UBMK'24

Bildiriler Kitabı
Proceedings

Editor Eşref ADALI

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı

9th International Conference on Computer Science and Engineering

26-27-28 Ekim (October) 2024 Antalya - Türkiye

9. Uluslararası Bilgisayar Bilimleri ve Mühendisliği Konferansı (UBMK'2024)

9th International Conference on Computer Science and Engineering

26-28 Ekim 2024 Akdeniz Üniversitesi Antalya Türkiye
26-28 October 2024 Akdeniz University Antalya Türkiye

IEEE Catalog Number : CFP24L97-CDR

ISBN : 979-8-3503-6587-0

Additional copies may be ordered from:
Curran Associates, Inc
57 Morehouse Lane Red Hook, NY 12571 USA
Phone: (845) 758 0400
Fax: (845) 758 2633
E-mail: curran@proceeding.com

Database of Grammatical Form and Meaning Correspondence for Uzbek-English Translation Software

Jamoliddinova Odina Rasulovna
Tashkent State University of Uzbek Language and Literature
Tashkent, Uzbekistan
jamoliddinovaodina@navoiy-uni.uz

Muhammedova Saodat Xudoyberdiyevna
Tashkent State University of Uzbek Language and Literature named after Alisher Navoi
Tashkent, Uzbekistan

Khamroeva Shahlo Mirdjanovna
Tashkent State University of Uzbek Language and literature
Tashkent, Uzbekistan
hamroyeva81@mail.ru

Ergasheva Guli Ismoil Qizi
Uzbekistan State University of World Languages
Tashkent, Uzbekistan
ergashevaguli327@gmail.com

Qodirova Madinabonu Murodjon qizi
Tashkent State University of Uzbek language and literature
Namangan, Uzbekistan
begonammmm@gmail.com

Kholova Muyassar
Department of Uzbek Linguistics, Termez State University
Termez, Uzbekistan
xolovam@tersu.uz

Abstract. **Creating a perfect Uzbek-English machine translation requires creating a perfect translation of grammatical forms in these languages. In this article, we analyze the units that we consider auxiliary words in the Uzbek language from a grammatical point of view, and several methods of their use in translation programs are proposed. The main focus of the research work is on the modeling of language units. The works of the world and Uzbek linguists dedicated to the modeling of language units were studied, and based on them, each grammatical meaning expressed by auxiliary words was approached individually.**

Keywords: machine translation, grammatical form, grammatical meaning, auxiliary language units, modeling, encoding-decoding, equivalence.

I. INTRODUCTION

The issue of classification of words in languages is of special importance. Because in this way concrete grammatical meaning, function and categorical features of the word are revealed. For this reason, from the earliest periods of linguistics, scientists have been paying attention to the issue of categorization of language units. However, it should also be noted that the perfect classification of units in many languages remains one of the main problems of linguistics. In particular, language units in the Uzbek language are classified according to three principles: lexical-semantic, syntactic and morphologic [1]. We based our research on the lexical-semantic classification of words. According to it, language units are divided into 3 groups:

1. Units with lexical meaning.
2. Units that refer to a concept.
3. Units that do not have a lexical meaning.

Grammatical form refers to words that do not have a lexical meaning, but express only a grammatical meaning. It follows that the object of our research is language units that refer to a concept and do not have a lexical meaning. It is possible to determine the amount of such units in the Uzbek language. The problem is translating them into English. Because one form can express several grammatical meanings. The main task of our research is to create a perfect set of grammatical forms in these two languages in English-Uzbek translator programs. In this article, problems of machine translation of Uzbek grammatical forms called connecting, auxiliary, predicate into English, cases of compatibility and incompatibility of grammatical form and grammatical meaning and ways to overcome these problems are shown.

II. LITERATURE REVIEW

The term morphological analysis used in traditional linguistics and the concepts of morphological analysis in the field of computational linguistics are not the same. In theoretical linguistics, morphological analysis means only the study of word form, while in applied linguistics, morphological analysis is a procedure that can be obtained from the form or appearance of information about different levels of the linguistic structure to convert into a structure, and the unit of morphological analysis is the word form into which the text is divided [2].

Morphological structure is a sequence of morphological structures of word forms included in the text, and these elements are arranged in the same order as the corresponding word forms in the text.

Comparing units that do not have a common lexical meaning in Uzbek and English, we refer to the grammar of these two languages. When analyzing word groups in machine translation, the following must be taken into account:

1. to identify the similarities of grammatical categories existing in both languages;
2. coordination of mutually different categories in both languages;
3. similar modeling of both groups above [3].

