MORPHOLOGICAL ANALYSIS BY FINITE STATE TRANSDUCER FOR UZBEK-ENGLISH MACHINE TRANSLATION

Abdurakhmonova Nilufar,
Tashkent state university of Uzbek language and literature named after Alisher Navoi
Tuliyev Ulugbek,
National University of Uzbekistan named after Mirzo Ulugbek

Key words: *morphological rules, morphophonological rules, automatic morphological analyser, machine translation.*

I. Introduction

Machine translation is the process of interaction between human and computer. It depends on not only computational technology but also interdisciplinary of sciences which belonging to for understanding text. Therefore, if the translation is for English and Uzbek, there are different structures and peculiarities make to study morphological aspects before translation stage.

Over the last 30 years, numerous researches have been carried out to create technologies for computational morphology. Morphological analyzer for Turkic languages proceeded in the beginning of 60s-years in 20th century [1]. Morphoanalyzer is necessary for machine translation to divide components of the words and identify the grammatical paradigms of target language. Uzbek language is one of agglutinative languages and English is inflection one. Therefore, there are a lot of morphemes like these languages. A morpheme is small meaningful unit of lexeme. It has two components as stem and affix. Stem gives main sense for lexeme and affix add grammatical or semantical meaning to the word. There are many ways to combine morphemes to create words. Four of these methods are common and play important roles in speech and language processing: **inflection**, **derivation**, **compounding**, and **cliticization** [2]. In Uzbek the number of possible inflectional affixes is rather big than other non-Turkic languages. Because nearly all parts of speech could be in inflected form in context: Noun: bola+jon+lar+im+dagi+lar+niki+mas+mi+ka n+a; Simple verb: o‘qi+t + tir + ma + yot + gan + lig + I + ni, Compound verb: mashq qil+dir+ish+ayot+gan+lar, verbal compound: ber+dir+tir+ib yubor+ma+yot+gan+dan+mi+kan+a *and so on.*

I.I. Morphotactic opportunity in Uzbek language.

Here morphotactics also plays main role for morphological parsing. After morphological parsing, the components of text are analyzed semantical approach. Consequently all legal and illegal positions morphemes are considered in spotlight. In Uzbek morphotactics of words are such as order position: (1) **prefix (**2**) root +** (3) **derivative affix +** (4) **lexical affix+**(5) **grammatical affix (**(*1)ham(2)qishloq (3)lik(4)lar(5)imiz(5)dan*). In English (1) **Prefix+** (2) **root +** (3) **lexical suffix +** (4) **grammatical suffix (**(*1)**co**(2)**work**(3)**er**(4)**s)**. *However the model is like each other Uzbek grammatical affixes match preposition and adverb in English.*

The most sub-problem of morphological recognition emerged in Turkic languages for machine translation. Because a morphological dictionary is a database, in which linguistic information could be stored.

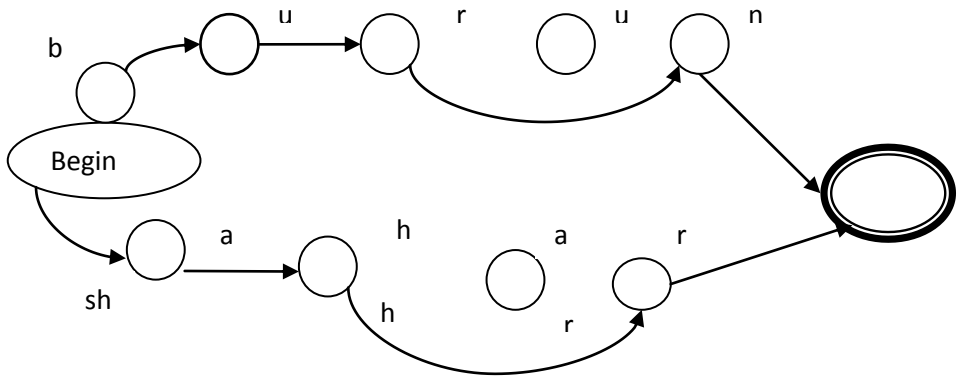
Some times to identify model of morphotactic knowledge of words is a bit problematic task if morphemes are compoundable: *yog’ingarchilik* and *zargarchilk, paxtachilik*. First word cannot be broken into parts, because there is not *yog’in+**garchilik***, but as a job there is *zar+**gar*** used separately from *+**chilik**, paxta+**chi+lik.*** As a result, it is three forms of morphemes: garchilik, gar+chilik, chi+lik.

