**UDC 81.26**
**DOI**

***Olga DEMYDENKO,***
*orcid.org/0000-0002-0643-5510*
*PhD in Pedagogy,*
*Associate Professor at the Department of Theory, Practice and Translation of English*
*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"*
*(Kyiv, Ukraine) olga.demydenko80@gmail.com*

***Liudmyla VLASIUK,***
*orcid.org/0000-0003-1020-0076*
*Lecturer at the Department of English for Engineering 1,*
*PhD student at the Department of Theory, Practice and Translation of English*
*National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"*
*(Kyiv, Ukraine) LyudmylaVlasyuk@ukr.net*

# DESCRIPTIVE LANGUAGE AS THE TYPES OF LINGUISTIC INDEXATION

*The article deals with the problem of conventional and advanced information search based on the phenomenon of linguistic indexation, The article pays special attention to information search languages and their diverse types as those are key aspects influencing fast and relevant information search. Moreover, the article reveals and describes the most widely used search intelligent language systems and means of their implementation. Apart from it, the advantages and disadvantages of building a system of each information retrieval language are presented.*

*One of the leading characteristics of the end of the first half of the twentieth century is the rapid emergence of large volumes of documents and publications that needed to be sorted out and properly managed. During this period, the first information search systems appeared. However, those systems were not advanced at that time and, thus, were initially performed manually. Over the years the successful development of computer technology influenced the introduction of the process of digitizing of textual information and the creation of automatic information search systems whose main goal is to find documents in the information array that meet the query criteria. This procedure is subject to certain algorithms, where the following main steps are performed: receiving a task, selecting documents, checking the completeness and accuracy of the search.*

*However, the high-quality performance of the aforementioned tasks is impossible without the proper study of linguistic indexation phenomenon. Linguistic indexation is characterized by an enormous variety of types, among which the most widely used is descriptive language – an information language, whose units are descriptors. Descriptor can be defined as: 1) one of the keywords that identifies a document in information search systems; 2) a unit of language of an information search system that corresponds to a certain basic concept included in the thesaurus of this system.*

***Key words:*** *linguistic indexation, keyword search, descriptive language, types of indexing, information search, information search language, structuring and standardizing texts' ecosystem, indexing system.*

***Ольга ДЕМИДЕНКО,***
*orcid.org/0000-0002-0643-5510*
*кандидат педагогічних наук,*
*доцентка кафедри теорії практики та перекладу англійської мови*
*Національного технічного університету України*
*«Київського політехнічного інституту імені Ігоря Сікорського»*
*(Київ, Україна) olga.demydenko80@gmail.com*

***Людмила ВЛАСЮК,***
*orcid.org/0000-0003-1020-0076*
*викладачка кафедри англійської мови технічного спрямування № 1,*
*аспірантка кафедри теорії, практики та перекладу англійської мови*
*Національного технічного університету України*
*«Київського політехнічного інституту імені Ігоря Сікорського»*
*(Київ, Україна) LyudmylaVlasyuk@ukr.net*

# ДЕСКРИПТОРНА МОВА ЯК ОДИН ІЗ ВИДІВ ЛІНГВІСТИЧНОЇ ІНДЕКСАЦІЇ

*У статті розглядається проблема звичайного та розширеного пошуку інформації на основі явища лінгвістичної індексації. Особлива увага у статті приділяється мовам пошуку інформації та їх різноманітним видам, оскільки вони належать до ключових аспектів, які впливають на швидкий та актуальний пошук інформації. Крім того, у статті розкриваються та описуються найбільш поширені пошукові мовні системи та засоби їх реалізації. Крім того, представлено переваги та недоліки побудови системи кожної інформаційно-пошукової мови.*