Studies on linguistic typology and comparative analysis of languages are considered in the aspect of generative linguistics. Language modeling for a computer base is the most convenient learning method. Computer morphology is necessary for high-level applications, so modeling within morphology is used in many studies [4]. In particular, at Carnegie Mellon University, M. Raab defended his thesis entitled "Language modeling techniques for machine

translation". Various methods of language modeling, comments on created language models, practical research of the researcher, results and conclusions are included in the research work [5]. University of Haifa scientists Gennady Lembersky, Noam Ordan, Shuly Wintner study the difference between language models built from original texts in the target language and language models built from texts manually translated into the target language. In this way, they propose to create language models for machine translation [6]. Research works are being carried out within the framework of machine translation in Turkish languages as well. Turkish scientists are conducting research on Turkish interlanguage and Turkish-English machine translation. For example, many scientists such as A.C. Tantuğ [7], Z. Sagay [8], I. Hamzao'glu [9], K. Altintash [10], M. Orhun can be listed. For example, in M. Orhun's research work entitled "Machine translation between uygur language and turkish", models of Turkish and Uyghur language units are created and compared [11].

In Uzbek computer linguistics, much attention is paid to modeling today. Because it is confirmed that it solves many problems of today's NLP. For this reason, scientific research work on modeling different levels of the language is being carried out. In particular, in this regard, B. Elov [12], N. Abdurahmonova, Sh. Hamroyeva [13], S. Bozorova [14], S. Muhammedova [15], S. Khudayarova [16], M. Suyunova [17] can be listed. They offer different ways of modeling language units in their work. In particular, models of simple Uzbek-English sentences for machine translation were developed and compared by N. Abdurahmonova. Sh. Hamroyeva's doctoral dissertation is devoted to the modeling of grammatical forms for the morphoanalyzer of the Uzbek language. And B. Elov offers a hidden dirixle method of modeling language corpus texts. S. Bozorova offers methods of syntactic analysis based on free context grammar and models Uzbek word combinations. M. Suyunova is engaged in modeling grammatical homonyms in the Uzbek language, and S. Khudayarova is engaged in modeling word combinations.

In morphological modeling, it is important to give a correct and precise linguistic description. The correctness of the morphological criteria is checked in the following stages:

1. analysis of only correctly formed grammatical word forms;
2. how to determine grammatically incorrect word forms [11].

The correctness of word forms is determined only from the models that are found to be correct, and the remaining cases confirm its incorrectness.

In Uzbek, conjunctions, auxiliaries and prepositions, which are generalized as auxiliary language units, have a specific form in Uzbek and appear in English through several different means. In this case, cases of mismatch between grammatical form and grammatical meaning are observed, namely, situations such as grammatical polysemy, homonymy, and synonymy appear. Such language phenomena affect the quality of translation in machine translation.

In this direction, the importance of rule-based, namely, grammatical analysis is that the analysis is consistent with linguistic phenomena of polysemy and homonymy, and another aspect is the separate study of different categories in languages belonging to different types [12].

Two types of algorithms are used to improve the quality of morphological analysis in the process of identifying a set of ambiguous words: statistical-probabilistic and rules governing words/codes.

In order to implement a perfect machine translation of Uzbek-English grammatical forms, it is necessary to create a linguistic database that takes into account linguistic phenomena. Below we will consider this model of linguistic support.

Solving the problems of synonymy, homonymy, polysemy encountered in grammatical forms requires the creation of a separate linguistic database for each case.

III. POLYSEMY IN GRAMMATICAL FORMS

Multiple meanings of a word is an independent system in language, and this phenomenon is called polysemy in general linguistics. The word was originally used for one meaning, but through the passage of time and various linguistic influences, it has several meanings. Polysemy is distinguished from other phenomena by the fact that it consists of lexical meanings within a category, their interdependence, and the fact that the meanings belong to the same word itself [13].For example,

Uxlasin deb qo'shiq aytib berdim(1).

Tanimaydi deb salom bermadim (2).

In both languages, the word deb has a grammatical meaning, but there is a difference in meaning. In the first example, the word deb expresses the meaning of the goal, and in the second, it expresses the reason. In English, they are equivalent to two words.

I sang a song for him *to* sleep(1).

I didn't say hello *because* I didn't know him(2).

In the first sentence, the word "deb" is translated as "to" and "because" in the second sentence.

It is important to fully reflect their grammatical meaning when creating the linguistic support of auxiliaries belonging to a group.

