№ 3(49) 2023

ISSN 2181-8460

HISOBLASH VA AMALIY MATEMATIKA MUAMMOLARI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

PROBLEMS OF COMPUTATIONAL AND APPLIED MATHEMATICS

RAQAMLI TEXNOLOGIYALAR VA
SUN'IY INTELLEKTNI RIVOJLANTIRISH
ILMIY-TADQIQOT INSTITUTI

ПРОБЛЕМЫ ВЫЧИСЛИТЕЛЬНОЙ И ПРИКЛАДНОЙ МАТЕМАТИКИ

№ 3(49) 2023

Журнал основан в 2015 году.
Издается 6 раз в год.

Учредитель:
Научно-исследовательский институт развития цифровых технологий и искусственного интеллекта.

Журнал зарегистрирован в Агентстве информации и массовых коммуникаций при Администрации Президента Республики Узбекистан.
Регистрационное свидетельство №0856 от 5 августа 2015 года.

ISSN 2181-8460, eISSN 2181-046X

Адрес редакции:
100125, г. Ташкент, м-в. Буз-2, 17А.
Тел.: +(99871) 231-92-45.
E-mail: journals@airi.uz.
Сайт: journals.airi.uz (www.pvpm.uz).

Дизайн и компьютерная вёрстка:
Шарипов Х.Д.

Отпечатано в типографии НИИ РЦТИИ.
Подписано в печать 26.06.2023 г.
Формат 60x84 1/8. Заказ №4. Тираж 100 экз.

Содержание

UDC 004.9+81'322.2:811.512.133

DESIGNING A SYSTEM OF MORPHOLOGICAL ANALYZER OF THE UZBEK LANGUAGE

Bakaev I.I.

i.i.bakaev@buxdu.uz

Bukhara State University,

705018, Muhammad Ikbol 11, Bukhara, Uzbekistan.

A brief review of existing morphological analyzers for agglutinative languages is given. An architecture of a morphological analyzer for the Uzbek language is proposed, an object-oriented model and a diagram of which are described using UML(Unified Modeling Language)".

Keywords: computational linguistics, UML, tokenization, stemming, state machine.

Citation: Bakaev I.I. 2023. Designing a system of morphological analyzer of the Uzbek language. *Problems of Computational and Applied Mathematics.* 3(49): 135-142.

1 Introduction

The morphological analyzer is an integral part of solving the subtasks of search engines (machine translation, information retrieval, information extraction), electronic libraries, archives, question-answer systems and workflow. The morphological analyzer is software that solves a number of complex tasks, such as determining grammatical features, stems, lemmas, morphemes and parts of speech of the word form Fig.1.

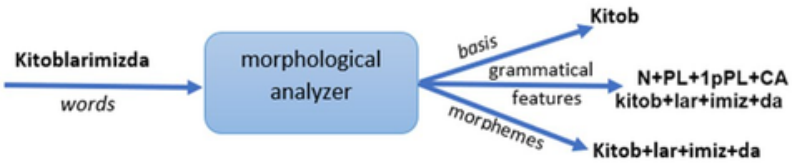

Figure 1 Formal scheme of the morphological analyzer.

There are many approaches to building a morphological analyzer: state machine, state machine with output, deterministic acyclic state machine, rule-based approaches, corpora, templates, two-level morphology and hybrid approaches. Based on them, a lot of ready-made solutions for groups of inflectional and agglutinative languages of the world have been created Fig.2.

Only a certain number of these natural languages are provided with the necessary linguistic resources and morphological analyzer tools for natural language processing. Most scientific research is based on a combinatorial mixture of several morphological approaches. Since in the article we will design a morphological analyzer on the example of the Uzbek language, some studies related to agglutinative languages are described below.