Therefore we length of string as morpheme in Uzbek. We assume that there is nine letter of longest morpheme like g+a+r+c+h+i+l+i+k. Linguistic database of Uzbek input software in morphological parsing.

Additionally, orthographic rules has important role for all agglutinative languages for morphological analysis. Because there are, so many phonetical changes in the words make usually a large number of rules. From right to left the first vowel is removed when it analyzes for deleting some possessive cases. So we can see this situation like this chart:

bur**u**n+im=>burnim-deleting

shah**a**r+im=> shahrim-deleting



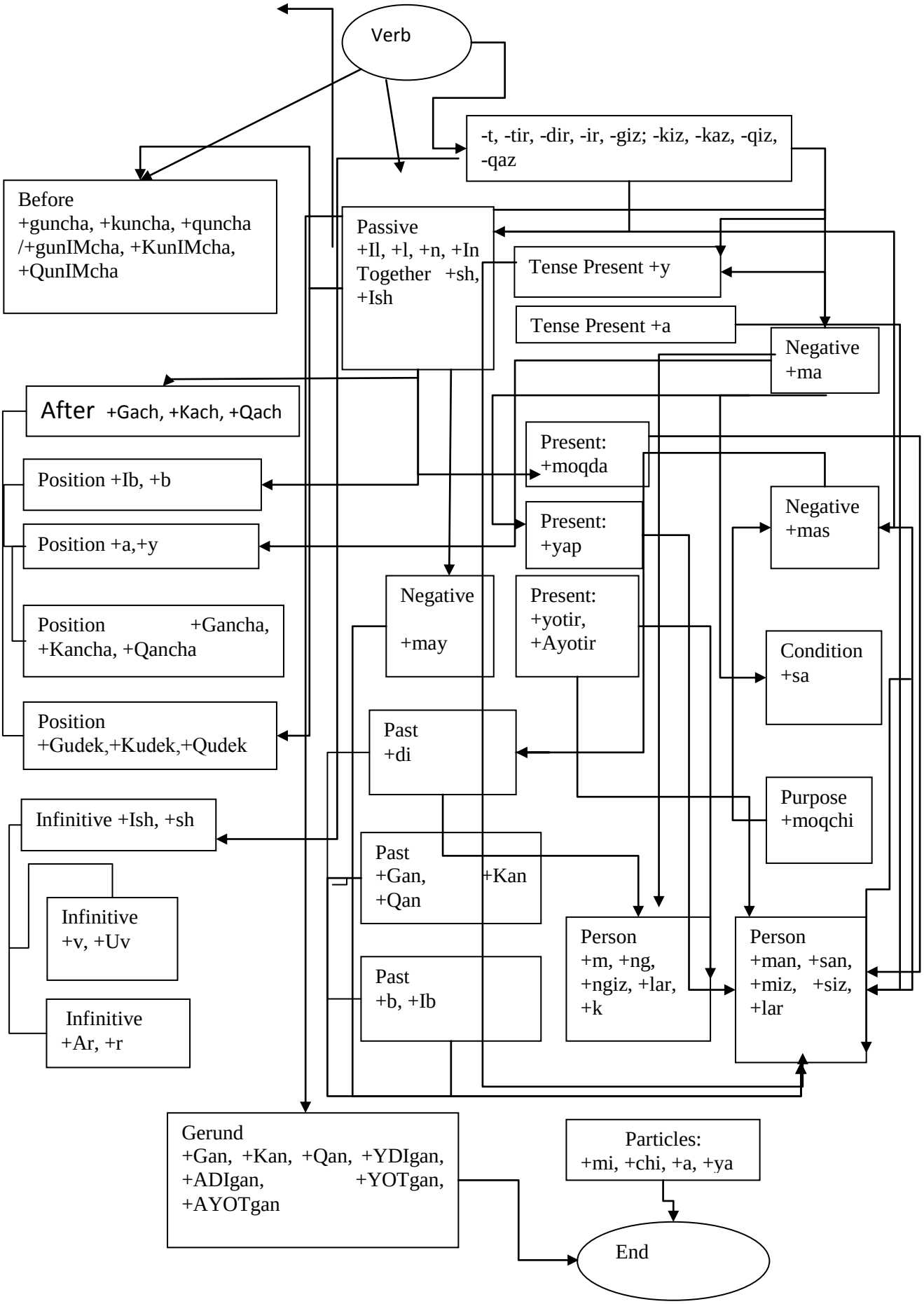
Other possibilities are epenthesis of a segment under phonological conditions. Take for example possessive case or dative case in Uzbek:

obro‘+im=>obro‘**y**im (my reputation); u+ga=> u**n**ga (he=> him)

