ПРИКЛАДНА ЛІНГВІСТИКА

UDC 811.111’33
DOI https://doi.org/10.52726/as.humanities/2025.2.12

L. S. VLASIUK
Lecturer, Postgraduate Student,
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine
E-mail: l.vlasiuk@kpi.ua
https://orcid.org/0000-0003-1020-0076

O. P. DEMYDENKO
PhD, Associate Professor,
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine
E-mail: olga.demydenko80@gmail.com
https://orcid.org/0000-0002-0643-5510

STRUCTURING INFORMATION ECOSYSTEM VIA LINGUISTIC INDEXATION

The article deals with the role of linguistic indexation in arranging the modern information ecosystem in the era of digital content and the ever-expanding use of media texts. Shift from manual to automated information retrieval highlighted the necessity of intelligent search technologies capable of efficiently processing large volumes of unstructured textual data. This paper outlines both the theoretical and practical ground of linguistic indexation, namely term indexing, metonymic indexing and syntactic indexing.

The study also views linguistic indexation within broader developments in computational linguistics, data mining and artificial intelligence, showing how these tools improve search outcomes by enabling semantic understanding and contextual awareness. A detailed analysis is presented on how indexing systems generate search images through keywords and descriptors, facilitating both clarity and reliability in content discovery.

Despite notable advances in the digitalized world, the authors stress that existing intelligent search systems often fall short in working with the specialized or domain-specific content, where nuanced linguistic features may not be subjected to generic indexing mechanisms. Consequently, the article strives for ongoing innovation in algorithmic development, indexing strategies and software tools tailored to specific sectors.

The article concludes that linguistic indexation serves as a foundational mechanism for structuring the information ecosystem, offering substantial improvements in speed, accuracy and user relevance of digital search environments. In light of the continual expansion of digital text, the authors underline the need for sustained interdisciplinary research to optimize linguistic processing tools and ensure they remain adaptive to the evolving demands of information retrieval and content analysis.

Key words: syntactic features, grammatical features, digital space, digital public discourse, human editing, text analysis, information systems.

Problem statement. A key characteristic of modern linguistics was the emergence of large volumes of documents and publications that needed to be sorted and unified. It was during this period that the first information retrieval systems were developed. At the first stages, such search was carried out manually, but the rapid development of the computer industry and, accordingly, the subsequent automation of business processes made a significant contribution to the implementation of the process of digitizing the text format of information, and subsequently to the development of automatic information retrieval systems.

Today, one of the most pressing issues for modern linguistics is the problem of structuring the information ecosystem of texts. That is why automatic text analysis and synthesis, text clustering, linguistic databases and their automation, and improvement of information retrieval systems are among the most important areas of linguistic research.

Particular attention should be paid to Internet resources, as today they are undoubtedly considered

one of the most used sources of information search, gradually replacing the paper format. This is also facilitated by the comprehensive digitalization of all spheres of human activity. This issue is extremely relevant for media texts, given the specifics of their use, which is increasingly expanding its horizons in the modern world. Gradually becoming a component of any field (journalistic, scientific, etc.), the issue of structuring media texts in such a way that they also reflect accurate search results is becoming more acute. This is because solving this issue would be a significant contribution to the unification and structuring of the information ecosystem of texts, which is an extremely relevant and acute issue today. In our research, we will focus on further studying the issue of indexing, in particular media texts, given its importance for building a single information space.

Research analysis. The underlying principles of linguistic indexation were studied by both national and foreign researchers. Particularly, the basic principles of applied linguistics were analyzed by John Catford, Charles Ferguson, Evgenia Karpilovska. The notion of information and search languages were studied by Daniel Batten, Moshe Taube, Greg Myers. The underlying principles of linguistic indexing were analyzed by Tony McEnery, David Crystal, Nikolas Coupland, Timo Lahtinen and others.

The aim of the article. The article is aimed at defining methods of linguistic indexation fostering the structuring of the information ecosystem.

Results and discussion. In order to search for the necessary information through information retrieval systems, it is necessary to create a search image of a document, which is usually an unstructured set of keywords (words that are presented in a standard lexicographic form and reflect key information about the document).