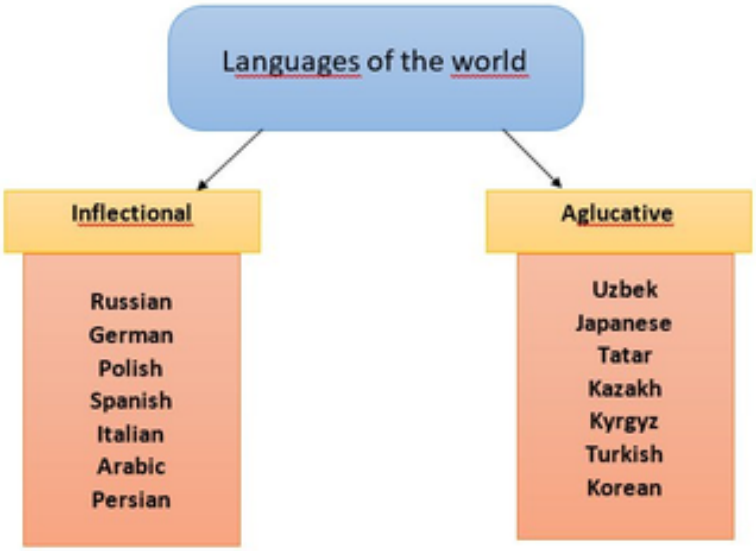

Figure 2 List of inflectional and agglutinative languages of the world.

2 Related works

A number of scientific works are being carried out on such languages as Turkish, Kazakh, Turkmen, Uzbek, Kyrgyz, Tatar, belonging to the family of Turkic languages, and are included in the group of agglutinative languages.

In scientific papers on the Turkish language, morphological analyzers such as TRmorph [1] and TRMOR [2] have been developed , based on the two-level morphology approach, which are implemented using SFST . TRmorph consists of three components: a state machine represented by regular expressions, a set of phonology and spelling rules, and a dictionary of 37,101 words belonging to 9 categories of parts of speech. The lexicon of the TRMOR morphological analyzer is compiled from the Wikipedia resource. When TRMOR and TRmorph were tested using 108 words, TRMOR gave 72% accuracy and TRmorph gave 38% accuracy [2].

The authors of [3] also described a morphological analyzer of two-level morphology for the Turkish language, which consists of five modules: a finite state machine with an output, a rules engine for suffixes (a file in .xml format), a lexicon (database), a data structure in the form of a tree, and a type cache LRU (least recently used). The lexicon consists of 54,000 words, of which 19,000 are nouns.

In studies [4], [5], morphological analyzers based on the rules for the Kazakh language were developed. To implement the rules of alternation and morphotactics (the set of rules that define how morphemes (morpho) can touch (tactics) each other) of two-level morphology, a state machine with the output of Foma (a finite-state compiler and C library) and XFST (Xerox Finite State Tools) is used. The main function of the morphological analyzer is disambiguation for the Kazakh language, which is a variation of the Brill tagger.

Also, for the Turkmen language [6], a morphological analyzer was developed with the approach of two-level morphology. The grammatical rules of the language are implemented using a state machine of the XFST type. The authors outlined [7] the design and development of a software package that includes linguistic resources (thesauri and ontologies) and text processing tools (word form generator, morphological analyzer and

morphological word disambiguation tool) based on the corpus approach for the Kazakh language.

The article [8] describes the developed morphological analyzer for the Kazakh, Tatar and Kumyk languages. To implement morphotactic rules, a state machine like HFST (Helsinki Finite-State Toolkit) is used. The total number of words in the lexicon for the Kazakh language is 11224, for Tatar - 10737, for Kumyk 4845 words.

In this article [9] a morphological analyzer tool for the Uzbek language called MorphUz is discussed. MorphUz works on the basis of a two-level morphology approach, this approach can determine the stem of words [10] and analyze suffixes.

The study [11] developed a morphological analysis tool UZMORPP for the Uzbek language. The tool includes morphotactic and morphophonemic rules of the Uzbek language, which is implemented using logic programming in the Prolog programming language.

In the work [12] an electronic dictionary of Uzbek word endings was proposed to determine the morphological bases of words. The electronic dictionary was analyzed using the Kazakh morphological analyzer of the stemming type [13], [14].

Although the works presented above provided important fundamental and practical results, they were not designed as a complete system for morphological analysis. Therefore, the purpose of this task is to design systems for all tasks of the morphological analysis of the Uzbek language.

3 Theoretical analysis

The vocabulary of the Uzbek language is constantly updated with new words. Replenishment occurs through the formation of new words with the help of affixation and composition. Affixation and composition are the main ways of forming words. The affixal way of forming words is the most productive in the Uzbek language. In this way, new words are created by adding derivational affixes to different types of bases. The compositional method is the creation of new words by combining two or more stems. As a result of the above two methods, the following types of words are formed in the Uzbek language in terms of structure and formation Fig.3.

