DOI: https://doi.org/10.26565/2304-6201-2025-68-11

УДК (UDC) 004.891

Shabelnyk Tetiana *Doctor of Economic Sciences, Professor;*
Head of the Department of Economic Cybernetics and System Analysis,
Simon Kuznets Kharkiv National University of Economics , ave. Sciences,
9-A, Kharkiv, 61166, Ukraine;
e-mail: Tanya.shabelnik17@gmail.com
https://orcid.org/0000-0001-9798-391X

Prokopovych Svitlana *Candidate of Economic Sciences, Associate Professor;*
Associate Professor of the Department of Economic Cybernetics and System Analysis,
Simon Kuznets Kharkiv National University of Economics , ave. Sciences,
9-A, Kharkiv, 61166, Ukraine
e-mail: prokopovichsv@gmail.com ;
https://orcid.org/0000-0002-6333-2139

An expert system for evaluating language patterns using nonparametric statistics

Abstract. The paper focuses on the research of language regularities by methods of non-parametric statistics. The emphasis is placed on how to resolve the common problems regarding the use of non-parametric statistical techniques, in particular the one of semantic marking of sentences in linguistics.
The aim of the scientific research is defined as the development of an expert system for evaluating language regularities using non-parametric statistical methods for any language.
Research methods. Methods of non-parametric statistics, IDEF4 notation, the programming language C #.
The research offers a specialized expert system that uses the C Programming Language to provide automatic questioning of the native speakers with further analysis of the lexical meaning of word combinations and phrases implemented as a Desktop Program for Windows. The program considers the possibility of taking into account of how to reveal the correspondence of phrases from dictionary files that have different expert assessments.

Keywords: *expert system, non-parametric statistical methods, language regularities, software.*

Як цитувати: Shabelnyk T., Prokopovych S. An expert system for evaluating language patterns using nonparametric statistics. *Вісник Харківського національного університету імені В. Н. Каразіна, серія Математичне моделювання. Інформаційні технології. Автоматизовані системи управління*. 2025. вип. 68. С.113-119. https://doi.org/10.26565/2304-6201-2025-68-11

How to quote: T. Shabelnyk., S. Prokopovych, "An expert system for evaluating language patterns using nonparametric statistics", *Bulletin of V. N. Karazin Kharkiv National University, series Mathematical modelling. Information technology. Automated control systems*, vol. 68, pp. 113-119, 2025. https://doi.org/10.26565/2304-6201-2025-68-11

Introduction

It is a well-known fact that the structure of any language and its development come to operate under the guidance of statistical laws. Language, as a system of communication characterized by the combination of discrete units, is represented by a certain way of organization, its particular structure and elements interconnected by subsystems of relationships. A constant increase in their number proves to be the result of the consequential evolution and development of the language. The use of statistical methods in linguistics is sure to increase the efficiency of linguistic observation, its accuracy and reliability as well as consistency and provability of linguistic assumptions.

1. Analysis Of Recent Researches And Publication

Prominent Ukrainian and foreign theorists and scholars (B. Strickland [1], V. Zhukovska [2], V. Zayats and M. Zayats [3], Ye. V. Kupriianov, N. S. Uholnikova, and O. M. Yurchenko [4], A. Sitar [6]) published their original works dedicated to the issues of using statistical methods in linguistic research.

The large number of papers comes to indicate an increasing level of interest in the subject in question. Thus, advancing statistical methods in linguistics as well as practical application of statistical criteria claims to be more forward looking and based on the latest scientific evidence can lead to generate relevant issues for further research and investigation.

It should be noted, that most of the research done in the fields of linguistics is of a qualitative nature. normal distribution law used in linguistic research [4] is not always observed in the frequency distribution of language elements. Consequently, when the latter differs from the norm, other methods to study language patterns presume to be essential. In this regard, we consider it relevant to use nonparametric statistical methods based on empirical data, which, however, do not depend on their distribution law [4]. Succinctly, to provide real patterns of particular phenomena when there is a need to study inaccurate numerical values (qualitative), valid results come to be provided by using non-parametric criteria to identify differences in language elements.

As an illustration of the problem in question, the paper attempts to evoke the necessity of studying the issues of semantic distinctiveness of sentences. In performing these, a psycholinguistic experiment, or in other words, native speakers' perception is applied as a certain valuable technique to determine the significance (not materiality) with respect to experts' differences of opinion.