In this context, information retrieval is undoubtedly one of the most important processes. Therefore, the creation and high-quality functioning of an information retrieval system is directly related to internal information and cognitive processes and information processing software, which indicates the need to use a large number of computer linguistics tools. An effective search for up-to-date and relevant information is made possible by artificial languages and information retrieval languages, and therefore the importance of analyzing modern information retrieval languages and their types in order to optimize the search is undeniable.

Indexing, in particular, coordinate indexing, can be defined as a basic means of linguistic indexation and document retrieval in the online environment. It is an effective tool that enables accurate and complete display of text content through the use of information retrieval language, keywords, and descriptors [Corazza 2004 : 47]. In the context of multidimensionality, volume, intersection of terms and relations between them, and intersection of semantic loads, indexing creates a unique search image of the text that allows users to immediately understand how the results of their search query meet their requirements, whether they are relevant, reliable and up-to-date.

Modern linguistic indexation relies on data mining for a number of reasons: 1) the lack of accuracy, completeness, and clarity, accompanied by the contradictory nature of the data and a large amount of information; 2) the availability of artificial intelligence tools in text analysis algorithms; 3) the need to use considerable computing resources and a comprehensive automated approach to thoroughly process large volumes of media texts.

Thus, it became necessary to create so-called intelligent search systems, in other words, technologies for in-depth text analysis.

We can define the technology of in-depth text analysis as a tool for analyzing large amounts of information aimed at finding relationships to select relevant information. This technology is a relatively new method of information retrieval [Steinbach 2011 : 54–55]. Previously, the user received a list of documents, books, texts, etc. in the search results that could potentially contain the required search results, which greatly complicated the process of selecting the necessary data. Now, thanks to further improvements in text mining technologies, the information search process is gradually becoming easier, as modern technologies are aimed at structuring the information ecosystem, which should lead to the most accurate results possible.

The first step in structuring the information ecosystem is to improve search engines in corporate networks, whose main function is to process arrays of text documents from individual

institutions. The size of such networks ranges from several gigabytes to several tens of gigabytes. In practice, such programs can be implemented in a networked version [Steinbach 2011 : 60]. In this case, the database will be available exclusively on a local network server. Corporate networks can be filled with the following resources: websites, individual Internet pages, publicly available databases, and various repositories of both structured and unstructured data. The use of these resources in combination leads to the fact that search engines are not able to transfer the entire amount of data to the corporate environment, and other methods must be used to accomplish such ambitious tasks.

The next method includes search engines that operate on the Internet protocol. Compared to the previous one, this option is more advanced and optimal, since both the main program and the database are located on the central server of the local network [Sukhyi 2005 : 20]. This means that access to the database is provided to absolutely all users of the intranet through a regular Internet browser. Information is searched for similarly to the global Internet. Having access to the Intranet, the user enters the addresses of the databases, gains access to them, and then searches for the necessary information according to the standard scheme. The interface of such a search engine is also quite standard. You should also understand that in this case we are talking about a multi-user mode, so search programs of this kind use Internet protocols for their work.

The last and, at the same time, the most advanced method is software systems with a client-server architecture, which is actually the client part of the program.

Any of the above methods serves as a basis for further development of intellectual analysis tools, including technologies for extracting factual information about the search object, taking into account anaphoric references to it (i.e. references to the object directly mentioned in the text), fuzzy search, thematic (precise) and tonal (complete) rubrication, cluster analysis of a selection of documents/texts/stores/databases, etc., building annotations and multidimensional frequency distributions of documents, and research of these documents based on OLAP technologies [Lobanovska 2011 : 32]. In addition, text mining methods can serve as a basis for determining the direction of research of large arrays of documents/texts and for extracting new, relevant information from these documents/texts.

The most relevant ways of extracting information from texts today include [Giorgi 2010 : 64–66]:

– analytical processing of facts;

– obtaining and structuring factual information;

– use of thesauri to search for information;

– searching for information in individual document repositories, text collections, databases, etc;

– annotating and abstracting documents, building digests based on objects;

– clustering and rubrication of documents (performing their thematic analysis);

– performing dynamic analysis of the structure of texts based on their semantics;

– identification of the main (key) topics and relevant information objects;

– determining the general and specific tone of information;

– conducting a study of the frequency characteristics of texts.

