**Past**
Present	-moqda, -yapti, -yotir, -yap	**Pres**
Future	-a, -y,-ar, -ajak	**Fut**
Imperative mood I Singular	-ay, -ayin	**Imp+PP1**
Imperative mood II Singular	-aylik	**Imp+PP1+PL**
Imperative mood III Singular	-gin, -ing	**Imp+PP2**
Imperative mood I Plural	-inglar	**Imp+PP2+PL**
Imperative mood I Plural	-sin	**Imp+PP3**
Imperative mood I Singular	-sinlar	**Imp+PP3+PL**
Condition	-sa	**Cnd**
Purpose	-moqchi	**Prp**
Person _1	-m	**PERS1+SG**
Person_PL_1	-k	**PERS1+PL**
Person _2	-ng	**PERS2+SG**
Person_PL_2	-ngiz	**PERS2+PL**
Person_1	-man	**PERS1+SG**

Person_PL_1	-miz	**PERS1+PL**
Person_2	-san	**PERS2+SG**
Person_PL_2	-siz	**PERS2+PL**
Past Participle	-gan, -kan, -qan	**Part+Past**
Present Participle	-ayotgan,-yotgan	**Part+Pres**
Future Participle	-adigan, -ydigan, -ar	**Part+Fut**
Infinnitive (positiv)	-sh, -ish, -v, -uv, -moq, -mak	**Vnoun**
Infinnitive (negative)	-maslik	**Vnoun+NEG**
Converb	-b, -ib,-a, -y, -gancha, -kancha, -qancha, -guncha, -kuncha, -quncha, -gach, -kach, -qach, -gani, -may, -mayin	**Conv**
Converb (negative)	-masdan	**Conv+NEG**
Passive or patient-focus voice	-n,-in,-l,-il	**Pass**
Reflexive voice	-n,-in,-l,-il	**Rfl**
Causative voice	-t, -dir, -tir, -ir, -ar, -iz,-sat,--gaz, -qaz, -kaz, -giz, -kiz, -qiz, -g'iz	**Cau**
Reciprocal voice	-sh, -ish	**Rcp**
Negative	-ma	**NEG**

In the table above, all the grammatical forms that perform morphosyntactic functions in the Uzbek language are listed with tags, and the morphotactic model of almost all of these forms is used in the corpus of the Uzbek language (https://uzbekcorpus.uz/). The morphotactic model is important in determining both the morphological and syntactic position of a particular text. The morpheme dictionary and morphotactic models created for the corpus of the Uzbek language that we present can also serve in computer syntactic analysis.

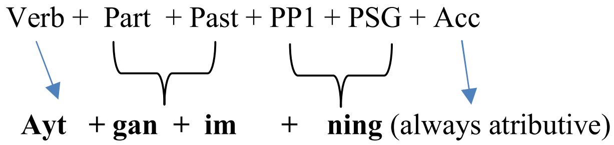
Modeling the grammatical forms of all word groups thanks to morpheme dictionary can serve as a tool for syntactic analysis. That is, it is known that the verb is formed with the additions of tense, person-number and mood, and it is participle in the syntax.



For example,Verb + Past + PP1

Ayt + **di** + **m**

Predicative form
(This form is predicate in the sentence)
or



Verb + Part + Past + PP1 + PSG + Acc

Ayt + **gan** + **im** + **ning** (always atributive)

Hence it has been formed word formation models and grammatical models of 2 different Uzbek languages.

In our previous works we compiled all derivational morpheme and it stated in following time table:

Root +	**Number of Noun Model**	**Number of Verb Model**	**Number of Adjective Model**	**Number of Adverb Model**
Verb	55	-	49	2
Adjective	8	15	6	10
Noun	33	17	47	6
Adverb	2	7	4	6
Pronoun	1	4	1	2
Numeral	1	3	-	-
Modal word	-	3	-	2
Imitative word	5	8	3	-
Exclamatory word	-	1	-	-
All	105	58	110	28

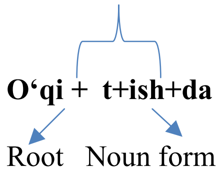
All derivational models: **306**

Here represented Noun model as example the following list:

Noun→Noun=>	-bin=>fol**bin**, -boz=>dor**boz**, -bon=>bog‘**bon**, -voy=>non**voy**, -gar=>zar**gar**, -gir=>fazo**gir** garchilik=>yog‘in**garchilik**, -goh=>qaror**goh**, -diq=>o‘rin**diq**, -don=>gul**don**, -dor=>chorva**dor**, -dosh=>sinf**dosh**, -do‘z=>gilam**do‘z**, -zor=>gul**zor**, -iston=>qabr**iston**, -kash=>arava**kash**, -kor=>paxta**kor**, -lik=>bola**lik**, -liq=>ota**liq**, -loq=>suv**loq**, -paz=>osh**paz** mand=>hunar**mand**, -soz=>soat**soz**, -xon=>kitob**xon**, -xo‘r=>meros**xo‘r**, -chi=>gul**chi**, -chilik=>qoramol**chilik**, -shunos=>o‘lka**shunos**, -furush=>chit**furush**, -iyat=>ruh**iyat**, -parast=>but**parast**, -vachcha=>amaki**vachcha**, -tarosh=>haykal**tarosh**

Compared to our previous work [11] grammatical categories divided two groups nominal (Noun, Adjective, Numeral, Pronoun) and verbal (Verb). Using FST technology there have been collected modelling nominal group 78 842 (edited 12452) and 26985 verbal models (edited 6851).

