$$\begin{vmatrix} K_1(\mathrm{D_{g1}}) - K_2(\mathrm{D_{g1}}); K_1(\mathrm{D_{g2}}) - K_2(\mathrm{D_{g2}}) \\ K_1(\mathrm{D_{g3}}) - K_2(\mathrm{D_{g3}}); K_1(\mathrm{D_{g4}}) - K_2(\mathrm{D_{g4}}) \ldots \end{vmatrix} = \begin{vmatrix} \tau_1 + \tau_2 \\ \tau_3 + \tau_{4\ldots} \end{vmatrix} = \delta \quad (22)$$

при $\delta = 0$ – равнокомпозиционность предложений;

$\delta \neq 0$ – неравнокомпозиционность предолжений.

III. ВЫВОДЫ

Таким образом, разработан метод формального определения смысла предложения. Данный метод позволяет ставить и решать вопросы: 1) создания типологии формального смысла предложения; 2) осуществления формального анализа семантики слова, при этом является важным инструментом в следующих практических приложениях:

1. Для создания числового семантического поля языка;
2. Для создания формального «образа» значения слова и лексемы.

Список литературы

1. Апресян Ю.Д. Языковая номинация (общие вопросы). М., 1977.
2. Моррис Ч. Значение и означивание//Семиотика М., 1983.
3. Рубинштейн С.Л. Принципы и пути развития психологии. М., 1959.
4. Сыдыков Т., Токтоналиев К. Азыркы кыргыз тили: фонетика жана фонология. Б., 2015.
5. Ф. де Соссюр. Курс общей лингвистики. М., 2006.
6. Luria A.R. The Making of Mind: A. Personal Account of Soviet Psychology M. Cole & S. Cole, eds. Cambridge, 1979.

УДК 811.11+811.512.133:81`322.4

THE BASES OF AUTOMATIC MORPHOLOGICAL ANALYSIS FOR MACHINE TRANSLATION

***Nilufar Abdurakhmonova**, doctoral student, Tashkent State university of Uzbek language and literature named after Alisher Navoi, Tashkent city, abdurahmonova.1987@mail.ru*

The aim of this article is to show how automatic morphological analyzer identifies clarification of the verbs in English and Uzbek languages. Verbs are very complex natured category in both of languages. The linguistic database of given program should not only include pure grammar, but also some morphological algorithms of different languages. English morphology depends on syntactic analyzing in machine translation. That is way the problems of machine translation in inflected and agglutinative languages is often required to be solved in morphological analyze.

Keywords: Uzbek language, automatic morphological analyze, natural language processing, lexicon

ОСНОВЫ АВТОМАТИЧЕСКОГО МОРФОЛОГИЧЕСКОГО АНАЛИЗА ДЛЯ МАШИННОГО ПЕРЕВОДА

***Нилюфар Абдурахманова** докторант, Ташкентский Государственный Университет Узбекского языка и литературы им. Алишера Навой abdurahmonova.1987@mail.ru*

Данная статья рассматривает классифицирование глаголов в английском и узбекском языках с помощью автоматического морфологического анализатора. Глагол в обеих языках является сложной характерной категорией. Кроме того, как утверждает автор, лингвистическая база любой переводческой программы должна включать не только чистой

грамматики, но и морфологические алгоритмы разных языков. В машинном переводе морфология английского языка зависит от синтактического анализа. Исходя из данной проблемы, в статье сделана попытка найти решение морфологического анализа машинного перевода флективных и агглютинативных языков.

Ключевые слова: узбекский язык, автоматический морфологический анализ, обработка естественного языка, лексикон, субкатегорическая парадигма

I. Introduction

In Uzbekistan Computational linguistics appeared as the subject at the beginning 2000s, it began to investigate new researches, scientific works. Now machine translation has become very important issue as one of the directions of computational linguistics. It has some problems depending on analysis. First of all, it needs morphological analyzer in translation of English texts into Uzbek. Then it should be subcategorized paradigms parts of speech. In spite of lack of resource formal language of Uzbek, it has rich linguistic database description of so many literatures for that.