We also based on the research of Wen Zhang, Taketoshi Yoshida, Xijin Tang [7] where the authors asserts that a series of experiments on classifying the Reuters-21578 documents using the representations with multi-words. We used the performance of representation in individual words as the baseline, which has the largest dimension of feature set for representation without linguistic preprocessing. Moreover, linear kernel and non-linear polynomial kernel in support vector machines (SVM) are examined comparatively for classification to investigate the effect of kernel type on their performances. Index terms with low information gain (IG) are removed from the feature set at different percentages to observe the robustness of each classification method [7].

Lytvyn, V., Vysotska, V., Hrendus M. [8] mentioned who the main part of KB is ontology as clearly structured SA's model, systematic set of terms, which explain the connections between objects of this SA. Ontologies are generally accepted and widely used in different branches of science such as nowledge engineering, presentation of knowledge, information search, knowledge management, database design, information modeling and object-oriented analysis [8].

2. The Main Tasks Of The Study And Their Significance

Thus, our main objective is defined as the development of an expert system for evaluating language regularities using non-parametric statistical methods for any language. In its general formulation, the problem is represented in the way when two sentences come to be studied by experts who further evaluate their correct and precise usage in the language. Based on the results of their evaluation, a conclusion is made regarding the perception of the distinction between two sentences (either there is one or not).

These can be illustrated by the examples in Ukrainian: "Підприємець зробив розрахунки" / "Підприємець виконав розрахунки" and accordingly in English: "The Entrepreneur made calculations" / "The Entrepreneur performed calculations".

To consider the example of expert analysis (Sentence 1 – The entrepreneur made calculations; Sentence 2 – The entrepreneur performed the calculations) one approach would be to demonstrate positive and negative assessments of sentences distinction made by 20 experts who are native speakers. The results of the evaluation are presented in Table 1.

Table 1. Expert Assessments of Sentence Distinction
Табл. 1. Експертні оцінки відмінності речень

Sentence	Expert assessment		
	Distinction available	Distinction nonavailable	Total
1. The entrepreneur made calculations	14 (a_1)	6 (b_1)	20
2. The entrepreneur performed the calculations	6 (a_2)	14 (b_2)	20

According to the Table 1, its first row is the inverse of the second. Consequently, the issues that come to reveal on how strong (weak) the difference in these rows look like have to be investigated using the criterion χ^2 [4] presented below:

$$\chi^2 = \frac{(a_1 - b_1 - 1)^2}{a_1 + b_1}, \quad (1)$$

where a_1 is the number of positive differences of the first sentence;
b_1 is the number of negative differences of the first sentence.
The result of appropriate calculations of criterion χ^2 according to the above example is represented as follows:

$$\chi^2 = \frac{(14-6-1)^2}{14+6} = 2{,}45. \quad (2)$$

In the next step, we determine the number of degrees of freedom that equals to:

$$f = (n-1) \cdot (m-1), \quad (3)$$

with n as the number of columns in the table;
whereas m is the number of rows in the table.
Thus we have the following calculation of degrees of freedom:

$$f = (2-1) \cdot (2-1) = 1. \quad (4)$$

Further on, we find the confidence limits of the criterion in the distribution table χ^2 with a reliability of 5% and 1%, thus, having values of 3.84 and 6.63, respectively. The calculated value of χ^2, which is equal according to expression (2) 2.45 is lower than the tabular value. As a result, we conclude that the differences in the answers of experts are random and with a probability of 99% it can be resolved that in native speakers' minds these sentences are distinguished by certain difference.

3. Major Research Results

Forming a database of results of expert evaluations of lexically unambiguous phrases is a very painstaking and time-consuming process. We also note that there are difficulties in further analysis of large volumes of texts, finding correspondences in it from the database of phrases as well as the introduction of new phrases and their evaluation by experts that are native speakers. The rise of digital challenges that result in accelerated digitalization of information with its rapid growth and development, has a considerable impact on linguistics and science, in particular linguistic analysis of the text.

The latter, as known, with all its semantic unambiguity requires maximum automation. With this aim, the study has developed specific software. To carry out the automatic survey, an expert system has been developed by querying native speakers and further analyzing the lexical meaning of the phrase in accordance with mathematical expressions (1) – (4) which is implemented in the form of a desktop program.

To select the programming language of the expert system, we have formulated the following requirements: the software product must be implemented for Windows, provide a high-quality and friendly interface, and thus, there is no need for any distributed application.