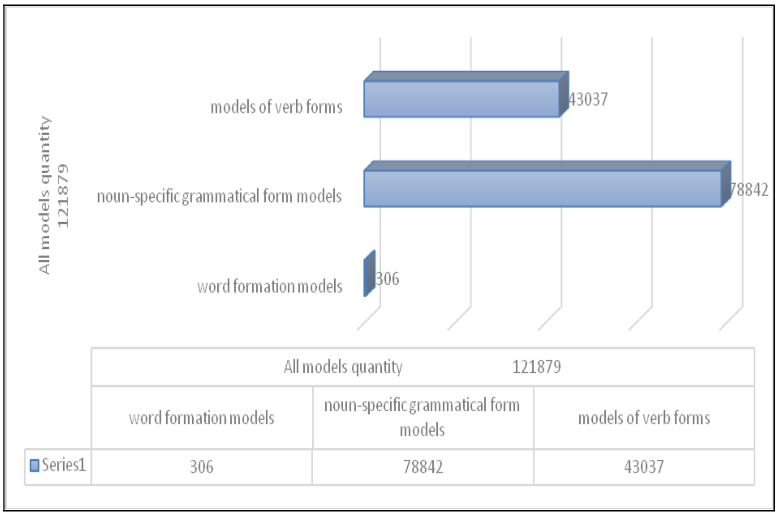
Verb form



O‘qi + t+ish+da

Root Noun form

We combined non-finite of the verb according with this technology. Using FST, we have developed models of grammatical forms of participle, infinnitive converb of the verb.



In the diagram, you can see that there are 121,879 common patterns in the Uzbek language, of which 78,842 are noun forms, 43,043 are verb forms, and 306 are word formation models.

The morpheme models we have developed serve to get an idea about Uzbek word formation and morphotactic models. The formation of a general model of grammatical formation is important in the implementation of morphosyntactic marking for the corpus. In addition, it leads to the improvement of the software of the Uzbek language corpus.

CONCLUSION

Uzbek corpus is as significant resource to develop NLP morpheme analysis is considered significant in the stage of text analysis. Consequently, inflectional and derivational morphemes in linguistic database are implemented to identify stemming stage accordingly. By obtained statistical results of morphotactic combination non-finite forms of the Uzbek and general models of derivational models are developed in tagging system in Uzbek corpus.

REFERENCES

1. Abdurakhmonova N., Modeling Analytic Forms of Verb in Uzbek as Stage of Morphological Analysis in Machine Translation, UCT Journal of Social Sciences and Humanities Research, 5(3) (2017) 89–100
2. Sulevmanov D., Gatiatullin A.,Prokopyev N. and Abdurakhmonova N. "Turkic Morpheme Web Portal as a Platform for Turkology Research," 2020 International Conference on Information Science and Communications Technologies (ICISCT), Tashkent, Uzbekistan, 2020, pp. 1-5, doi: 10.1109/ICISCT50599.2020.9351500.
3. Abdurakhmonova N.,Alisher I. and Sayfulleyeva R. "MorphUz: Morphological Analyzer for the Uzbek Language," 2022 7th International Conference on Computer Science and Engineering (UBMK), Diyarbakir, Turkey, 2022, pp. 61-66, doi: 10.1109/UBMK55850.2022.9919579.
4. Mengliev D., Barakhnin V., Abdurakhmonova N. 2021. "Development of Intellectual Web System for Morph Analyzing of Uzbek Words" Applied Sciences 11, no. 19: 9117.
5. Abdurakhmonova N., Boysarieva S. Morpheme analysis of occasionalisms in natural language processing (NLP) //Mejdunarodnyy zhurnal iskusstvo slova. - 2023. - T. 6. – no. 3; Boysariyeva S. Electronic dictionaries as a new stage of lexicography // Science and Innovation. – 2024. – T. 3. – №. 2. – C. 32-34.
6. Jeongwon Ch., Geunbae L.,Jong-Hyeok L. 1998. Generalized unknown morpheme guessing for hybrid POS tagging of Korean. In Sixth Workshop on Very Large Corpora.
7. Cotterell R., Müller T., Fraser A., and Schütze H. 2015. Labeled Morphological Segmentation with Semi-Markov Models. In Proceedings of the Nineteenth Conference on Computational Natural Language Learning, pages 164–174, Beijing, China. Association for Computational Linguistics.
8. Kenneth R. Beesley and Karttunen L. 2000. Finite-State Non-Concatenative Morphotactics. In Proceedings of the Fifth Workshop of the ACL Special Interest Group in Computational Phonology, pages 1–12, Centre Universitaire, Luxembourg. International Committee on Computational Linguistics.
9. OrhunM., TantuğA., Adali E. (2018). Rule Based Tagging of the Uyghur Verbs.

10. Rice K. 2011. Principles of affix ordering: An overview. Word Structure 4(2).169–200.; InkelasSh. 2016. Affix ordering in Optimal Construction Morphology. In Daniel S., Heidi H. (eds.), Morphological metatheory, 479–511. Amsterdam/Philadelphia: John Benjamins
11. Abdurakhmonova N. Computer models of the electronic corpus of the Uzbek language (monograph). Tashkent, 2021.
12. https://uzbekcorpus.uz/

SAODAT BOYSARIYEVA
LECTURER,
TERMEZ STATE UNIVERSITY TERMEZ,
UZBEKISTAN.
E-MAIL: <SBOYSARIYEVA@TERSU.UZ>