Within the process of automatic analyzing of the each word is to be considered morphological surface form of the word as well. We are of the opinion that, the active and passives morphological lexicon is used in translation system. Active morphological form contains the list of word stem and suffixes combinations. Grammatical rules and analysis are input to passive morphological analysis.

This paper presents the first step to lay the foundation for automatic morphological analyzer. Naturally, to input all forms of words is impossible, because there are so many combinations of word forms. That's way that is necessity to study combination word and suffixes of agglutinative language. Paradigms of part of speech are handful to adapt languages. It should be taken the specificity and general rules of those languages and it needs to be given their formal definition, especially in machine translation for not related languages. That's the reason, some problematic situations between Eastern (agglutinative) languages and European (inflected) languages have created the new and big critical approach in linguistics. Mainly these types of problems are observed in Krivonosov's works [1] concerning syntaxes and translating eastern languages into European languages and vice versa in Marchuk's works [2].

Uzbek language is morphologically rather complicated and rich in inflectional form (form with endings). For example, Noun: bola+jon(1)+lar(2)+im(3)+dagi(4)+lar(5)+ niki(6)+mas(7)+mi(8)+kan(9) (shortened form of ekan) +a(10); **Verb**: o'qi+t(1)+tir(2)+ ma(3)+ gan(4)+lig(5) (changed form of -lik)+im(6)+dan(7)+mas(8)+mi(9)+kan(10)+a(11). As we see, there is long enough chain of suffixes of inflection. Particularly when the text is translated from Uzbek into English, it will be difficult to give all the meaning of the sentence because of diverse structure. The English words are rather transparent as the morphemes are easily segmented and associated with appropriate grammatical meanings. The Uzbek word forms are considerably less transparent. Very often it is impossible to decide unambiguously whether we have morphological formative and stem element: burnim=>burun+im. Morphological analysis lies at the found of all programs of automatic text processing of the Uzbek language. Next we would like to point out a few spheres in which a morphological analyzer is indispensible.

The morphological analyzer is used in various systems for special functions: text editors (spell checker), information retrieval, automatic annotation, speech recognition and machine translation. On the one hand, there are certain linguistic problems. A text body would offer a better opportunity for studying actual language usage, a field still rather poorly cultivated in Uzbek. New prospects are opened up in the studying of a) grammar: the usage of word forms, phrases and collocations; b) lexicography: frequency dictionaries, dictionary of individual styles, authors, dialect, concordance instead of card files, as well as; c) textology and stylistics: grammatical and

lexical peculiarities of different text types[1]. The morphology part of every Uzbek grammar are, without exception, synthesis-oriented. They provide rules for the formation of the inflectional forms, but say nearly nothing about the usage those forms. In the actual usage, one word form usually dominates over its parallel forms. Of some words only the singular, or plural, or just a couple of concrete forms are used. Some forms occur only in certain fixed word of large text corpora.

The different strategies underlying morphological analyses are based on the following properties of morphological units and their relationships [3]:

-integrity of word forms

-segmental structure of word forms

-variability of units

Regularity and irregularity of relation

In order to analyze a word form it is first necessary to segment it into units which will then have to be **transformed** into the shape of their initial forms to be, in turn, searched for in dictionaries. The result of the analysis is generated from information attached to the initial forms. In morphological analysis it is not possible to consider unit variation on the level of an individual unit, instead, the word form must be treated as a member of a paradigm. The paradigmatic approach serves as basis for the model of classificatory morphology. The variability of the stem appears within the paradigm, i.e. it is revealed if we compare different inflectional forms of the same word. The variability of the formatives, in the contrary, is revealed inter paradigmatically, i.e. if we compare one and the same inflectional form across different words.

