UDC 811.111'33
DOI https://doi.org/10.32782/2409-1154.2025.74.1.6

Vlasiuk L. S.,
Senior Lecturer, PhD Student
Department of Theory, Practice and Translation of English
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
https://orcid.org/0000-0003-1020-0076

Demydenko O. P.,
PhD, Associate Professor
Department of Theory, Practice and Translation of English
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
https://orcid.org/0000-0002-0643-5510

AI-BASED TEXT ANALYSIS: STRUCTURAL AND SEMANTIC ASPECT

Summary. The article explores the potential and limitations of AI-based approaches to automatic text analysis, focusing on the structural and semantic dimensions of language processing. In the context of global digitalization, the exponential growth of textual data demands advanced analytical tools capable of ensuring efficiency, precision, and adaptability across multiple domains. Traditional linguistic methods, though valuable, are increasingly unable to keep pace with the dynamic information environment, which necessitates the adoption of artificial intelligence and natural language processing (NLP) technologies. The study reviews existing AI-driven systems, including LanguageTool, Grammarly, Turnitin, Linguakit, Stilus, Delph-in, SDU, and Link Grammar, emphasizing their ability to support tasks such as morphological, syntactic, and semantic analysis. While these systems provide valuable assistance in grammar checking, stylistic evaluation, and structural parsing, the research demonstrates that their accuracy remains limited, especially in addressing complex semantic phenomena such as idioms, metaphorical constructions, polysemy, and phraseological units.

A significant focus is given to the role of language resources in determining system effectiveness.

The findings highlight that effective AI-driven text analysis requires not only algorithmic sophistication but also comprehensive linguistic training resources. Building corpora and encoding structural, lexical, grammatical, and semantic patterns are prerequisites for enhancing the reliability of automatic systems. The research concludes that while current AI-based tools have achieved remarkable progress in automating routine linguistic tasks, they still fall short of fully replicating the complexity of human text comprehension. Improving their performance depends on resource enrichment, algorithmic refinement, and the integration of structural-semantic models. Ultimately, AI-based text analysis represents a transformative yet evolving field, with the potential to optimize information processing, support scientific and educational tasks, and contribute to the creation of a more structured and accessible digital information environment.

Key words: artificial intelligence, innovative technologies, English, foreign languages, information literacy, lexical unit, lexical-semantic fields, digital environment, information reliability.

Problem Statement. The modern era is characterized by the rapid development of AI-based technologies in different areas. With the information ecosystem overflowing with the excessive amount of information, linguistics is also in need of advanced tools which would assist in the process of the text analysis as efficiently as possible. This necessity is particularly acute due to the high level of the destructurization of the digital information ecosystem caused by the uncontrolled amount of texts and other data.

However, the existing systems do not work well for a profound language analysis, particularly in terms of the language structure and semantics. Automatic text analysis is an important area of development in artificial intelligence language technologies, focused on improving the efficiency of processing large volumes of information. In the context of global digitalization, the volume of text data requiring systematization, analysis, and concise presentation is growing rapidly.

Traditional approaches to such analysis require significant time and human resources, which limits their applicability in the context of highly dynamic information processes. Artificial intelligence, in particular natural language processing methods, offers tools for automating these processes, ensuring accuracy, speed and adaptability to different languages and topics.

Automatic text analysis is an important tool for solving current scientific and practical problems. Scientific aspects relate to the development of algorithms capable of taking into account contextual semantics, multilingualism and the specifics of texts from different domains. Practical tasks focus on integrating these algorithms into information systems used in scientific, educational, legal and other fields. Such technologies contribute to improving the efficiency of data analysis, automating routine tasks, and expanding the possibilities for working with large amounts of text information.

Theoretical background. Over the last few years the issue of automatic text analysis has been in the center of the research works of both national and foreign researchers. In particular, the underlying concepts of automatic text analysis we can find in the works of Robert Moore, Kevin Oliver, Scott Crossley, Natalia Dyachuk, Yulia Batko and others. However, with the some of the core principles of automatic text analysis having been out-

lined in these works, there is still a number of issues that require further analysis.

The **aim** of the article is to highlight the potential of AI-based approaches to text analysis, with a particular focus on structural and semantic dimensions. It seeks to examine how artificial intelligence techniques can identify, represent, and interpret linguistic structures and semantic relations within texts, highlighting their effectiveness, limitations, and implications for linguistic research and practical applications.