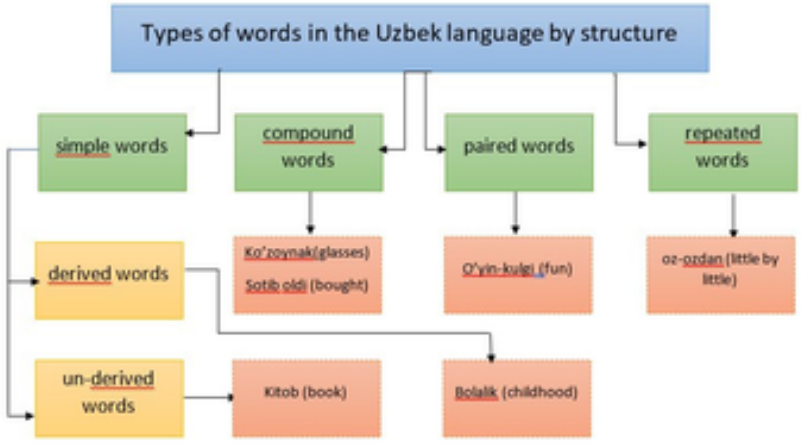

Figure 3 Types of words in the Uzbek language according to their structure.

There are six independent parts of speech in the modern Uzbek language: noun, adjective, numeral, pronoun, verb, adverb. Each part of speech in structure has the

above types of words. Based on the above theoretical information, we have identified the following entities:

- words;
- types of words;
- parts of speech;
- affix;
- types of affixes.

Next, we define entities for the tasks of the morphological analyzer. These include the following:

- Splitting text into tokens;
- Definition of the basis of words;
- Definition of morphemes and grammatical features.

4 Formulation of the task

Taking into account the analysis carried out, the following tasks are solved in the article:

- Development of a class model for a morphological analyzer;
- Building a use case diagram for specific user groups.

5 Solution method

Based on the above theoretical material, we determine the following classes in Fig. 4:

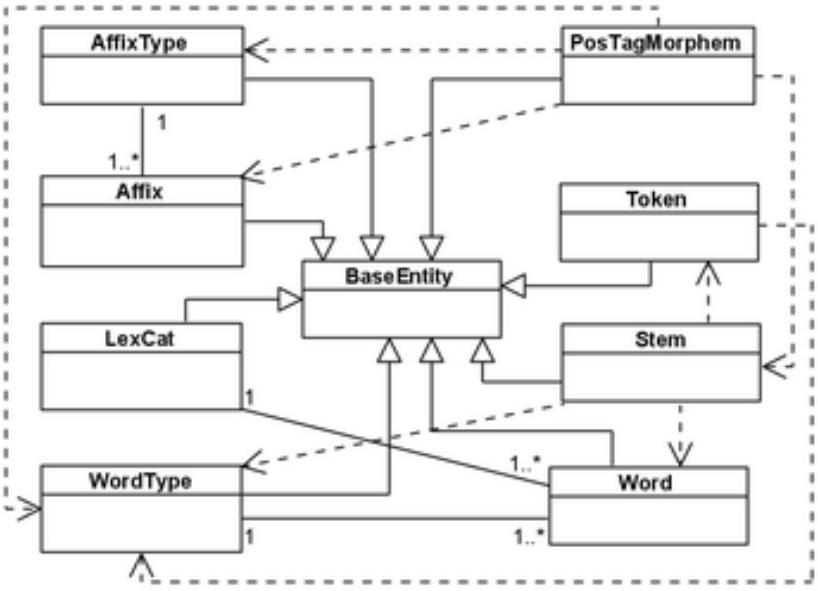

Figure 4 Object-oriented model of the morphological analyzer.

"BaseEntity"is an abstract class, all classes inherit from this class. This class consists of fields representing an identifier (id) and a value (value).

"Words"is a class representing the "Word"entity. The class has its own "WordParent "field to determine what word it was created from.