There are about 207 types suffixes (including variation) of parts of speech in Uzbek language and 130 of them are defined as verbs. In order to add endings to the bases of each words it needs to separate one or another part of speech into paradigms. We separated the verb into following paradigms:

1. According to the features of adding voice endings:

1.1. Causative voice of verbs:

V_1:-ar is added only two verbs=>V: chiq+ar, qayt+ar

V_2:-giz{-g‘iz}is added verbs that is ended voiced consonant =>yur+giz,tur+g‘iz

V_3:-dir {-tir} is added to verbs are ending with vowel and voiced consonant=>ye+dir, yoy+dir

V_4:-ir is added to verbs are ending with **t, ch, sh** consonants=>ich+ir, shosh+ir, tush+ir

V_5:-iz is added to verbs are ending **q, m** consonants=>oq+iz, tom+iz, em+iz

V_6:**-t/it** –is added ending vowels of two or many-syllabled words: ishla+t, tuga+t, boshla+t, o‘qi+t

The causative voice ending in Uzbek language looks like into the following grammatical form in English language. **Have / Get** something V_{III}=>Uzbek verbs vocabulary similar to the above mentioned groups V1, V2, V3 are entered into the linguistic database. For example, I have my lesson done –Men darsimni qildirdim. In translation process for Uzbek language we use left-to right structure. The verb is translated. According which group does it belong to: one or multy syllabled, ended with voiced consonant, ended with vowel we can put correct endings.

1.2. The endings of passive and reflexive voices.

The passive voice in English language looks like to a “category” in Uzbek language. That’s way their formula is entered into the database.

S+am/is/are+V_{III}=>S+V+PV+TS+PS[2]: The book is written- Kitob o‘qiladi.

During translation into English it should be input to lexicon due to homonym suffixes passive and reflexive voices in Uzbek. For instance, I wash-Men yuvinaman=>yuv+in+a+man. In

[1]ÜlleViks. A morphological analyzer for the estonian language: the possibilities and impossibilities of automatic analysis http://www.eki.ee/teemad/morfoloogia/viks1.html

[2]1. PV(passive voice), 2. TS(tense suffix), 3. PS(personal suffixes), 4) $V_{v\text{-verb voice}}$, 5) NP_1-noun plural, 6)NA- animate object (boy-boys), NIP-noun irregular plural (child-children)

English the meaning must be like wash-1) yuvmoq, 2) yuvinmoq; close-ochmoq, ochilmoq, begin-boshlamoq, boshlanmoq and others. And they are put in the discrete paradigms (V_v) such kind of verbs. But it should be given some grammars for these verbs: if $S+V_v+Noun$=>active, if $S+V_v+nonNoun$=>reflexive voice

1.3. Cooperative voice (Birgalik nisbat)

–sh (-ish) suffixes are belongs to the voice. What kind of English verb forms suit to the voice. It can be synonym to the plural form in Uzbek as well. For example, Bolalar kelishdi (keldilar)-The children came. So we use plural form as it comes like: NP_1+V => they/ NA+s {-es}/ NIP

2. Functional forms of Uzbek verbs (non-finite forms of the verb)

There are three types functional forms of Uzbek: participle (sifatdosh), harakat nomi, adverbial participle (ravishdosh). English has three types: gerund, infinitive and participle. The characteristics and capacity both of the languages are dissimilar. And they are not suit for each others. In the chart pointed out versions different functions of languages.

	Suffixes	Infinitive (to)	Gerund (v+ing)	Participle (V_{III})
Harakat nomi	-sh,-ish,-v,-uv, -moq, -maslik	+	+	-
Sifatdosh	-gan,-kan,-qan, -yotgan,-ayotgan,-ydigan, -adigan, -mas	-	+	-
Ravishdosh	-guncha, -kuncha,-quncha, -gach, -kach, -qach, -b, -ib, -a, -y, -may, -mayin, -ma	-	Till (until), by, after	+
Harakat nomi	O‘qish foydali	To read is useful.	Reading is useful.	-
Sifatdosh	Yonayotgan olov ajoyib.	-	Burning fire is wonderful.	
Ravishdosh	Ish qilinguncha vaqt tugaydi. Ish tugatilgach uyga ketdik.	-	Till doing work, time will be over.	After having finished work, we went home.

The morphologic analysis of English is identified coming words in order. It should be responded so that to solve some matters.