Results and discussion. At this stage there is a great a range of AI-based systems that allow to perform an automatic text analysis. These systems differ in the context of their functions and their capacities to work with different languages. Below we're listing some of the most widely used programs while carrying out an automatic text analysis.

LanguageTool (https://languagetool.org) is a software tool for checking grammar and writing style. It supports many languages and can detect grammatical errors, incorrect word usage, stylistic flaws, etc.

Grammarly (https://www.grammarly.com) – a program for automatic checking of English grammar and spelling. It can be used as a browser extension or as a standalone program.

Turnitin (https://www.turnitin.com) – with this program, you can get an idea of which part of the submission is authentic, written by a human, and which is generated by artificial intelligence using ChatGPT or other tools.

Linguakit (https://linguakit.com/en/syntactic-analyzer) – a software tool for processing and analyzing text data, including counting words, phrases, sentence length, lexical analysis, etc. This can be useful for solving various tasks related to text analysis.

Stilus (https://www.mystilus.com)– an online tool for morphological and syntactic analysis of text. Using this tool makes it possible to automatically parse texts into individual words and determine their parts of speech, forms, and dependencies.

Delph-in (Deep Linguistic Processing with HPSG Initiative) (https://delphin.github.io/delphin-viz/demo/) – a software tool for visualizing and analyzing the results of deep syntactic analysis of text (parsing). The program supports dynamic mode, syntactic tree visualization, linguistic annotations (parts of speech, dependencies, semantic roles, etc.) and highlighting of linguistic features.

SDU (https://visl.sdu.dk/visl/en/parsing/automatic/parse.php) – an online tool for automatic syntactic analysis of text (parsing). Developed based on grammar and linguistic resources at the University of Denmark. The software is available online and does not require installation, but there is a certain limitation on the length of the text entered.

Link Grammar – The official website of the Link Grammar program can be found at the following link: http://www.link.cs.cmu.edu/link/.

Automated text analysis tools are extremely helpful in analyzing patterns in text, identifying relevant words and phrases, and minimizing the search for irrelevant studies. However, our research has shown that such systems have significant shortcomings and often lack a high level of effectiveness.

When performing automatic text analysis and linguistic indexing, it is necessary to take into account a number of factors that govern any system. These factors include the semantic, structural, and syntactic features of a particular language. In addition, it is important to consider how well these features are known to the system [1, p. 120]. This depends on the type of language: high-resource or low-resource. In applied linguistics, particularly computer linguistics, these terms are used to refer to the amount of linguistic data and technological support available for a given language.

High-resource languages are languages that have large digital and linguistic resources. Their key features include large corpora (text, speech, parallel translations), highly developed linguistic analysis tools (tokenizers, parsers, taggers), a high level of presence in Internet sources and academic research, and broad support from natural language processing programs (online translators, speech recognition systems, etc.).

Low-resource languages are characterized by limited or insufficient digital and technological support. These languages can be characterized by the presence of small corpora or their absence (as a rule, there are few text datasets, and speech data is very limited), a lack of annotated resources (no large dictionaries or corpora with POS tags), and a small number of natural language processing tools (there are no spell checkers or parsers for such languages, and translation programs do not provide reliable results) [2, p. 342].

There is also one more category of languages which can be referred to as a middle-resource language. Those are the languages that can be branded neither as high-resource language nor as low-resource language. That is there are certain systems that can work with these languages to a certain extent, but the results are not as exact as they should be.

Whether a language is classified as high-resource, medium-resource, or low-resource plays a key role in the subsequent automated linguistic analysis. This is because the effectiveness of any language in an automatic linguistic indexing system depends on a number of factors [3, p. 54]. These factors include the availability of corpora, the ability to work with linguistic tools, representation in academic research and the digital space, as well as the ability to support natural language processing systems.

When we talk about automatic language analysis systems (such as machine translation, speech recognition, or text mining), their effectiveness depends heavily on the level of language resources. This is because any linguistic analysis system works on the basis of pre-programmed algorithms and uses available materials (including semantic, structural, syntactic, and other linguistic features) during the analysis [3, p. 64]. Therefore, when working with highly resourced languages, the effectiveness of the analysis will be higher than when working with low-resourced systems. This is due to the fact that there are significantly more resources available in the information space, on the basis of which it is possible to track patterns characteristic of a particular language and, accordingly, perform an analysis with higher accuracy.