Although grammatical forms do not have a lexical meaning, their grammatical meaning and function is realized in the speech process. For this reason, it is not correct to consider them as a morphological form separated from speech.

In world and Uzbek linguistics, several theories have been developed to solve the problems of lexical polysemy, homonymy and synonymy. In a sense, they partially solved the problem. However, this is still not enough. Perfect machine translation mechanisms have not yet been developed. Studying the opinions and theories of world scientists and drawing conclusions from them, we would like to propose several ways to solve the problem of inconsistency of grammatical form and grammatical meaning.

A. Method 1 – the encoding method.

The linguistic support base is aimed at explaining such words and clarifying their meaning. It contains all the grammatical meanings of a word. The meanings represented by conjunctions, auxiliaries, and prepositions are coded to translate these grammatical forms into machine language and determine the English equivalent. A separate code is given for each grammatical meaning. The general ID number of a

morphological tool is equal to the codes represented by the grammatical meanings in it.

The first character in the code of the morphological form is its grammatical category ((bog‘lovchi – B, ko‘makchi – K, yuklama – Y), if the language unit we are studying does not have polysemous features, the next three characters are the code of its grammatical meanings, the next two characters represent the special code of the word. If the auxiliary language unit has several grammatical schemes, a two-character code is given for each scheme. For example, the atab assistant will have the ID number K003A6.

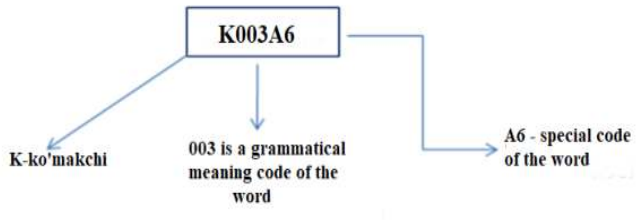

Fig.1. Encoding a word with one grammatical meaning

Some units work together in two different categories. A conditional sign of both categories is included in the coding. For example, the code of the word “bo‘yicha” is K4567F9. See figure 2.

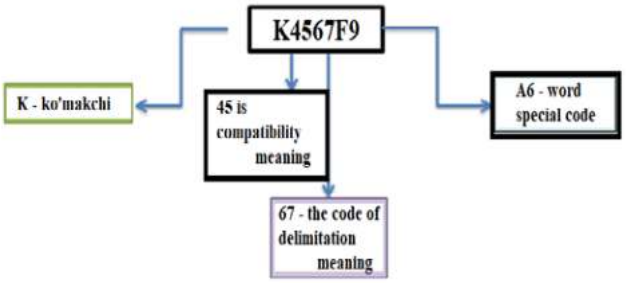

Fig.2. Encoding a word with multiple grammatical meanings

Some units work together in two different categories. A conditional sign of both categories is included in the coding. For example, the code of the word “ham” is YB-056308A4. This is given in figure 3.

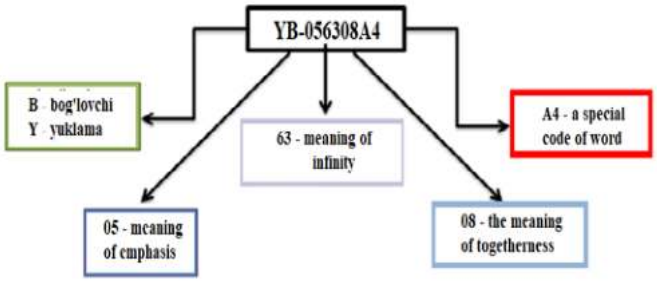

Fig. 3. Coding the word within two different categories

The grammatical meanings of auxiliary language units in the Uzbek language have the following conditional sign. See Table I.

TABLE I. SPECIFIC GRAMMATICAL MEANINGS OF AUXILIARY WORDS

definition	001
amount, measure	002
designation	003
subtraction	004
highlight	005
gallic	006
sequence	007
togetherness	008
cooperation	009
expressing togetherness	010
connection	011
emotion	012
surprise	013
scaring	014
topic of thought	015
suspicion, guess	016
the object that is the basis for the action	017
direction	018
manner of performance of action	019
the occurrence of movement along a certain place	020
occurrence of the action for a certain period of time	021
action start time	022
the initial place in the occurrence of the action	023
the initial source of action	024
duration of action	025
that only one of the two actions takes place	026
that neither of the two actions takes place	027
completion of action	028
lack of activity	029
time before action	030
level of activity	031
past tense	032
present tense	033
future tense	034
question	035
command	036
desire	037
strengthening	038
exaggeration	039
reason	040
result	041
target	042
contradiction	043
selection	044
compatibility, proportion	045
compatibility	046
exception	047
uncertainty	048
place	049
demarcation relative to place	050
time	051
analogy	052
comparison	053
clarifying on the basis of comparison	054
mixing	055
compare	056
weapon, tool	057
something that is a tool for action	058
a place that is instrumental in the occurrence of action	059
equality	060
speed	061
used in relation to a sudden action	062
barrier	063
surprise, amazement	064
condition	065
generalization	066
limitation	067
limitation in relation to time	068
demarcation relative to place	069

On the same basis, special codes of grammatical forms in English are created. They are compared and the

proportionality of grammatical forms is determined in percentages.