*Однією з провідних характеристик кінця першої половини двадцятого століття є швидка поява великих обсягів документів і публікацій, які потребують сортування та належного управління. На цей період припала поява перших інформаційно-пошукових систем. Однак, на той час ці системи не були досконалими, а тому спочатку пошук здійснювався вручну. Згодом, успішний розвиток комп'ютерних технологій вплинув на впровадження процесу оцифрування текстової інформації та створення систем автоматичного пошуку інформації, основною метою яких є пошук в інформаційному масиві документів, які відповідають критеріям запиту. Ця процедура підпорядковується певним алгоритмам, де виконуються такі основні етапи: отримання завдання, відбір документів, перевірка повноти та правильності пошуку.*

*Проте якісне виконання вищезазначених завдань неможливе без належного вивчення феномену лінгвістичної індексації. Лінгвістична індексація характеризується величезною різноманітністю типів, серед яких найпоширенішим є дескрипторна мова – інформаційна мова, одиницями якої є дескриптори. Дескриптор можна визначити як: 1) одне з ключових слів, що ідентифікує документ в інформаційно-пошукових системах; 2) одиниця мови інформаційно-пошукової системи, що відповідає певному базовому поняттю, включеному в тезаурус цієї системи.*

***Ключові слова:*** *лінгвістична індексація, пошук за ключовими словами, дескрипторна мова, види індексування, пошук інформації, мова пошуку інформації, структурування та стандартизація екосистеми текстів, система індексування.*

Nowadays the issue of structuring and standardizing the information space occupies the central place in modern linguistic studies. The process of summarizing information is quite complex and extended as, firstly, it is needed to carry out an analytic and synthetic analysis of information including its selection, evaluation and generalization. The phenomenon of linguistic indexation is an integral part of the aforementioned process.

Linguistic indexation is defined as the process of identifying linguistic features which directly refer to the context where the utterance unfolds. In other words, it's the process of conferring particular keywords and codes to the text (Giorgi Alessandra: 2010, p. 21). These keywords and codes take on the role of document's content identifier and are used to search for this document.

Linguistic indexation ensures high effectiveness of searching for the necessary documents and data through the use of two most meaningful criteria 'accuracy' and 'completeness'. Its primary aim is fast and highly-effective search of required information in texts' ecosystem. Thus, linguistic indexation is mostly studied within structural paradigm. Being based upon the language consistency, its level hierarchy, the existence of systemic relations at all language levels, duality of language units' variants and invariants, it enables the study of diverse linguistic indexation types, peculiarities of keyword search as well as an effective model of linguistic indexation (Селіванова О.О.: 2008, p. 137).

Thus, information search is a set of operations necessary to obtain information that corresponds to the user's request, while the information search system is defined as an organized set of documents and information technologies targeted at the storage and retrieval of information, texts or data.

Algorithms of performing actions in an information search system unfold at the level of software. It, in turn, consists of a logical and semantic apparatus and search array. A search array is a collection of documents equipped with search images, which contain those that correspond to the query. The logical and semantic apparatus is a processor that creates tools for search implementation and performs the search itself. It contains the following blocks (Кушнаренко Н.М.: 2006, p. 47):

– information search language;
– indexing system;
– logic.

The main component of an information search system is an information search language. This is a special formalized artificial language that expresses the main semantic content of a document and is intended for performing information search. Its main tool is the indexing process – a formal description of textual information. Depending on the method of constructing an algorithmic search system, information retrieval languages are divided into classification and descriptive.

The structure of languages of the first type is based on a classification system. Its key characteristic is that an object has a number of features, each of which belongs to a certain class (subject area) with its own code. All these characteristics are connected

by genus-species relations, since they are carriers of a general, broader meaning in comparison to the previous one. The first manual information search systems were created with the help of classification languages. They are the basis for organizing documents in library affairs. Classification information retrieval languages include (Сухий О.Л., Міленін В.М., Тарадайнік В.М.: 2005, p. 35):

– enumerative;
– analytical and synthetic;
– faceted.