Word error rate (WER) is the sum of insertions, deletions, and substitutions normalized by the length of the reference sentence. A slight variant (WERg) normalizes this value by the length of the Levenshtein path, i.e., the sum of insertions, deletions, substitutions, and matches: this ensures that the measure is between zero (when the produced sentence is identical to the reference) and one (when the candidate must be entirely deleted, and all words in the reference must be inserted) [3].

In a parser, morphological analysis of words is an important prerequisite for syntactic analysis. Properties of a word the parser needs to know are its part-of-speech category and the morphosyntactic information encoded in the particular word form. Another important task is lemmatization, i.e. finding the corresponding dictionary form for a given input word, because for many applications a lemma lexicon is used to provide more detailed syntactic (e.g, valency) and semantic information for deep analysis.

Alternation and adjacency of morphemes is important to analyze automatically for finite state transducers. Following scheme shows morphotactic order of the verb in Uzbek.


Verb
-t, -tir, -dir, -ir, -giz; -kiz, -kaz, -qiz, -qaz
Before +guncha, +kuncha, +quncha /+gunIMcha, +KunIMcha, +QunIMcha
Passive +Il, +l, +n, +In Together +sh, +Ish
Tense Present +y
Tense Present +a
Negative +ma
After +Gach, +Kach, +Qach
Present: +moqda
Position +Ib, +b
Position +a,+y
Present: +yap
Negative +mas
Negative +may
Present: +yotir, +Ayotir
Position +Gancha, +Kancha, +Qancha
Condition +sa
Position +Gudek,+Kudek,+Qudek
Past +di
Purpose +moqchi
Infinitive +Ish, +sh
Past +Gan, +Kan +Qan
Infinitive +v, +Uv
Person +m, +ng, +ngiz, +lar, +k
Person +man, +san, +miz, +siz, +lar
Past +b, +Ib
Infinitive +Ar, +r
Gerund +Gan, +Kan, +Qan, +YDIgan, +ADIgan, +YOTgan, +AYOTgan
Particles: +mi, +chi, +a, +ya
End

II. Derivative possibility of Uzbek

Hitherto owing to lack of resources of Uzbek language in database, we may see some problems like verbal categories in morphology. In order to analyze correctly morphemes in the context it should be construct classification and structure of verbs. Derivation is also productive in Uzbek:

Stem (Noun)	**Derivative affixes**	**Part of speech**
Gul (Flower)	-chi (florist)	Noun
	-dor	Adj.
	-li (floral)	Adj.
	-siz (without flower)	Adj.
	-chilik	Noun
	-la (blossom)	Verb
	-don (flowerpot)	Noun

There are some issues on the types of affixes in the approach of inflection and derivation. For instance in derivational diversity of we can see the models of morphotactics in the verbs:

Noun+	**-a** =>sana, **-an** =>kuchan, **-i**=>ranji, **-ik**=>ko‘zik, **-ir**=>gapir, **-y**=> kuchay, **-ka**=>iska, **-la**=>gulla, **-lan**=>faxrlan, **-lash**=>ommalash, **-lashtir** =>sahnalashtir, **-sit**=>aybsit, **-sira**=>suvsira, **-iq** => yo‘liq, **-g‘ar**=> jamg‘ar, **-qar** =>boshqar
Adjective+	**-a**=>qiyna, **-i**=>tinchi, **-ay**=>toray, **-la** =>maydala, **-lan**=>shodlan, **-lash** =>osonlash, **-lat**=> **-lashtir**=>soxtalashtir, **-r**=>qisqar, **-ar** =>oqar, **-si** =>garangsi, **-sin** =>yotsin, **-sira**=>begonasira**, -t**=>to‘lat, **-it**=>berkit, **-iq**=>namiq
Numeral+	**-ik**=>birik, **-lan**=>ikkilan, **-lash**=>birlash
Pronoun+	**-la** =>sizla, **-si =>mensi, -sira=>sensira**
Adverb+	**-ik**=>kechik**, -ir**=>ko‘pir, **-ay**=>ko‘pay, **-la**=>tezla**, -lash**=>birgalash, **-sit**=>kamsit, **-chi**=>ko‘pchi
Imitative words +	**-a**=>shildira**, -illa** =>guvilla**, -ur**=>tupur, **-ira**=>yaltira, **-la**=>gumburla, **-ra**=>ma’ra**, -shi**=>g‘ingshi, **qir=>**hayqir
Modal words+	**–la**=>yo‘qla, **-ol** =>yo‘qol**, -ot=>**yo‘qot
+modal affixes+	**-imsira=>**kulimsiramoq, **-inqira=>**oqarinqiramoq, **-kila=>**tepkilamoq, **-qila=>**chopqilamoq, **-gila=>**yugurgilamoq, **-g‘ila=>ezg‘ilamoq, -ish=>**to‘lishmoq, **-q=>**tutaqmoq, **-iq=>**toliqmoq, **-k=>**junjikmoq, **-ik=>**ko‘nikmoq, **-la=>**savalamoq, **-ala=>**quvalamoq, **-qi=>**yulqimoq, **-g‘i=>**to‘zg‘imoq, **-a=>**buramoq