The difference of this method from other encoding-decoding, comparison-contrast methods is that it emphasizes the dominance of one language. That is, the language that is relatively more complex and has more exceptions is taken as dominant. Units of a second language are considered separate units, detached from the grammatical rules of that language. Separate units of the second language are analyzed based on the rules of the dominant language. In our research, the Uzbek language was taken as the main criterion. English words are analyzed based on Uzbek grammar. This method gives us the following advantages:

1. Eliminates inter-category differences. Since we are dealing with the grammatical meanings of words, we cannot avoid the issue of their categorization. The border of independent and non-independent units in English differs from that in Uzbek. Therefore, it is necessary to bring them together.
2. Facilitates unit modeling. Allows both languages to use the same encoding method.
3. The modeling of words in English is the same as in Uzbek. See figure 4.

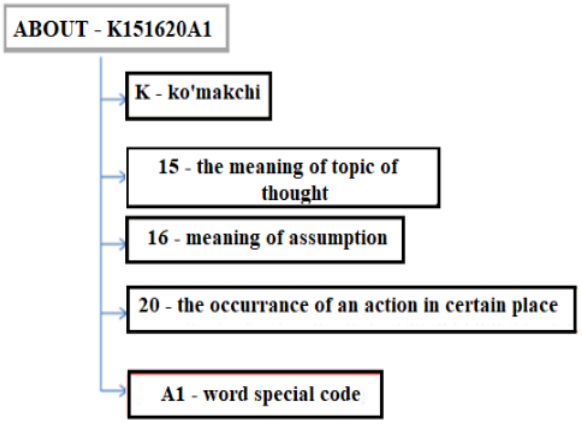

Fig.4. Encoding a word with multiple grammatical meanings in English.

The results are compared.

B. Method 2. Encoding each form and meaning separately.

This method is a logical continuation of the first method. The difference is that in this method, a separate code is given for each meaning of a specific language unit. Each grammatical meaning and the grammatical form representing it are given a separate place in linguistic support. For example, the word with has five grammatical meanings: togetherness, cooperation; condition; unobstructed; time, tool. They are reflected in the linguistic base as shown in Table II.

TABLE II. GIVING THE WORD FORM IN THE LINGUISTIC BASE.

Id	Word formative	Word classes	Grammatical meaning	Examle
B-010008	bilan	Conjunction	Cooperation	Doim birga kuylaymiz do‘stim *bilan* men
K-010051	bilan	Auxillary	Time	Qo‘ng‘iroq chalinishi *bilan* dars boshlandi
K-010019	bilan	Auxillary	Condition	Alam *bilan* kuyladi
K-010063	bilan	Auxillary	Unobstructed	Aytganim *bilan* baribir kelmaydi
K-010057	bilan	Auxillary	Tool	Qalam *bilan* yozdi

In this case, the first character represents the general grammatical meaning of the word, namely the category of the word, the next three characters represent the special code of the word, and the last three characters represent its grammatical meaning. Given in figure 5.

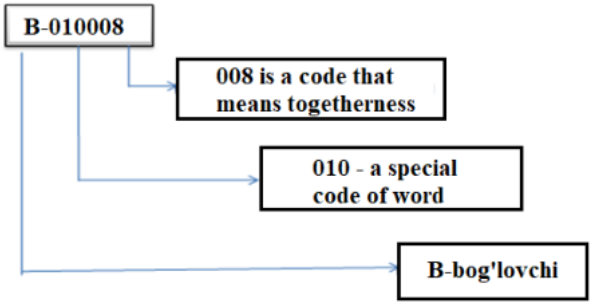

Fig.5. Encoding only one meaning of a word.

A similar database of code words is created in English. Units that correspond to each other are considered compatible. See table III.