At the present stage of software development tools, there have been many possible options to choose different resources to implement the task.

According to specialized ratings, the most popular programming languages are Python, Java, C #, C ++, each of them has a number of advantages and peculiar features. Python programming language is the option, which is convenient enough when one deals with the first language most frequently used for academic purposes. However, this language significantly loses in performance to C and Java, so it is not used to develop highly loaded applications and, consequently, to process large amounts of text.

The C # language is based on Java, only with a syntax closer to C ++ and is used mainly for the development of software products for the NET platform. Windows Framework [9]. Thus, the most effective way to solve this problem is to use the programming language C #. The Microsoft Visual Studio IDE has been developed especially for the C # language.

The algorithm of the program is presented in the form of an IDEF4 diagram (Fig. 1).

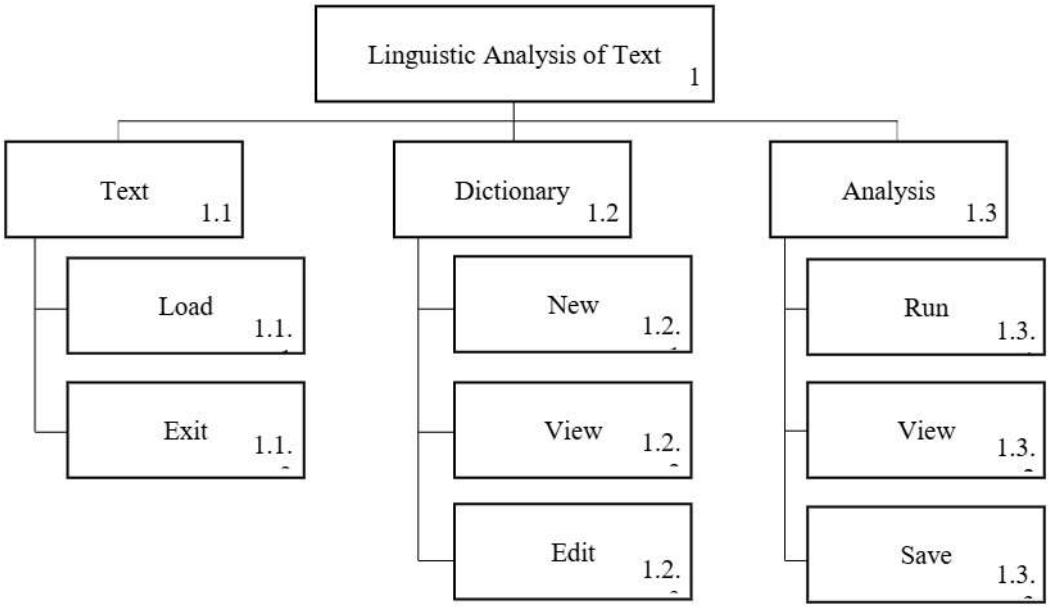

Fig. 1. Algorithmic Model of Expert System
Рис.1. Алгоритмична модель експертної системи

In the linguistic analysis of the text, there are three main actions – reading the text, compiling dictionaries of phrases with their expert assessments and directly assessing the text for linguistic homogeneity. The software does not provide the ability to edit the source document. The program reads text in docx, rtf or txt formats.

The algorithm provides the ability to create a new dictionary of phrases with the introduction of their expert assessments, viewing and downloading existing dictionaries.

Fig. 2 represents information model of the program. The text file and the dictionary file are loaded into RAM in a dialog mode. Depending on the search results of phrases in the text, the criterion χ^2 and the degree of freedom f are calculated.