Overall 56 types of lexical affixes that made by other parts of speech. In our lexicon includes 50 000 entries and their subdivision of categorical parameters.

Some multifunctional affixes of them come as homonyms. They make other parts of speech like noun, adjective, adverb and so on. In most cases, the words may be ambiguous apart from discourse. Therefore, to point out the certain places in syntactic position is also crucial for computational analysis. For example, the word ***och*** has different senses: *och rang –light colour, qorin och – be hungry.* Besides the word *“och”* comes as a component of idioms or compound verbs.

Ishtahani **och +ib** {ber, bo‘l, chiq, ket, ko‘r, qo‘y, tashla}

+a {bil, boshla, ol}

Ko‘gilni **och+ib** { ber, ko‘r, o‘tir, qo‘y, tashla, yubor}

+a {ol}

Finite state transducers read their input symbol by symbol and each time they read a symbol, they give a corresponding output and move to a new state. This improves the processing speed fundamentally. Practically, the processing speed is independent of the size of the rules [5]. A lexicon compiler is a program that reads sets of morphemes and their morphotactic combinations in order to create a finite-state transducer of a lexicon [6].

Sirni **och (divulge)**
Yo‘l **och (open the way)**
Fol **och (guess)**
Gul **och (flourish)**

III. Approaches to morphological analysis

An inflectional form is a combination of a stem with an inflectional affix. According to Cerstin Mahlow, Michael Piotrowski showed four approaches to restrict combination of affixes [7]: naive, affix, stem, indirection approaches.

Morphological analysis for machine translation includes morphonological rules as well. For instance English and Uzbek languages have own rules: big=>bigger; quloq (ear)=>qulog‘im (my ear)

In the early of 90s years there were three types of morphological analizators based on three models: generative model, paradigmatic model, the two-level morphological model for Tatar language [8].

IV. Algorithm for morphological

The earliest algorithms for automatically assigning part-of-speech were based on a two stage architecture (Harris, 1962; Klein and Simmons, 1963; Greene and Rubin, 1971). The first stage used a dictionary to assign each word a list of potential parts-of-speech. The second stage used large lists of hand-written disambiguation rules to winnow down this list to a single part-of-speech for each word.

It is known that machine translation is a huge problem for any language if there is lack of resources. But it can be considered as a very large problem for Uzbek language than others. Because as other Turkic languages Uzbek is very non structured language and applying some strike method to it is very difficult. Some of its difficulties has been mentioned above. According to these issues, it can be useful that if we will create a method or program for this language which analyze its parts. That, it should identify type and meanings of words in sentences. For this, we should analyze only words very first. It is called **morphoanalyzer**. Using this analyzer we can make a decision about words and their meanings, morphological or other changings in it as well.

So, creating this analyzer also can be divided several steps:

- Identifying a stem of lexemes;
- Identifying parts of speech type of stem;
- Parsing all affixes added to the word according to stem as token;
- Identifying types of all parsed affixes and noticing them.