TABLE III. THE APPEARANCE OF THE ENGLISH ALTERNATIVE OF THE WORD IN THE LINGUISTIC DATABASE

Id	Word formative	Examle
B-010008	And	My friend *and* I always sing together
K-010051	As soon as	*As soon as* the bell rang, the lesson began
K-010019	With/ -	He sang *with* pain
K-010063	Even if	*Even if* I say him, he will not come
K-010057	With	He wrote *with* a pen

IV. CONCLUSION

The following conclusions were drawn from the study:

1. Unlike lexical units, grammatical forms can be analyzed on the basis of certain forms. Because their meanings are limited.
2. When comparing morphologically different languages, the units of one language can be reduced to the grammatical criteria of the second language. This makes them easier to understand.
3. Analyzing words using the modeling method in grammatical forms gives good results.

Thus, the concept of morphological analysis was formed in applied linguistics and widely used in machine translation. The concept of morphological analysis is an operational concept, if in traditional linguistics it includes only the description of word forms and "what is classified?" to answer the question, in computer linguistics it is not important, but the main attention is focused on how to get this or that information from the word form in the text. However, these two areas cannot be separated from each other. Because computational linguistics is based on theoretical linguistics.

REFERENCES

[1] Z.T. Kholmanova Introduction to Linguistics. Tashkent -2007, p. 300.

[2] I.Zaliniskaya Morphological analysis. Vidya ma. // Linguistic support for morpnoanalyzers. Slovakia,2020.

[3] N. Abdurahmonova Linguistic support of machine translation Academy, Tashkent - 2018, page 31

[4] N. Abdurahmonova Morphological analysis in machine translation (as an example of a noun phrase model). // Scientific methodical electronic journal of foreign languages in Uzbekistan. Tashkent, 2017. p. 40-51.

[5] M.Raab Language model techniques in machine translation // Diplomarbeit. Karnegi, Germany, 2006.

[6] G. Lembersky, N.Ordan, Sh. Wintner Language Models for Machine Translation: Original vs. Translated Texts // Association for Computational Linguistics: Volume 38, Number 4. USA, 2012 - p. 799-826

[7] C.Tantuğ, E. Adalı, K.Oflazer,. Lexical Ambiguity Resolution for Turkish in Direct Transfer Machine Translation Models, The 21st International Symposium on Computer and Information Sciences, ISCIS 2006, İstanbul, Turkey, Lecture Notes in Computer Science, Vol. 4263, pp. 230-238, Springer. 2006

[8] Z.Sagay A computer translation from English to Turkish. Yüksek Lisans Tezi, Bilgisayar Mühendisliği Bölümü, ODTÜ. Ankara. 1981

[9] İ.Hamzaoğlu Machine translation from Turkish to other Turkic languages and an implementation for the Azeri languages, in Institute for Graduate Studies in Science and Engineering. MSc Thesis İstanbul: Bogazici University 1993.

[10] K.Altıntaş, Turkish to Crimean Tatar Machine Translation System, in Bilgisayar Mühendisliği Bölümü. MSc Ankara: Bilkent Üniversitesi 2000.

[11] M.Orhun Machine translaton between uyghur language and turkish // dissertation of Dsc. – Turkey, 2010.

[12] B.B. Elov thematic modeling by SVD and NMF methods. Language and culture of Uzbekistan // Computer Linguistics. Tashkent, 2023 Vol 2(6) p. 55-66.

[13] Sh. Hamroyeva Linguistic support of the morphological analyzer of the Uzbek language // Diss. of DSc. Tashkent, 2021.

[14] S. Bozorova Methods of syntactic analysis based on free context grammar // International Journal of Word Arts, Vol 6. Tashkent, 2021. p. 71-76

[15] S. Muhamedova Creating linguistic support for computer programs based on verbs of action.-Tashkent, 2006.

[16] S.Khudayarova The problem of modeling of phrases in the Uzbek language // "Computer linguistics: problems, solutions, prospects" international scientific and practical conference: Vol. 1 no. 01 (2023) Tashkent, 2023. p. 289-294.

[17] M. Suyunova On the modeling of grammatical homonyms. // Perspectives of Uzbek applied philology // Vol 1, Tashkent, 2023.

[18] N. Abdurahmonova Morphological analysis in machine translation (as an example of a noun phrase model). // Scientific methodical electronic journal of foreign languages in Uzbekistan #2-2017. Tashkent, 2017 p. 40-51.

[19] R.Zangenfeind, B. Sonnenhauser Russian Verbal Aspect and Machine Translation. Vypusk 13. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue" (2014), P. 744–745.)

[20] M. Mirtojiev Polysemy in the Uzbek language. -T.: FAN, 1975. -p. 47