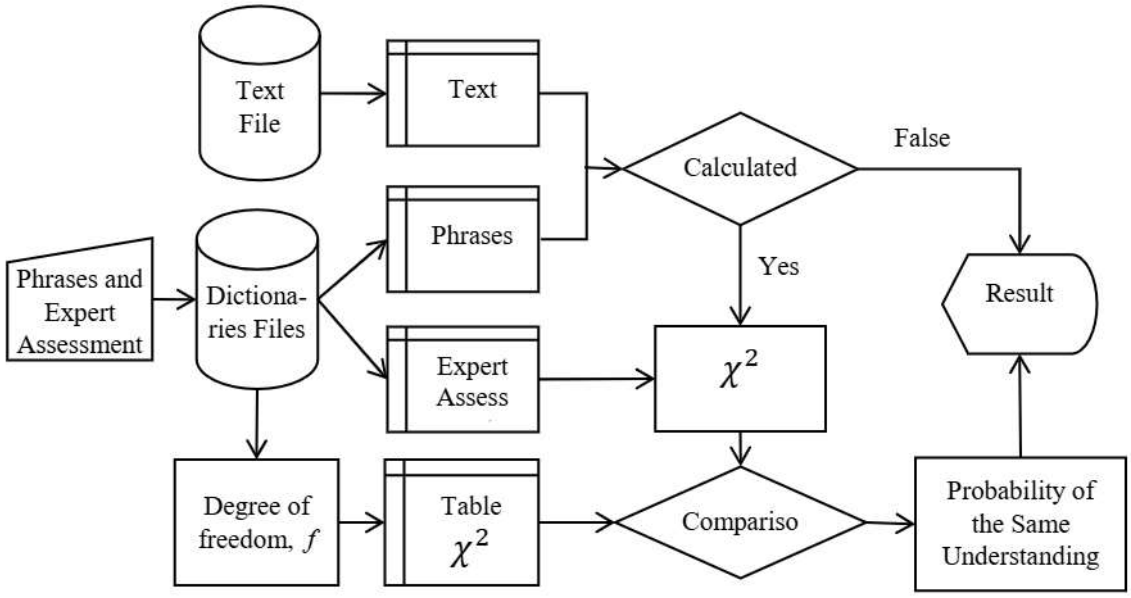

Fig. 2. Information Model of the Expert System
Рис.2. Інформаційна модель експертної системи

When analyzing the text, it is possible to match several phrases from dictionary files with different expert assessments. In this case, the total probability of the same understanding of the text is calculated as the product of the probabilities of understanding each separate phrase:

$$P_{com} = P_1 \cdot P_2 \cdot \ldots \cdot P_n, \quad (5)$$

where n is the number of identical phrases found.

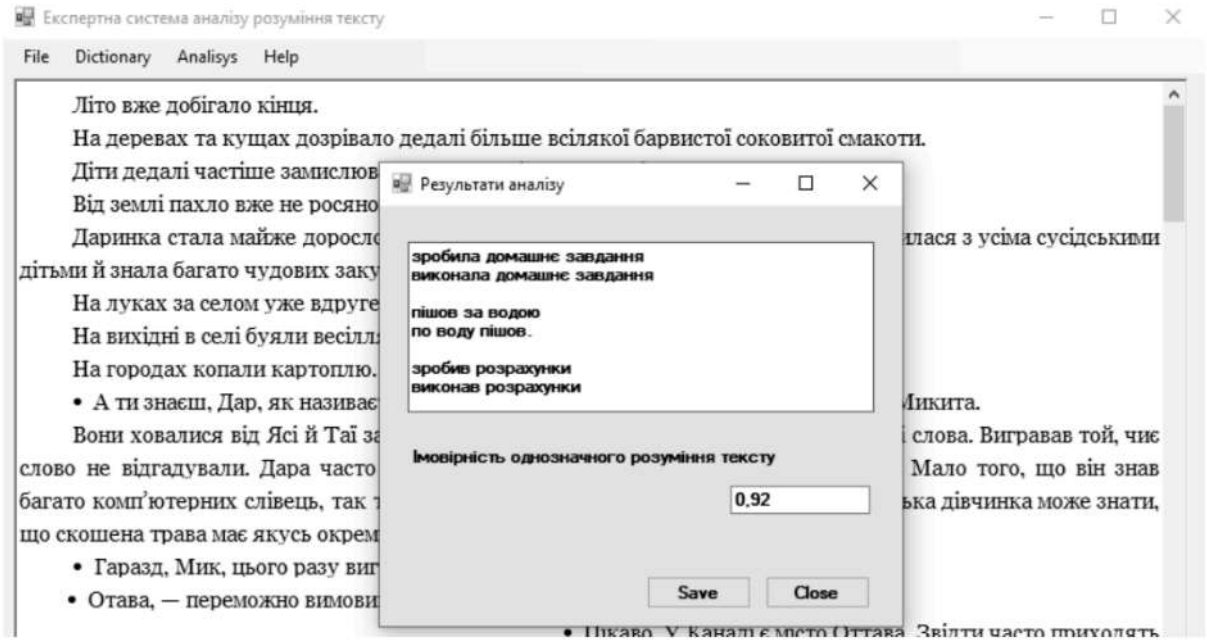

Fig. 3. The Results of Text Analysis Using an Expert System
Рис.3. Результати аналізу тексту з використанням експертної системи

The probability of unambiguous understanding of the text was 0.92. This means that in the minds of native speakers the sentences under consideration are not distinguished by difference.