Enumerative classification languages consist of numbered classes that are based on a certain characteristic. The top of such system is occupied by a general feature, from which the lower features depart, so enumerative classifications are also called hierarchical classifications. The advantage of languages of this type is the ease of using the created schemes. However, there are significant disadvantages of such languages. First and foremost, they are unable to update and add classes as new industries emerge and old ones expand. Secondly, the repetition of the same concepts in different fields as well as formal and logical principles of the system construction do not allow for a flexible hierarchical classification, which makes it cumbersome.

Currently, the most common universal enumerative classification languages are as follows (Лобановська І.Г.: 2011, p. 24):

– Dewey Decimal Classification (DDC). The classification was developed in 1876 by Melville Dewey. It has a large number of reprints, and at the end of the twentieth century it received an electronic version. The classification is based on decimal division: 10 main classes contain 10 sections of 10 subsections each. Indexing is fulfilled in Arabic numerals from 0 to 9. The main classes (the first level of division) denote general concepts that cover the sciences and branches of knowledge subordinate to them. The DDC was widely recognized, and it was later used as the basis for many classifications;

– Library and bibliographic classification (LBC). It was created to organize library collections and catalogs. The LBC system consists of basic tables and a system of typical divisions, sorted out by sciences and phenomena of reality. The main table contains a number of levels. The first level is divided into levels of the following order, according to the ratio of general fields of activity to their subordinate sciences. The division of the initial level for scientific libraries is indicated by letters of the alphabet, and for mass libraries – Arabic numerals, classification by which has an additional division – the unification of sciences into larger classes of the first level, and the first level of classification, which is presented for scientific libraries, is automatically transferred to the second level. The system of typical divisions creates a thematic division of textual information for quick and easy search for documents. In addition to numbers and letters, in LBC encoding is realized through string characters (dot, two dots, hyphen, parentheses, slash) and an alphabetical subject index as an additional tool for indexing.

– Another classification of information search language is faceted. It consists of a set of facets, a combination of common figurative features that are classes for certain generalizing categories. The set of facets in a search query is called 'facet formula'. The advantage of facet classification is the creation of new thematic complexes and any combination of them. However, there are also disadvantages: the complexity of the structure, inability to enter all subject features due to the large number of them. The faceted classification was first created by the Indian librarian and mathematician Shyali Ramamrita Ranganathan. It was called the Colon classification. It consists of an alphanumeric ordering of the main principle of combining classes of major subjects and 5 general categories: P (Personality, Individuality), M (Matter, Matter: MM (Matter-Material), MP (Matter-Property), E (Energy), S (Space), T (Time). Each category has its own facets – a digitally systematized set of common concepts, features of an object, called isolates. The search is based on the following principle (Лобановська І.Г.: 2011, p. 27):

1) the search query highlights the main subject that is in the list of classification;

2) finding the appropriate general categories, facets, and isolates for the subject;

3) creating a facet formula using special characters.

– Another type of classification information search languages is analytical and synthetic. The peculiarity of this system is the division of the document into separate independent features (classes), which will be combined while searching for the necessary textual information combining into a whole, and characterizing it from general to more specific (searching for the number of the last class). Thanks to the possibility of assigning its own number to a new subject area, the system has the ability to be updated, and with the help of additional tables and notations, the issue of massiveness is addressed. However, this classification complicates its creation, as you need not just to select the desired feature of the ready-made list, but also create a new class number. The bright example of analytical and synthetic language is Universal Decimal Classification (UDC). UDC is based on faceted associations – the first facet

is the main facet, the others are auxiliary. The first facet (main table) contains classes of general sciences, which are divided by analogy to the Dewey Decimal Dewey's classification. The auxiliary facets provide some additional information about the relationships between classes of knowledge fields and are divided into general and special indices according to the type of indexes. General indices can combine concepts from different sections and are indicated by symbols, respectively to their own functions. Special indexes indicate links only within a certain section of the main table and, thus, refine the information. The UDC system uses a certain sequence of rules and linking symbols, which are used to create a faceted formula. Despite the high flexibility of the classification, it also has disadvantages. The system has many indices that have a common symbol, but denote different relationships, which leads to ambiguity in the rules of building classification. Nowadays, the UDC is one of the most common classification systems in the world (Giorgi Alessandra: 2010, p. 64).

