ACADEMICIA
An International Multidisciplinary Research Journal
(Double Blind Refereed & Peer Reviewed Journal)

DOI: 10.5958/2249-7137.2021.01324.0

# THE MORPHOLEXICON OF THE UZBEK LANGUAGE AS A SOURCE FOR THE CORPUS

**Ural Menglievich Kholiyorov***

*Senior Lecturer of the Department of Uzbek Linguistics,
Termez State University, UZBEKISTAN
Email id: xoliyorov@tersu.uz

## ABSTRACT

*The article analyzes the issue of morpholexicon, which is considered a necessary database for corpora. Morpholexicon is compared with grammatical dictionaries. World experience, theoretical views and research in the Uzbek language have been studied. In addition, the initial works on morpholexicon is described, and in the article the number parts of speech of the Uzbek language is described for the first time. The results are presented on the basis of tables and graphs, and proposals for further work are developed.*

**KEYWORDS:** *Corpus, Grammatical dictionary, Morpholexicon, Automatic Morphological Analysis, Digital Indexes, Morphoclass, Morphological Dictionary.*

## INTRODUCTION

Attention to the state language in our country has increased to the level of one of the priority tendencies. Therefore, preserving our mother tongue, to enrich, along with increasing the effectiveness of practical use of it, further improving the status of the state language, to achieve wide usage of Uzbek language in modern information and communication system, for this to study of our national cultural heritage, increasing electronic resources in education of mother tongue, as well as, o achieve the unrestricted use of these resources by educators, to collect cultural heritage serving national and spiritual education –native language materials on a platform have become an urgent task. In the implementation of this task, exactly the language corpora, especially the Educational corpora takes the main place.

For thisthe formation of the most necessary morpholexicon for the corpora is an important component.

**Analysis of literature on the topic**

A number of monographic studies and articles have been discussed about the principle of creating a morpholexicon in world corpus linguistics,method,theoretical basis and software tools.[8, 13, 12, 10, 9].In world linguistics there is also the experience of creating grammatical dictionary, in practice it is observed that grammatical dictionary of several languages is created. We discuss about grammatical dictionary of natural language, morpho lexicon, its features as well as the possibilities of using them in the processing of natural language

**Grammatic dictionary**-a dictionary covering the total lexeme of a particular language, all their grammatical forms. A.A. Zaliznyak's grammatical dictionary [12] not only shows the grammatical changes of modern Russian words (noun, adjective, number, declension, conjugation of verbs), but also serves as a reverse dictionary. In addition to there are all the variants of diversification and conjugation about large volume of theoretical and descriptive information except list of words in introduction part of dictionary.[19]. The features of diversification proper nouns are expressed in the next filled edition. Although the dictionary was published in 1977, it has been reprinted several times with additions and corrections. The dictionary is available in paper and electronic versions, the electronic version is used in instruments of processing natural language: orthographic editing, machine translation, automatic referatization. A.A.Zaliznyak'sgrammatic dictionary is a fundamental study of morphology. Information about group, stress, genus, view of every words, transitive and intransitive ( for verbs) are attached[18].The author himself gives the dictionary the following definition:“The grammar dictionary shows the grammatical changes of modern Russian words.The dictionary contains 100,000 words arranged in reverse / inversion:the word is in the order of the letters at the end of the word, not the beginning letter.Each word is marked with an index / index referring to the grammatical information: the user diversify and conjugate interesting words through this index based on the rules of diversification and conjugation.The first edition of the dictionary to the present, the changes in the Russian language are taken into account, filled[18].

The information on the word changes of the general morphology of the Russian language given in Wikislovar is based on the dictionary of AA Zaliznyak.The indexes proposed by A. Zaliznyak were taken into account in the formation of the paradigm of word changers: numbers, written letters in the Latin alphabet are based on the classical method of indexing words (Zaliznyak`s principles).Even in the automatic morphological analysis of the word assimilation, the same system is preserved[20].According to A. Zaliznyak'sGrammar dictionary and Wikislovar, numerical indices and accents ending in hard vowels, ending in soft vowels, as well as *г, к, х, ж, ш, ч, щ,ц*and y multi-letter indexes are included.There is a need to enter this data into the index, which provides information about thematic property in the process of compiling the morpholexicon of the Uzbek language.