Fig. 4 represents the results of studies of texts with different frequencies v coincidences of identical phrases are given.

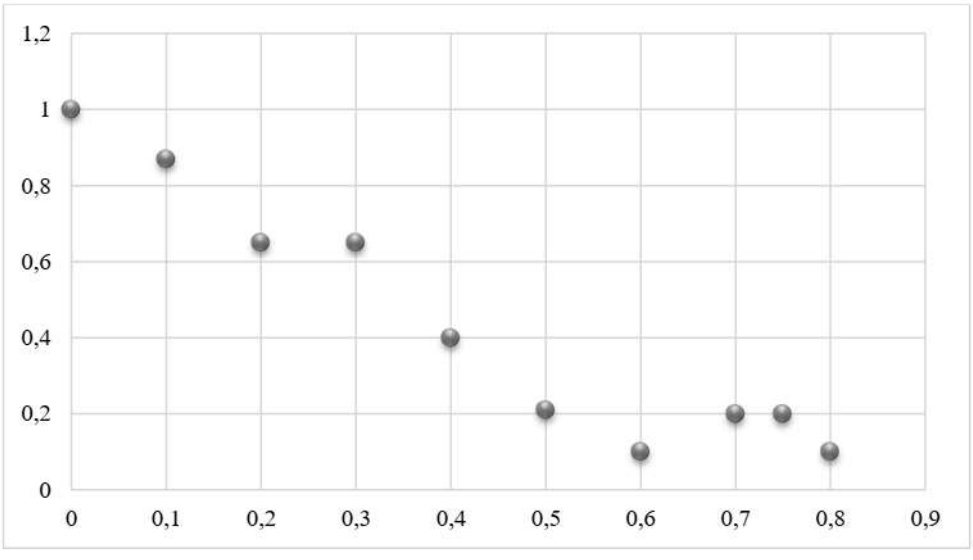

Fig. 4. f graphics
Рис.4. f графік

With sufficient accuracy, the probability of unambiguous understanding of the text P can be described by the function:

$$P = e^{-2,644 \cdot v}, R^2 = 0,8535 \quad (6)$$

The frequency of matching phrases from the dictionary is calculated as the ratio of the number of k matching phrases from the dictionary with different probabilities of understanding to the total number K of sentences in the text.

4. Conclusion

In linguistic research of qualitative characteristics of texts, nonparametric criteria of differences of language elements prove to provide more precise and rigorous results. In the context of smart integration of digital technologies in all spheres of science as well as a huge amount of information available, the process of linguistic analysis of the text, and in particular the analysis of the text for semantic unambiguity requires maximal automation.

The research developed a specialized expert system using the C # programming language for providing an automatic survey of native speakers and analyzing lexical meaning of phrases to be implemented as a desktop program for Windows. This takes into account the possibility of considering the determination of the correspondence of several phrases from dictionary files that require diverse expert assessments.