Improving the productivity and efficiency of any automated system depends on how broad and comprehensive the material for further work is. In other words, the system must first be "trained," that is, provided with sufficient resources for analysis and key language patterns, including its semantic and structural features.

The carried out analysis shows that even in the case of English, which is a highly resourceful language, none of the automatic text analysis systems are 100% accurate; the average effectiveness of these systems is 70-85%. The greatest difficulties arise when analyzing lexical and semantic components. None of the programs has sufficient knowledge and resources to work effectively with idioms, phraseological units, and phrasal verbs. The greatest difficulty is posed by metaphorical constructions, as well as working with

homonyms, homographs, and homophones. In addition, systems often demonstrate a low ability to work with complex grammatical constructions. In turn, the inability to work with such text elements indicates the need for further improvement of these systems.

The results of these systems when working with Ukrainian-language sources are significantly lower. More than 80% of systems do not have any tools for working with the Ukrainian language. Those systems that do have such tools demonstrate extremely inaccurate results: their average effectiveness is 40%. The programs do not have enough material to effectively analyze grammatical, semantic, and lexical aspects.

Improving the effectiveness of automatic linguistic indexing systems is possible, first and foremost, by providing sufficient resources to set basic grammatical, lexical, semantic, and syntactic patterns for the system. The automatic text analysis is a highly complex process which requires the development of the relevant algorithm.

The first step is to create a corpus that allows us to collect a sufficient amount of resource data to subsequently provide the automatic text analysis system with material. The next step is to conduct a linguistic analysis of the text in order to form typical grammatical, lexical, semantic, and syntactic patterns, on the basis of which automatic indexing programs will be able to perform further analysis and provide more accurate results.

Text processing remains one of the key tasks. Our knowledge of reality is expressed in verbal form. Teaching automated systems to "understand" and "analyze" text means ensuring their ability to obtain the information necessary to perform various tasks. Such "understanding" and "analysis" of text includes the ability to interpret it at various levels of information representation, such as morphological, syntactic, logical-semantic, as well as to summarize the results of the analysis in a specific, predefined form.

Today, automatic text processing (ATP) systems, also known as automated text processing (ATPS) systems, occupy a prominent place among linguistic intelligent computer systems. Such systems simulate human mental activity in the process of solving theoretical and/or practical problems. The main task of ATP and/or ATPS systems is to analyze text at various levels, such as morphological, syntactic, and logical-semantic, as well as to identify text components using appropriate computer grammar modules [4, p. 81].

The strategy for creating computer analysis systems for text information involves the use of two main technologies. The first technology, known as dictionary technology, is based on the development of auxiliary linguistic databases, such as dictionaries, compilation of rules, changes in word forms, verification and/or identification of these word forms for the purpose of practical implementation of the created information processing algorithms. The other technology is dictionary-free, "independent" and is aimed at using algorithmic rules to represent the necessary information about linguistic units.

These two technologies are not mutually exclusive. The choice of the leading approach in the process of developing a specific APT and/or ATPS system depends on a number of factors, namely the type of text, the type of task, technical capabilities and characteristics of the available software. The most effective is the use of systems that combine the advantages of both technologies. This is explained by the fact that the complete rejection of auxiliary databases can result in a complicated structure of algorithms [4, p. 85]. Accordingly, the use of such databases can lead to an increase in the level of complexity of these algorithms.

The initial module of APT and/or ATPS systems is an automatic morphological text analysis module designed to automatically determine the grammatical class of each word in the text, as well as its grammatical subclass. A grammatical class determines the part-of-language belonging of a word, and grammatical subclasses are categories of words that have common substantive, formal, and functional properties. Typically, these are words that can be attributed to different grammatical categories within different parts of languages.

At the phrase level, automatic syntactic analysis (ASA) aims to automatically select phrase combinations, assign syntactic connections to these phrase combinations, and automatically create corresponding phrase dictionaries. At the sentence level, it is assumed to create a complete syntactic analysis – a dependency tree. The main goal of syntactic analysis is to identify connections between sentence members, establish semantic meaning and sentence segmentation. In turn, ASA pursues similar goals, using computer syntax, the main task of which is to determine syntactic structures in the text and their corresponding representation. In fact, the text is decomposed into minimal syntagms – words that are interconnected by means of a syntactic connection.