"Word type"is a class that represents the entity "Type of words". "Affix "is a class that represents the "Affixes"entity. "AffixType "is a class that represents the "Affix

Types"entity. "LexCat "is a class that represents the Lexical Categories entity, that is, parts of speech.

All classes represented in class diagrams are inherited from the BaseEntity class. Fields inherited from the BaseEntity class are sufficient for "WordType "Affix "AffixType "LexCat"classes. Therefore, these classes do not have their own fields. These fields can be used to store the name of the word type, affixes, affix types, lexical categories and their identifiers.

"Word_LexCat"is a class that represents the relationship between the "Words"and "LexCat"classes. The link between the "Words"and "LexCat"classes uses an additional class "Word_LexCat"to avoid homonymy of some words.

The link between the Affix and LexCat classes uses the Affix_LexCat class to define homonymous suffixes. "Stem"is a class that represents the "stem"entity. The class consists of the following fields and methods:

- "OrfoTypeWord method for determining the type of a word;
- "StemmingWordW2 a method that finds the stem of a word (stem) of type w_{sb} ;
- "StemmingWordW3W4 method for finding stems (stems) of words w_j and w_t;
- the "StemmingWordW5"method, which finds the stem of a word (stem) of w_{ab} type;
- the "StemmingWord"method, which finds the stem (stem) of words of type w_{st}, w_{sy} and w_{psy};
- "Lemmatization a method that converts word forms into a type lemma w_{st}, w_{sy} and w_{psy};
- "isprefix a method that determines whether the word has a prefix or not;
- "prefixLen method for determining prefix length;
- "ReadRule"– method for reading rules;

The algorithms of the StemmingWordW2, StemmingWordW3W4, StemmingWordW5 and StemmingWord methods were implemented in [15] earlier.

"Token"is a class that breaks text into words. The class contains the "Tokenlash" [16] method , which splits the text into words.

"POSTagMorphem"is a class that defines a word into morphemes and morpheme properties [17]. The class consists of the following methods:

- "isbelogswordlang a method that checks if a word belongs to a language;
- "POSTaggingMorphemic method of marking affixes with grammatical tags;
- "MorphemicList a method that generates a list of affixes;
- "MorphemicTypeList a method that generates the type of affixes;

The "POSTagMorphem"class uses a "Dependency"relationship with the classes "AffixType "Affix"and "WordType that is, as input and output values. Next, we will build a use case diagram for certain user groups. The following groups of users can be distinguished for the morphological analyzer system:

- users (student, teacher, API user) - can conduct a morphological analysis (tasks tokenization, definition of word stems, morphemes and grammatical features) of words using the system Fig.5.
- - expert linguist - a specialist in the field of computational linguistics who can perform the CRUD operation (an abbreviation of the English words create, read, update, delete) Fig.6.

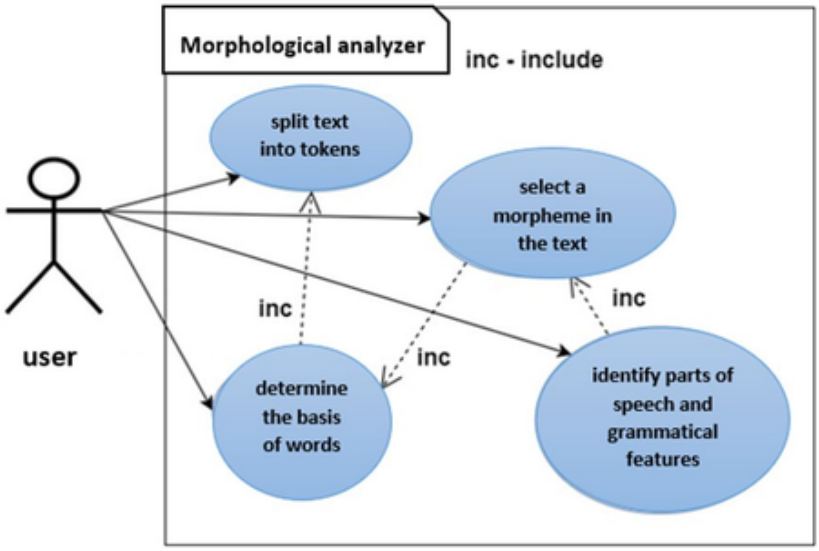

Figure 5 Diagram option using morphological analyzer for users.