1. Verb comes after subject, and then it is considered as a predicate. Then checked simple and complexity of verbs (gerund, infinitive, modal, phrasal verb, have+noun, make+noun, do+noun, take+noun, have+noun+verb)

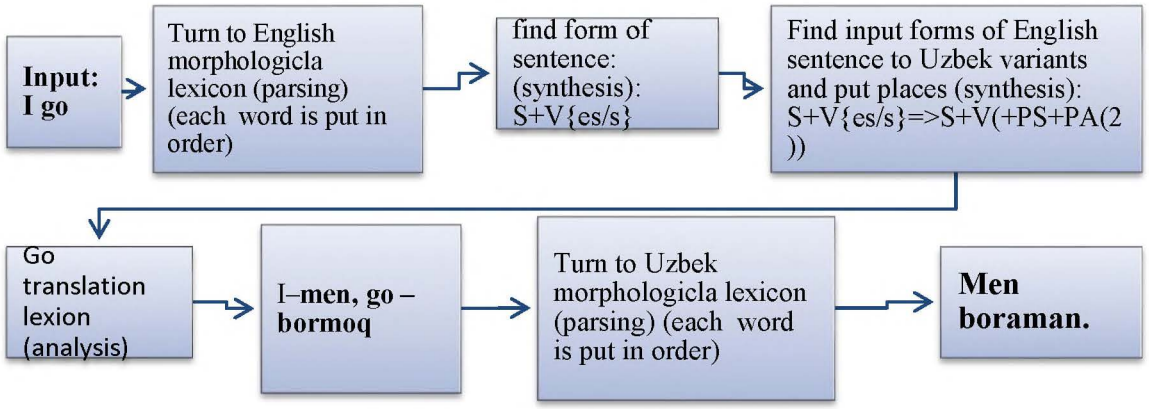
2. To identify tenses (present, past, future, future in the past)

3. Types of sentences (darak-declerative (Dec.), inkor-negative (Neg), so‘roq-interrogative (Int.), buyruq-imperative (Imp), so‘roq-inkor (IN), undov-exclamatory (EX).

4. To identify transitive and intransitive verbs. It helps to clarify accusative (case) in English. For example, to read Øthe book–kitob**ni** o‘qimoq, go Øhome-uy**ga** bormoq. As we see there isn’t any preposition in English but in Uzbek two different cases.

5. Next step to formulate sentence in two languages (Firstly we take simple sentences). Some of them are below:

	PS PS	**PC**	**PP**	**PPC**
Dec.	S+V{es/s}=> S+V(+PS+ $PA_{(2)}$)	S+am{'m}/is{s}/are{'re}+V(-ing) => S+V+(PC+ $PA_{(2)}$)	S +have{'ve}/has {'s}+ $V_{(III,-ed)}$=> S+V+(PP+ $PA_{(1)}$)	S+ have{'ve}/has {'s}+been+V(ing)
Neg.	S+do not {don't}/does not {doesn't}+V=>+V(+ NA+PS +$PA_{(2)}$	S+am/is/are+not / {isn't/aren't}+V(-ing) => S+V+(NA+PC+ $PA_{(2)}$+	S+have+not/{haven't}/has +not/{hasn't}+ $V_{(III,-ed)}$=> S+V+(PP+NA+PastA+ $PA_{(1)}$)	S+have+not+/{haven't}/has +not/{hasn't}+ been+$V_{(ing)}$=> S+V+(PP+NA+PastA+ $PA_{(1)}$)
Int.	Do/Does+S+V=?=> S+V(+PS+ $PA_{(2)}$+QA=?)	Am/is/are +S+ V (-ing)=? => S+V+ (PC+$PA_{(2)}$+QA=?)	Have/has +S+ $V_{(III,-ed)}$=> S+V+(PP +PastA+ $PA_{(1)}$+QA=?)	Have/has +S+ been+$V_{(-ing)}$=> S+V+(PP +PastA+ $PA_{(1)}$+QA=?)
Int. N	Don't/Doesn't+S+V= ?=> S+V(+NA+PS+ $PA_{(2)}$+QA=?	Am/is/are+S+not+ V(-ing)=? => S+V+(NA+PC+$PA_{(2)}$+QA=?)	Have/has +S+ not+$V_{(III,-ed)}$=> S+V+(NA+PP PastA+$PA_{(1)}$+QA =?)	Have/has +S+been+$V_{(-ing)}$=> S+V+(NA+PP +PastA+ $A_{(1)}$+QA=?)