Despite their undoubted advantage in the context of simplifying the search for relevant information on the global network, intelligent systems are currently unable to prove their effectiveness in searching specialized systems and, in particular, in analyzing information. Based on this, we can state that improving existing methods and developing new ones, as well as software products whose functionality would allow for automatic search and analysis of information in accordance with the parameters of a particular industry, is more relevant than ever.

The main problem of search engines today is not only finding relevant information, but also obtaining highly accurate results in a short period of time. Search engines cover a huge array of documents, so achieving accurate results requires performing a series of operations, which leads to a decrease in speed; in turn, maintaining speed causes a decrease in the quality of the results obtained. The only way to maintain both speed and quality is linguistic indexation.

Linguistic indexation of texts involves the classification of unstructured text arrays in the media sphere, which can be achieved using different approaches. Typically, text structures

include term indexing, metonymic indexing, and syntactic indexing.

First, let's look at term indexing, which is most often accomplished by using inverted indexes. There are two ways to invert an index: the first method involves placing the index in RAM; the second method involves using a disk where the index size is in the range of 50%–150%. If you do not take into account the free space for further updating the index and large amounts of information, the index size can be 20% of the total size of the texts [Lobanovska 2011 : 35].

Another important component of the indexation process is metonymic indexing. This is primarily due to the fact that the effectiveness of indexing only terms is justified when working with scientific sources, where the term is actually the basis of relevant indexation. However, when working with other types of sources, this approach alone will not be very effective. Understanding a text can often depend on the world model of the participants in the communication act and on the interest factor. It is important to consider the text not only as an immanent structure, the meaning of which can be interpreted in different ways, but also as a metastructure that allows you to identify cultural codes and implicit meanings in it and recognize relevant keywords accordingly. When decoding the meaning of a metonymically organized text, you need to rely on the internal context, which includes the speaker's past experience programmed into his or her mind.

The next extremely important aspect is the syntactic indexation of texts, which can be carried out according to the following principles:

1. Leonardo Bloomfield's analysis, i. e. analysis based on direct components.

2. Lucien Tenier's method, which involves building a dependency tree.

Both of these approaches to the choice of principles are important because they are presented within the framework of structuralism and correspond to a purely formal approach to sentence analysis, which makes them the most suitable for automatic analysis.

The next step is to automatically identify syntactic (semantic-syntactic) relations within the identified components. These relations can be displayed within the constructed dependency trees. The corresponding automated process will be based on the construction of a logical and linguistic model of a natural language sentence that reproduces the syntactic structure of the sentence, taking into account the semantic relationship that makes the meaning of the text clear [Corazza 2004 : 70]. For a more accurate study of natural language sentences and, accordingly, linguistic indexation, it is extremely important to compare logical and linguistic models of textual information, which takes place in several stages:

– creating logical and linguistic models, which involves writing a logical formula for each natural language sentence. The order of elements in the logical formula corresponds to the syntactic role of individual words in the sentence;

– identification, where we analyze predicates, predicate variables and constants, search for synonyms, antonyms, cognates, passive forms, etc. This allows us to track how different elements with the same syntactic role reproduce the meaning of a sentence;

– substitution of identical variables, which is a necessary step for unifying the meaning;

– logical inference, which is the stage at which the comparison of logical and linguistic models is actually performed. In fact, we establish the relationship between words and content components using component analysis.

– Component analysis allows us to find synonyms and hyperonyms, analyze units that are at the same level of the hierarchy, create a list of components that distinguish a particular element from all others, and formulate a word definition.

Conclusions. Structuring the information ecosystem is a complex task which encompasses many stages of text analysis and, consequently, requires an alignment of the processes in automatic systems for language analysis. The key component in the process of information ecosystem structuring is linguistic indexation, which requires a comprehensive and multifaceted approach. Thus, we need to carry out an analysis of smart search systems, automatic text analysis programs and indexing mechanisms. This helps to ensure an efficient, accurate and relevant process of information retrieval.

Since the amount of digital content continues to grow, the issue of linguistic indexation leaves ground for future research. This will contribute towards the replenishment of information analysis programs to make them more capable of understanding and interpreting the needed information.