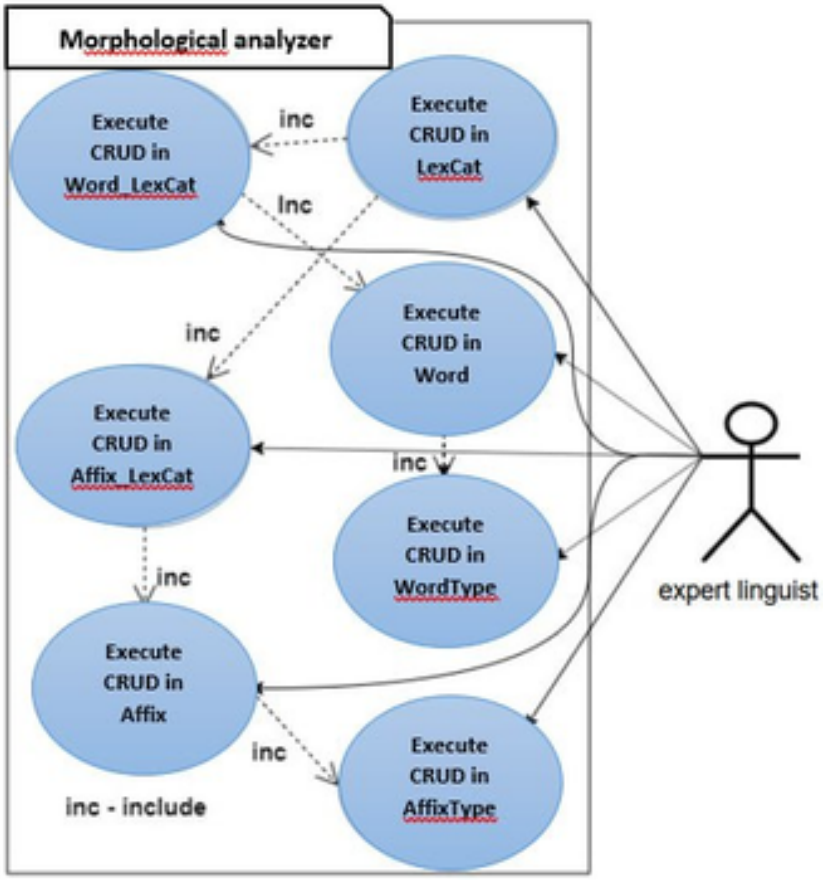

Figure 6 Diagram option using morphological analyzer for expert linguists.

6 Results

Based on the application of the developed morphological analysis of the Uzbek language words for text queries in the Uzbek language when searching for information re-

sources in the electronic library system of the Bukhara Regional Information and Library Center named after Abu Ali ibn Sino and in the information resource center of the Bukhara State University, it is possible to reduce time and labor costs for 9-11% when performing the operations of compiling bibliographic descriptions of documents in the Uzbek language in the electronic catalog system due to automatic correction of spelling errors and the issuance of the correct spelling of words.

7 Conclusion

According to the results of the scientific research analysis presented in the work, associated with the creation of a morphological analyzer for the Uzbek language that belongs to the family of agglutinative languages, models of classes of a morphological analyzer have been developed, and a diagram of the use case of a morphological analyzer for certain user groups has been built.

References

[1] Gagri G. 2010. A Freely Available Morphological Analyzer for Turkish *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. Place of publication: European Language Resources Association (ELRA). – P. 820–827.

[2] Ayla K., Helmut T., Ahmet E., Özkan K. 2019. TRMOR: a finite-state-based morphological analyzer for Turkish. *Turkish Journal of Electrical Engineering and Computer Sciences* 27(5): – P. 3837–3851. doi: http://dx.doi.org/10.3906/elk-1902-125.

[3] Olcay Taner, Yıldız. Begüm, Avar Gökhan, Ercan. 2019. An Open, Extendible, and Fast Turkish Morphological Analyzer *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*. Place of publication: INCOMA Ltd. – P. 1364–1372.

[4] Kessikbayeva G., Cicekli I. 2016. A Rule Based Morphological Analyzer and A Morphological Disambiguator for Kazakh Language *Linguist. Lit. Stud.* 4(1): – P. 96–104. doi: http://dx.doi.org/10.3906/elk-1902-125.