1. I go=>:S+V{es/s}=> S+V{es/s}=>S+V(+PS+PA(2))=>Men boraman.

1.1. I don't go=>S+do not {don't}/does not {doesn't}+V=> S+V(+NA+PS+PA(2))=> Men bormayman.

1.2. Do I go?=> Do/Does+S+V=?=> S+=> S+V(+PS+PA(2)+QA=?)=> Men boramanmi?

1.3. Don't I go? => Don't/Doesn't+S+V=? =>S+V(+NA+PS+PA(2)+QA=?) => Men bormaymanmi?

2. I am going=>S+am{'m}/is{s}/are{'re}+V(-ing)=> S+V+(PC+PA(2))=> Men boryapman.

2.1. I am not going=>S+am/is/are+not / {isn't/aren't}+V(-ing) => S+V+(NA+PC+PA(2))=> Men bormayapman.

2.2. Am I going ?=> Am/is/are +S+ V(-ing)=? => S+V+(PC+PA(2)+QA=?)=>Men boryapmanmi?

2.3. Am I not going ?=>Am/is/are +S+not+ V(-ing)=? => S+V+(NA+PC+PA(2)+QA=?)=>Men bormayapmanmi?

3. I have gone=>S+have{‘ve}/has {‘s}+$V_{(III,-ed)}$=> S+V+(PP+PA(1))=> Men borib bo‘ldim.

3.1. I have not gone=>S+have+not/{haven’t}/has +not/{hasn’t}+V(3,-ed)=> S+V+(PP+NA+PastA+PA(1))=> Men borib bo‘lmadim.

3.2. Have I gone?=> Have/has +S+ $V_{(III,-ed)}$=> S+V+(PP+NA+PastA+PA(1))=> Men borib bo‘lmadim.

3.3. Have I not gone?=> Have/has +S+ not+$V_{(III,-ed)}$=> S+V+(NA+PP PastA+$PA_{(1)}$+QA=?)=> Men bormaganmidim?

We have presented a rule-based morphological analysis system for English-Uzbek translation system. As we admitted that it is initial (opening) stage of translation system. Using theories of typological grammar we create deep principles of morphological analysis. Rich lexicon, full based grammar rules, the base of terms are all of them help to improve analyzing text translation process. And we hope the next researches on linguistic database of translation program will be advanced within the next few years.

References

1. Кривоносов А.Т. (2001) Система классов слов как отражение структуры языкового создания. Москва –Нью –Йорк: Че-ро, 846 с.
2. Marchuk Y.N. (2003) The Burdens and Blessings of Blazing the Trail. In Journal of quantitative linguistics. Trier, Swets, Zeitlinger, Vol. 10, No. 2 Aug. p 81-87
3. ÜlleViks. A morphological analyzer for the estonian language: the possibilities and impossibilities of automatic analysis http://www.eki.ee/teemad/morfoloogia/viks1.html
4. Idem
5. Miriam Butt, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The parallel grammar project. COLING’02, Workshop on Grammar Engineering and Evaluation.
6. Eugene Charniak, Kevin Knight, and Kenji Yamada. 2003. Syntax-based language models for statistical machine translation. MT Summit IX.

УДК 81.32

EXPERIMENTS WITH RUSSIAN TO KAZAKH SENTENCE ALIGNMENT

Zhenisbek Assylbekov, *Nazarbayev University, Astana, Kazakhstan zhassylbekov@nu.edu.kz*
Bagdat Myrzakhmetov, Aibek Makazhanov, *National Laboratory Astana, Astana, Kazakhstan bagdat.myrzakhmetov@nu.edu.kz, aibek.makazhanov@nu.edu.kz*

Sentence alignment is the final step in building parallel corpora, which arguably has the greatest impact on the quality of a resulting corpus and the accuracy of machine translation systems that use it for training. However, the quality of sentence alignment itself depends on a number of factors. In this paper we investigate the impact of several data processing techniques on the quality of sentence alignment. We develop and use a number of automatic evaluation metrics, and provide empirical evidence that application of all of the considered data processing techniques yields bitexts with the lowest ratio of noise and the highest ratio of parallel sentences.

Keywords: sentence alignment, sentence splitting, lemmatization, parallel corpus, Kazakh language