BIBLIOGRAPHY

1. Corazza E. (2004). Reflecting the Mind: Indexicality and Quasi-Indexicality. Oxford : Oxford University Press.
2. Giorgi Alessandra. (2010). About the Speaker: Towards a Syntax of Indexicality. New York : Oxford University Press.
3. Steinbach M. A. (2011). Comparison of Document Clustering Techniques. Minnesota : Minnesota Publishing.
4. Ticher S & Mejer M. (2009). Methods for analyzing text and discourse. Oxford : Oxford University Press.
5. Лобановська І. Г. (2011). Індексування документів ключовими словами. Київ : Нілан-ЛТД.
6. Сухий О. Л., Міленін В. М., Тарадайнік В. М. (2005). Алгоритми пошуку в інформаційних системах. Київ.

REFERENCES

1. Corazza E. (2004). Reflecting the Mind: Indexicality and Quasi-Indexicality. Oxford : Oxford University Press.
2. Giorgi Alessandra. (2010). About the Speaker: Towards a Syntax of Indexicality. New York : Oxford University Press.
3. Steinbach M. A. (2011). Comparison of Document Clustering Techniques. Minnesota : Minnesota Publishing.
4. Ticher S & Mejer M. (2009). Methods for analyzing text and discourse. Oxford : Oxford University Press.
5. Lobanovska I.H. (2011). Indeksuvannia dokumentiv kliuchovymy slovamy. Kyiv : Nilan-LTD.
6. Sukhyi O. L., Milenin V. M., Taradainik V. M. (2005). Alhorytmy poshuku v informatsiinykh systemakh. Kyiv.

Л. С. ВЛАСЮК
викладач, аспірантка,
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна
Електронна адреса: l.vlasiuk@kpi.ua
https://orcid.org/0000-0003-1020-0076

О. П. ДЕМИДЕНКО
кандидат педагогічний наук, доцент,
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», м. Київ, Україна
Електронна адреса: olga.demydenko80@gmail.com
https://orcid.org/0000-0002-0643-5510

СТРУКТУРУВАННЯ ІНФОРМАЦІЙНОЇ ЕКОСИСТЕМИ ШЛЯХОМ ЛІНГВІСТИЧНОЇ ІНДЕКСАЦІЇ

У статті розглядається роль лінгвістичної індексації в організації сучасної інформаційної екосистеми в епоху цифрового контенту та дедалі ширшого використання медіатекстів. Перехід від ручного до автоматизованого пошуку інформації актуалізував потребу в інтелектуальних пошукових технологіях, здатних ефективно опрацьовувати великі обсяги неструктурованих текстових даних. У цій статті викладено теоретичні та практичні засади лінгвістичного індексування, а саме термінологічного, метонімічного та синтаксичного індексування.

Дослідження також розглядає лінгвістичну індексацію в контексті ширшого розвитку комп'ютерної лінгвістики, інтелектуального аналізу даних і штучного інтелекту, показуючи, як ці інструменти покращують результати пошуку завдяки семантичному розумінню і контекстуальній обізнаності. Представлено детальний аналіз того, як системи індексування генерують пошукові образи за допомогою ключових слів і дескрипторів, сприяючи чіткості та надійності у виявленні контенту.

Незважаючи на значний прогрес в оцифрованому світі, автори підкреслюють, що існуючі інтелектуальні пошукові системи часто не справляються зі спеціалізованим або доменним контентом, де нюанси лінгвістичних особливостей можуть не піддаватися загальним механізмам індексації. Таким чином, у статті наголошується на необхідності постійних інновацій у розробці алгоритмів, стратегій індексування та програмних інструментів, пристосованих до конкретних галузей.

У статті зроблено висновок, що лінгвістична індексація слугує фундаментальним механізмом для структурування інформаційної екосистеми, пропонуючи суттєве покращення швидкості, точності та релевантності для користувачів цифрових пошукових середовищ. У світлі постійного розширення цифрового тексту автори підкреслюють необхідність постійних міждисциплінарних досліджень для оптимізації інструментів лінгвістичної обробки та забезпечення їхньої адаптивності до мінливих вимог пошуку інформації та контент-аналізу.

Ключові слова: синтаксичні особливості, граматичні особливості, цифровий простір, цифровий дискурс, людське редагування, аналіз тексту, інформаційні системи.