In order for automatic syntactic analysis to be performed correctly and effectively, it is first necessary to perform pre-processing of the source information. Such processing requires not only the isolation of semantic elements in the text and their marking, but also the analysis of a number of linguistic phenomena.

The multitasking of automatic analysis allows us to eliminate both morphological and syntactic ambiguity by using information from the semantic and syntactic levels. Thus, we obtain more accurate results, since the system analyzes a lexical unit taking into account all possible options.

Conclusions. The process of automatic text analysis is complex and requires comprehensive approaches. Its effectiveness directly depends on the level of "training" of the automatic linguistic text analysis system and the volume of the resource database. It is important to take into account a number of features: structural-semantic, lexical, grammatical and syntactic features, text type, general concept of the text, linguo-cultural and ethno-cultural features, politically correct expressions, features of terms, metaphorical constructions. The more data is contained in the database, the more effective will be the creation of a logical-linguistic model, and therefore the process of transforming natural language into information-search language. This, in turn, allows automatic analysis systems to read information more accurately, thereby increasing the level of quality of linguistic indexing. Thus, it becomes possible to provide a structured digital information space, and therefore more clear and accurate search results.

Bibliography:

1. Corazza E. Reflecting the Mind: Indexicality and Quasi-Indexicality. Oxford : Oxford University Press, 2004. 400 p.
2. Crossley S. A. Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal*. 2018. Vol. 102, No 2. P. 333-349.
3. Giorgi Alessandra. About the Speaker: Towards a Syntax of Indexicality. New York : Oxford University Press, 2010. 320 p.
4. Steinbach M. A. Comparison of Document Clustering Techniques. Minnesota : Minnesota Publishing, 2011. 247 p.

Власюк Л., Демиденко О. Аналіз тексту за допомогою штучного інтелекту: структурно-семантичний аспект

Анотація. У статті досліджується потенціал та обмеження підходів на основі штучного інтелекту до автоматичного аналізу тексту, зосереджуючись на структурних та семантичних вимірах обробки мови. У контексті глобальної цифровізації експоненціальне зростання текстових даних вимагає передових аналітичних інструментів, здатних забезпечити ефективність, точність та адаптивність у багатьох областях. Традиційні лінгвістичні методи, попри значний їхній потенціал, все частіше не встигають за динамічним інформаційним середовищем, що вимагає впровадження технологій штучного інтелекту та обробки природної мови (NLP). У статті розглядаються існуючі системи на основі штучного інтелекту, включаючи LanguageTool, Grammarly, Turnitin, Linguakit, Stilus, Delph-in, SDU та Link Grammar, з акцентом на їхній здатності підтримувати такі завдання, як морфологічний, синтаксичний та семантичний аналіз. Хоча ці системи надають значну допомогу в перевірці граматики, стилістичній оцінці та структурному розборі, дослідження демонструє, що їхня точність залишається обмеженою, особливо при розгляді складних семантичних явищ, таких як ідіоми, метафоричні конструкції, полісемія та фразеологічні одиниці.

Значна увага приділяється ролі мовних ресурсів у визначенні ефективності системи.

Результати дослідження підкреслюють, що ефективний аналіз тексту на основі штучного інтелекту вимагає не лише алгоритмічної складності, але й комплексних лінгвістичних навчальних ресурсів. Створення корпусів та кодування структурних, лексичних, граматичних та семантичних шаблонів закладають основу для підвищення надійності автоматичних систем. Попри те, що сучасні інструменти на основі штучного інтелекту досягли значного прогресу в автоматизації рутинних лінгвістичних завдань, вони все ще не здатні повністю відтворити складність розуміння тексту людиною. Покращення їхньої продуктивності залежить від збагачення ресурсів, алгоритмічного вдосконалення та інтеграції структурно-семантичних моделей. Зрештою, аналіз тексту на основі штучного інтелекту являє собою трансформаційну галузь, яка має потенціал для оптимізації обробки інформації, підтримки наукових та освітніх завдань, а також для створення більш структурованого та доступного цифрового інформаційного середовища.

Ключові слова: штучний інтелект, інноваційні технології, англійська мова, іноземні мови, інформаційна грамотність, лексична одиниця, лексико-семантичні поля, цифрове середовище, надійність інформації.

Дата першого надходження рукопису до видання: 12.08.2025
Дата прийнятого до друку рукопису після рецензування: 11.09.2025
Дата публікації: 21.10.2025