But since the emphasis in the Uzbek language often does not differ in meaning, it is not necessary to enter an index indicating the word accent. Morpholexicon is a database that covers a specific language lexical, in which information about its category affiliation, grammatical feature is attached to the lexeme, adapted to the natural language processing process. Such a database is available in electronic form and serves as a linguistic support of various means of processing natural language (automatic translation, linguistic corpus, morphological, semantic, graphemic analyzer).

Based on our observations, we can say that the morphological dictionary contains indicators of change of the word belonging to a category, belonging to a grammatical category, variation of the stem and its relation: these characters serve as a search filter; returns the result to the query with the exact, desired, information. A grammatical dictionary serves to describe the grammar of a particular language; its main function is not automatic processing of natural language. And the morph lexicon is distinguished by the fact that it fully covers the language lexical, is available as a database, is a linguistic supply of natural language processing tools. The morphological dictionary of language and morpholexicon are complementary information banks.

A.M.Galieva, A.R.Gatiatullin, studying the importance of the morphological category of the verb in the construction of the model of Turkish word form suffixes [9], describe the possibility of performing morphological analysis through the base.

If a word in one of the specified Turkic languages is entered in the data entry window of the software, the result window shows the sequence of morphemes and morphological categories according to the word morpheme structure. It may seem that the designation of morphemes and morphological categories in a certain word forms is superfluous. There are certain reasons for this: In Turkic languages, the same morpheme (grammatical form) can represent several categories [9].

For example, in Uzbek, *-ган, -иб, -ди*represent different grammatical categories depending on their position. Therefore, the suffixes morpholexicon such forms require a special description. Depending on its position, it is the task of the morphoanalyzer, not the morpholexicon, to show what the same form represents.There are special studies on the position of additives in the morpho analyzer [17],therefore we will not dwell specifically on this issue. There are several articles on the theoretical foundations and practice of morpholexicon formation[14, 1].D.Selegey, T.Shavrina, V.Selegey, S.Sharovlar in the article“Автоматическаяморфоразметкакорпусоврусскоязычныхсоциальныхмедиа: обучение и оценкакачества”[14]problems of categorization of words in the language of media materials,the types of automatic category detectors, their advantages and disadvantages are discussed and develops suggestions and recommendations for their use, categorization of words that occur in the language of these materials. I.A.Bolshakov, E.I.Bolshakova in the article on the creation of an automatic morphoclassifier of Russian proper nouns [6] expressed valuable views on the automatic differentiation of word groups; they can be used in the formation of the morpholexicon of the Uzbek language. I.A.Bolshakov, E.I.Bolshakova writes about giving information about the belonging of compound names to the category:«42 percent of the morpholexicon, known as the Crosslexics of the Russian language, consists of compound names. Compound nouns can consist only of a noun, but also of a noun, adjective, number, pronouns, adverb. For example: *точка зрения, вид на жительство, право быть избранным, свободная экономическая зона, свободно конвертируемая валюта, семь смертных грехов, теория вероятностей, математическаястатистика.*

During the gradual development of cross-lexicon, it became clear that this morpholexicon was enriched with 15 percent of the words present in A. Zaliznyak's dictionary.When the user performs a survey of compound names or enter compound names for analysis, in the result window he can see the change in the paradigm of the agreements of this word, even the names that do not exist on the basis of Crosslexics”[6].The authors also share their experience in

morphosystem separation.The morphosins they suggest are specific to the Russian language: genus,одушевленность /неодушевленностьand etc.

A.Fadoua, B.Siham's article «Morpho-Lexicon for standard Moroccan Amazigh» also describes the experience of composing the morpholexicon of the Moroccan language[1].

**Research methods**

Methods of classification, description, comparison, statistical analysis were used to cover the research topic.

**Analysis and discussion of results**

The following is a description of the practice of composing the morpholexicon of the Uzbek language for the educational corpus of the Uzbek language on the basis of the study of the theory and experience of morpholexicon in computer linguistics.

In order to perform a search based on morphological features in the educational corpus of the Uzbek language, the corpus database must contain a table containing information indicating that all words in the Uzbek language belong to a category – the morpholexicon of the Uzbek language. We used the dictionary "Explanatory dictionary of the Uzbek language" [15] (more than 80,000 words and phrases) published in 2000-2006 to form the morpholexicon dictionary. This dictionary lists a total of more than 80,000 words and phrases, but the dictionary - those listed as keywords – makes up 32,500 lexemes; the remaining 47,000 units are compound words, phrases, and commonly used phrases based on the same lexeme. Therefore, the morphology of the Uzbek language is developed in the following four stages.