[5] Kessikbayeva G., Cicekli I. 2014. Rule Based Morphological Analyzer of Kazakh Language *SIGMORPHON/SIGFSM* 4(1): – P. 46–54. doi: http://dx.doi.org/10.3115/v1/W14-2806.

[6] Tantug A.C., Adalı E., Offlazer K. 2006. Computer Analysis of the Turkmen Language Morphology *International Conference on Natural Language Processing (in Finland)*. Place of publication: Springer, Berlin, Heidelberg. – P. 186–193.

[7] Akhmed-Zaki D., Mansurova M., Madiyeva G., Kadyrbek N., Kyrgyzbayeva M. 2021. "Development of the information system for the Kazakh language preprocessing *Cogent Engineering* 8(1): – P. 46–54. doi: http://dx.doi.org/10.1080/23311916.2021.1896418.

[8] Washington J., Salimzyanov I., Tyers F. 2014. Finite-state morphological transducers for three Kypchak languages *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. Place of publication: European Language Resources Association (ELRA).3378–3385.

[9] Abdurakhmonova N., Ismailov A., Sayfulleyeva R. 2022. MorphUz: Morphological Analyzer for the Uzbek Language *2022 7th International Conference on Computer Science and Engineering (UBMK)*. Place of publication: IEEE. – P. 61–66.

[10] Mengliev D., V. Barakhnin, and N. Abdurakhmonova 2021. *Development of Intellectual Web System for Morph Analyzing of Uzbek Words*. Appl. sci., vol. 11, no. 19. 528 p.

[11] Matlatipov G., V Zygmunt 2009. Representation of Uzbek Morphology in Prolog *Aspects of Natural Language Processing* 5070:54–57. doi: http://dx.doi.org/10.1007/978-3-642-04735-0

[12] Matlatipov S., Tukeyev U., Aripov M. 2022. Towards the Uzbek Language Endings as a Language Resource *Advances in Computational Collective Intelligence*. Place of publication: Springer, Cham. – P. 729–740.

[13] Tukeyev U., Turganbayeva A. Abduali B., Rakhimova D., Amirova D., Karibayeva A. 2018. Lexicon-free stemming for Kazakh language information retrieval *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*. Place of publication: SAlmaty, Kazakhstan, Cham. – P. 1–4.

[14] Tukeyev U. Turganbayeva A. Karibayeva A. Amirova D. 2021. Universal Programs for Stemming, Segmentation, Morphological Analysis of Turkic Words *International Conference on Computational Collective Intelligence*. Place of publication: Springer, Cham. – P. 643–654.

[15] Bakayev I. 2021. Development of a stemming algorithm based on a linguistic approach for words of the Uzbek language *International Conference on Scientific, Educational & Humanitarian Advancements*. Place of publication: Turkey. – P. 195–202.

[16] Bakaev I.I. 2021. Linguistic features tokenization of text corpora of the Uzbek language *Bull. TUIT Manag. commun. Technol.* 4: – P. 1–8.

[17] Ravshanov N.K. Bakaev I.I. Shafiyev T.R. 2021. Linguistic features tokenization of text corpora of the Uzbek language *Problems of Computational and Applied Mathematics* 4(28): – P. 121–131.

Received June 19, 2023

УДК 004.9+81’322.2:811.512.133

РАЗРАБОТКА СИСТЕМЫ МОРФОЛОГИЧЕСКОГО АНАЛИЗАТОРА УЗБЕКСКОГО ЯЗЫКА

Бакаев И.И.

i.i.bakaev@buxdu.uz

Бухарский государственный университет,
705018, Узбекистан, Бухара, ул. М. Икбол дом 11.

Приведен краткий обзор существующих морфологических анализаторов для агглютинативных языков. Предложен архитектура морфологического анализатора для узбекского языка, объектно-ориентированная модель и диаграмма вариант использовании которого описана при помощи UML.

Ключевые слова: компьютерная лингвистика, UML, токенизация, стемминг, конечный автомат.

Цитирование: *Бакаев И.И.* Разработка системы морфологического анализатора узбекского языка // Проблемы вычислительной и прикладной математики. – 2023. – № 3(49). – С. 135-142.