These processes also does not go easily. Because there are also many problems we can face according to linguistical approach. For example, to identify a base of word we need the database of all simple words, which are not include any affixes, in Uzbek language. Then we should compare almost all words in database with the word. There are some idea to apply our work. Firstly, we take a letter from the end of word every time and compare with all words in database. So, we can get base cutting all affixes in the ending of word. For example: bolalarim (is not be found) -> bolalari (is not be found)-> bolalar (is not be found)-> bolala (is not be found)-> bolal (is not be found) -> bola (is found and finishes). Until we get “bola” six times we compare all words, which has less length than nine (because “bolalarim” has nine letters, and every step we can decrease for one the number of variants of words), in database. But, if the word has prefix, such as “serg’ayratlar”, “noodatiylik”, “beg’am-liging”, this method does not work: serg’ayrat (is not be found) -> serg’ayra (is not be found) -> serg’ayr (is not be found) -> serg’ay (is not be found) -> serg’a (is not be found) -> serg’ (is not be found) -> ser (is not be found) -> se (is not be found) -> s (is not be found

and finishes unsuccessfully). Because until the end of the word we cannot find a word in database similar the word which we cut. If we start cutting a letters from the beginning of the word, the same problem can be faced anyway.

Next, another idea is using *contains* method of the programming. To do this: we identify a length of the word; select words from the database that have less length than the words'; search all words in the component of the word; if not found then decreasing the length of selected words and repeating the process until getting to success. However, in this case we have more and more combinations.

Despite these problems above if we get a base using some methods, we can identify a type part of speech of the base. But, parsing all appendixes is also not easy. As our approach to morphological analyzing from left to right is appropriate for Uzbek language. Firstly, stem is taken according to parts of speech database, then identifying Taking example of some lexeme and wordforms we obtained like this algorithm by python.

```
k=1
for i in range(0, len(word)):
   if(otlar.__contains__(word[0: i+1])):
       k=i+1
print(word[0: k])
word=word[k:]
k=10
while(len(word)>0):

if(qoshOtYas.__contains__(word[0:k]))
:
print(word[0:k])
word=word[k:]
        if(len(word)>10):
           k=10
        else:
           k=len(word)

    elif(qoshimchalarOt.__contains
    __(word[0:k])):
        print(word[0:k])
        word = word[k:]
        if (len(word) > 10):
           k = 10
        else:
           k = len(word)
```

RESULT: ***BOLAJONLARIMGAMI***
(to my dear children?)

bola
jon
lar
im
ga
mi

Conclusion

As we showed above the model of morphotactic of Uzbek is crucial for analysis. Uzbek verbs have grammatical categories which are should be clarified stage of segmentation each of them. Segmentation of morphological parsing is multilevel process, so there are a number of notable approaches in the world. Each grammatical and orthographical rules are important for finite state transducer. The current article presents some ways to resolve morphoanalyzer issue for machine translation.

References:

1. Jurafskiy D. Speech and language processing. 2007. – P. 4.
2. Mitkov R. The Oxford handbook of computational linguistics. -P. 62.
3. Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster Learning Machine Translation Cambridge, Massachusetts, London, England, 2009. - P.6.
4. Raül Canals, Anna Esteve, Alicia Garrido et.al., interNOSTRUM: A Spanish Catalan Machine Translation System, Machine Translation Review, Issue No. 11, December (2000) –PP. 21-25.
5. Krister Lindén, Miikka Silfverberg, and Tommi Pirinen HFST Tools for Morphology – An Efficient Open-Source.

6. Package for Construction of Morphological Analyzers / – Computational Morphology in the Framework of the SLIM Theory of Language / State of the Art in Computational Morphology. – Zurich, 2009 P. 30.

7. Cerstin Mahlow, Michael Piotrowski (eds.). JSLIM – Computational Morphology in the Framework of the SLIM Theory of Language / State of the Art in Computational Morphology. – Zurich, 2009. –P. 15.

8. *D. Suleymanov, R. Gilmullin, R. Gataullin* Morphological analysis system of the Tatar language based on the two-level morphological model / Turklang 2017. Kazan, 2017. pp. 6-26.

Abdurakhmonova N., Tuliyev U. Morphological analysis by finite state transducer for uzbek-english machine translation. *This article describes brief results of the stages of an automatic morphological analyzer for Uzbek language, which used for machine translation system. The paper analyzes ordering of segment and the rules of the Uzbek wordforms generation in the frame of morphological aspect.*

Abduraxmonova N., Тулиев У. Инглизча-ўзбекча машина таржимасида морфоанализатор таҳлили. *Ushbu maqolada mashina tarjimasida foydalaniladigan avtomatik morfoanalizatorning bosqichlari amalga oshirish natijalarini qisqacha yoritib o‘tilgan. Shuningdek, o‘zbek tilidagi so‘zlarining segment birliklari tartibi va qoidasi morfologik aspektda tahlilga tortilgan.*