REFERENCES

1. B. Strickland, "Language Reflects 'Core' Cognition: A New Theory About the Origin of Cross-Linguistic Regularities," *Cognitive Science*, vol. 41, pp. 70–101, 2017. Available: https://onlinelibrary.wiley.com/doi/10.1111/cogs.12332
2. V. V. Zhukovska, "Cognitive-quantitative parametrization of positional properties of English detached impersonal/non-verbal constructions with explicit subject," *Zakarpatski Filolohichni Studii*, vol. 17, no. 2, pp. 121–128, 2021. [in Ukrainian] Available: https://dspace.uzhnu.edu.ua/items/54938f63-cea9-43bd-b849-4c925e87811f
3. V. M. Zayats and M. M. Zayats, "Methods of comparing statistical characteristics in the formation of samples in linguistics," *Scientific Bulletin Journal of Lviv Polytechnic National University "Information Systems and Networks"*, no. 673, pp. 296–305, 2010. Available: https://ena.lpnu.ua/handle/ntb/6753
4. Ye. V. Kupriianov, N. S. Uholnikova, and O. M. Yurchenko, *Structural linguistics in theory and practice*. Kharkiv, Ukraine: NTU "KhPI", 2024. [in Ukrainian]. Available: https://repository.kpi.kharkov.ua/handle/KhPI-Press/76037
5. W. Zhang, T. Yoshida, and X. Tang, "Text classification based on multi-word with support vector machine," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 879–886, 2008. Available: https://www.sciencedirect.com/science/article/abs/pii/S0950705108000968?via%3Dihub
6. A. Sitar, "Statistical analysis of phraseologized sentences: an indicator of mutual information association," *Ukrainian Linguistics*. Kyiv: Taras Shevchenko National University. vol. 1, no. 46, pp. 114–125, 2016. [in Ukrainian] Available: https://doi.org/10.17721/um/46(2016).103-125
7. W. Zhang, T. Yoshida, and X. Tang, "Text classification based on multi-word with support vector machine," *Knowledge-Based Systems*, vol. 21, no. 8, pp. 879–886, 2008. Available: https://www.researchgate.net/publication/222219356_Text_classification_based_on_multi-word_with_support_vector_machine#fullTextFileContent
8. V. Lytvyn, V. Vysotska, and M. H. Hrendus, "Method of data expression from the Ukrainian content based on the ontological approach," *Radio Electronics, Computer Science, Control*, no. 3, vol. 2362, pp. 53–70, 2018. Available: https://ric.zp.edu.ua/article/view/149991
9. J. Skeet, *C# in Depth*, 4th ed. Shelter Island, NY: Manning, 2019. Available: chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://dl.ebooksworld.ir/motoman/Manning.Csharp.in.Depth.4th.Edition.www.EBooksWorld.ir.pdf

Шабельник Тетяна Володимирівна *Доктор економічних наук, професор; завідувачка кафедри економічної кібернетики і системного аналізу, Харківський національний економічний університет імені Семена Кузнеця, просп. Науки, 9-А, Харків, 61166, Україна;*

Прокопович Світлана Валеріївна *Кандидат економічних наук, доцент; Доцент кафедри економічної кібернетики і системного аналізу, Харківський національний економічний університет імені Семена Кузнеця, просп. Науки, 9-А, Харків, 61166, Україна;*

Експертна система оцінювання закономірностей мови методами непараметричної статистики

Анотація. В роботі аргументовано, що при проведенні лінгвістичних досліджень мають місце завдання, вирішення яких потребує використання інструментарію непараметричних статистичних методів, зокрема проблеми семантичної відмінності речень.

Будь-яка мова, як система комунікації, характеризується поєднанням дискретних одиниць, особливою структурою та елементами, що пов'язаними між собою. Постійне збільшення їхньої кількості є результатом послідовної еволюції та розвитку мови. Використання статистичних методів у лінгвістиці, безумовно, підвищить ефективність лінгвістичного спостереження, його точність та надійність, а також узгодженість та доказовість лінгвістичних припущень.

Мету наукового дослідження визначено як, розробка експертної системи оцінки закономірностей мови методами непараметричної статистики для будь-якої мови.

Методи дослідження. В роботі використано методи непараметричної статистики, нотація IDEF4, мова програмування C#.

В роботі розроблена спеціалізована експертна система з використанням мови програмування C # для здійснення автоматичного опитування «носіїв» мови та аналізу лексичного значення фраз, яка реалізована у вигляді Desktop-програми для ОС Windows.

Для вибору мови програмування експертної системи сформульовані наступні вимоги: програмний продукт повинен бути реалізований для ОС Windows, забезпечувати якісний і доброзичливий інтерфейс, крім того немає необхідності в розподіленому додатку. Програмний продукт не надає можливість редагувати вихідний документ. У програму зчитується текст в форматах docx, rtf або txt. В алгоритмі передбачені можливості створення нового словника фраз із введенням їх експертних оцінок, перегляд і завантаження існуючих словників.

При роботі програми врахована можливість урахування визначення відповідності декількох фраз з файлів словників з різними експертними оцінками.

У лінгвістичних дослідженнях якісних характеристик текстів досить точні результати дають непараметричні критерії відмінностей мовних елементів. В умовах суцільної діджіталізації інформації та її величезних обсягів процес лінгвістичного аналізу тексту, зокрема аналізу тексту на смислову однозначність потребує максимальної автоматизації.

Ключові слова: *експертна система, методи непараметричної статистики, закономірності мови, програмне забезпечення.*

Надійшла у першій редакції 03.10.2025, в останній - 07.11.2025.
The first version has been received on 03.10.2025, the final version – on 07.11.2025.