First stage: 32,500 lexemes from the Uzbek dictionary were included in the morpholexicon database; they were accompanied by information about belonging to the category.

The second stage: 4700 words were allocated from the dictionary of the Uzbek language from The Dictionary pair and repeated words of T.Jumaev in co-authorship with T.Valievwhich were published for schoolchildren and contain 8000 units,[4]and included in the morpholexicon. Since the compound words belong to a different category, they were also given information about their belonging to the category. A new 32500 joint, double and repeated words were introduced into the morpholexicon base with 4700 Lexis: as a result, the base of the morpholexicon, prepared in the second stage, amounted to around 37200 lexical quantities.

Statistical analysis of the morpholexicon base has so far shown the following result:

| № | Parts of speech | amount | percent |
|---|---|---|---|
| **1** | **noun** | 22599 | 60,62 |
| **2** | **adjective** | 7179 | 19,26 |
| **3** | **verb** | 3949 | 10,59 |
| **4** | **adverb** | 1856 | 4,98 |
| **5** | **imitation word** | 800 | 2,15 |
| **6** | **Interjection** | 244 | 0,65 |
| **7** | **pronoun** | 182 | 0,49 |
| **8** | **Modal word** | 132 | 0,35 |
| **9** | **polunctional word** | 108 | 0,29 |

| | | | |
|---|---|---|---|
| **10** | **postposition** | 106 | 0,28 |
| **11** | **conjunction** | 52 | 0,14 |
| **12** | **Particle** | 45 | 0,12 |
| **13** | **number** | 30 | 0,08 |
| | **total** | **37282** | **100** |

The above statistic is the first morpholexicon to be implemented, which may certainly be free of minor errors and omissions. But these indicators give a preliminary idea of the category of lexemes in the Uzbek language. When you give the words of the existing base in percentages, the following form the view:

The third stage still requires processing, so the amount of content in this lexicon is unknown. At this stage, the Explanatory Dictionary of the Uzbek language and all the phrases given in the explanatory phrase logical Dictionary of the Uzbek language are distinguished and information about their belonging to the category is attached.

In the fourth stage, the compound names given in the explanatory dictionary of the Uzbek language are selected and the information on belonging to the category is attached. The content of this lexicon is also determined when the database is ready. All of these processes require manual labor: if all of these units are included in a database, they can be used in a variety of linguistic operations. In addition to information about the affiliation of words to the morpholexicon of the Uzbek language, it is necessary to attach grammatical symbols (morphos ins) based on the characteristics of the Uzbek language.

The group of morphos ins is different for each parts of speech. Sh.Hamroeva develops a system of morphological, syntactic and semantic tags while studying the problems of grammatic tapping of corpus material[16].The complexity of the system of tags proposed by Sh. Hamroeva makes it difficult to automatically annotate words.The experience of corpus linguistics shows that the simplification of the system of tags simplifies the practice of annotation (tagging).

## CONCLUSIONS AND SUGGESTIONS

During the examination and preliminary testing of morpholexicon, the following problematic cases were identified.

First of all, the word homonym cannot be identified only by information about belonging to the category on the basis of morpho lexicon. The morpholexicon database shows the homonyms present in the language; a tag belonging to the category is assigned, but this information is not sufficient to identify the homonymous unit in context (in the case of a grammatically formed).Grammatical formation is the most important factor that helps to distinguish the homonymous unit; only this requires the activation of a morphological-syntactic analyzer, in particular the development of a homonymous unit filter.This work will be the subject of a separate study.

Second, in the morpholexicon, language units that reflect the feature of polyfunctionality in the Uzbek language were identified, and they were marked as polyfunctional in the database on their belonging to the category.Naturally, this also leads to a lack of clear indication of affiliation to the category. The meaning of such unity is realized only in the text.The issue of automatic

determination of the category of polyphonic words also requires special study. The primary and secondary functions of the word categories in the Uzbek language [7], computer processing of polyphonic words and suffixes [3, 11, 2, 16] were the objects of Special Research. However, in Uzbek computer linguistics there is no development of the algorithm for automatic detection of polyfunctional words. Therefore, we will analyze this issue specifically in subsequent studies.

**REFERENCES:**

1. Fadoua A., Siham B. Morpho-Lexicon for standard Moroccan Amazigh /MATEC Web of Conferences 210, 04024 (2018). – CSCC 2018 // https://doi.org/10.1051/matecconf/201821004024
2. Gulyamova Sh. An'anaviy tilshunoslik va kompyuter lingvistikasida polifunksionallik masalasi // Davlat tili − ijtimoiy taraqqiyot va milliy yuksalish mezoni (O'zbek tiliga Davlat tili maqomi berilganligining 31 yilligiga bag'ishlangan respublika ilmiy-amaliy konferensiya materiallari). – Buxoro, 2020-yil, 16-oktabr. – B. 369-377.
3. Gulyamova Sh. The problem of polyfunctionality in the traditional and computational linguistics // Middle European scientific bulletin. – Volume 5, October 2020. ISSN 2694-9970. IFSIJ Impact Factor: 5,985. – P. 104-107.
4. Jumayev T, Valiyev T. Qo`shib yoziladimi, ajratib (Maktab o'quvchilari uchun). – Тошкент, 2006. – 76 b.
5. Абжалова М. Таҳрир ва таҳлил дастурларининг лингвистик модуллари. Монография. – Тошкент: Нодирабегим, 2020. – 176 б.
6. Большаков И.А., Большакова Е.И. Автоматический морфоклассификатор русских именных групп // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 30 мая–3 июня 2012 г.). Вып. 11 (18): В 2 т. Т. 1: Основная программа конференции. – М.: Изд-во РГГУ, 2012.– 774 с. – С. 81-93.
7. Ботирова А.Э. Ўзбек тилида сўз туркумларининг бирламчи ва иккиламчи вазифасининг функционал-синтактик таҳлили: филол. фан. бўйича фалсафа доктори (PhD) диссер. автореф. – Қарши, 2018. – 56 б.
8. Володин А.П., Храковский В.С. Об основаниях выделения грамматических категорий // Проблемы лингвистической типологии и структуры языка / Отв. ред. В.С.Храковский. – Л.: Наука, 1977. – С.42-54.
9. Галиева А.М., Гатиатуллин А.Р. Обозначение морфологических категорий глагола в моделях окончаний тюркских словоформ // Компьютерная обработка тюркских языков. Первая международная конференция: Труды. – Астана: ЕНУ им. Л.Н.Гумилева, 2013. – 340 с. – С. 171-181.
10. Гибадулин Р.Я., Гибадулин Я.Н., Сакаев А.Р., Закиев М.З., Саламатин И.М. Электронные словари тюркских языков // Компьютерная обработка тюркских языков. Первая международная конференция: Труды. – Астана: ЕНУ им. Л.Н.Гумилева, 2013. – 340 с. – С. 156-159.

**11.** Гулямова Ш. Полифункционаллик хусусида // “Ўзбекистонда илмий-амалий тадқиқотлар” мавзусидаги онлайн конференсия. 15-август. № 19. – Тошкент, 2020. – Б. 46-48.

**12.** Зализняк А.А. Грамматический словарь русского языка. Словоизменение − М.: Русский язык, 1980. − 880 с.

**13.** Плунгян В.А. Общая морфология: Введение в проблематику. – М.: Едиториал, 2003. – 384 с.

**14.** Селегей Д., Шаврина Т., Селегей В., Шаров С. Автоматическая морфоразметка корпусов русскоязычных социальных медиа: обучение и оценка качества // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2016» Москва, 1-4 июня 2016. / URL: https://www.academia.edu/26571774/Автоматическая_морфоразметка_корпусов_русскоязычных_социальных_медиа_обучение_и_оценка_качества.

**15.** Ўзбек тилининг изоҳли луғати. 5 томлик. – Тошкент: Ўзбекистон Миллий энциклопедияси, 2000-2006.

**16.** Хамроева Ш. Ўзбек тили морфологик анализаторининг лингвистик таъминоти. Монография. (Электрон китоб) – GloeEdit, 2020. – 244 б. – Б. 168.

**17.** Хамроева Ш. Ўзбек тили муаллифлик корпусини тузишнинг лингвистик асослари. Филология фанлари бўйича фалсафа доктори (PhD) диссертацияси. – Бухоро, 2018. – 165 б.

**18.** https://gramdict.ru/

**19.** https://ru.m.wikipedia.org/wiki/Грамматический_словарь_русского_языка_А._А._Зализняка

**20.** https://ru.wiktionary.org/wiki/Викисловарь:Использование_словаря_Зализняка