Descriptive information search languages form a group of languages of a different type. The system is based on the descriptive method. It is carried out with the help of keywords that represent the main content of the document or query. Keywords can include words and phrases of a nominative nature. A keyword that expresses the most general, main meaning, which can be used to accurately describe the content of a document or request is called a descriptor. Being organized in an alphabetical order, descriptors and their synonyms form a descriptive dictionary. By its nature, it is only a list of lexical items that may appear when indexing a particular information text (Сухий О.Л., Міленін В.М., Тарадайнік В.М.: 2005, p. 56).

More complex relationships between descriptors and their meanings are reflected in the information search thesaurus, which has become the main tool for automatic search. It is a structured vocabulary for vocabulary control, organized in a systematic and alphabetical principle that conveys the basic semantic relations (equivalences, hierarchical and associative) between natural language terms and is capable of being modified and updated. In addition to the descriptor, the thesaurus also has an ascriptor (nondescriptor), which is a lexical unit that cannot be used for indexing in a search image (query) and must be replaced by the corresponding descriptor. Therefore, each keyword is not a descriptor, but with the help of semantic relations in the thesaurus, it is associated with a descriptor of its class (Кушнаренко Н.М.: 2006, p. 157).

Most information search languages of the classification type are characterized by limited search and rigidity of the structure itself. As a result, any query must be categorized into a particular class, which does not always give a positive answer. Whereas descriptive languages analyze the document by the features that are relevant to the user's and speed up the search process. However, even here there is not always an exact the exact answer to the query. Currently, information search systems of the thesaurus-free type are being researched, the process of which occurs at the level of natural language, which significantly increases the completeness and accuracy of the textual information found (Лобановська І.Г.: 2011, p. 17).

When using a descriptive language, the main content of a query or document is expressed as a set of natural language words or phrases. Words and phrases are the names of certain classes of concepts. A word or phrase that is part of a search image specifies the coordinates of a document in a multidimensional feature space.

A descriptive language is used for coordinate indexing of documents and queries using thesauri (descriptive dictionaries) or keywords using natural language. The basis of descriptive language is based on an alphabetical list of lexical items. A set of keywords is a kind of lexical model of a research text. The functional significance of which is determined by the fact that they are one of the most optimal ways of classifying, storing, and transmitting information. Reflecting development and terminology dynamics of a particular research field, it is also a system for tracking and disseminating modern terminology (Лобановська І.Г.: 2011, p. 27).

As it is known, coordinate indexing is a type of indexing in which the content of a document or query is expressed in many ways by a set of keywords or descriptors. For instance, in the abstract of research article, the number of such words is much less than the standard (500–600 characters). In this way, the information contained in the document is collapsed and presented in the form of an index.

A descriptor is a lexical unit, word, phrase of an information search language and is a descriptive element, as well as keywords, which should reflect the content by sections, preface or annotation of the document (Селіванова О.О.: 2008, p. 72). Therefore, the keywords of an abstract are an illustration of the main parts of a research article, those meta-elements that the author is trying use in order to influence us, the recipients.

When using a descriptive language, the main content of a query or document is expressed as a set of natural language words or phrases. Words and phrases are the names of certain classes of concepts. A word

or phrase that is part of a search image specifies the coordinates of a document in a multidimensional feature space.

For coordinate indexing of documents or queries, it is possible to use words that are directly selected from the indexed texts. Such words and phrases are called keywords.

Such searching is reduced to a formal comparison of the search image of a document and the search instruction (query). However, a simple selection of keywords from the text is difficult for the following reasons (Кушнаренко Н.М.: 2006, p. 91):

1) the same words may be spelled differently;

2) there are many synonyms and homonyms among the keywords;

3) keywords do not determine the generic-specific relations between words.

To eliminate these shortcomings, a special lexical and grammatical control is performed when developing a descriptive language, and special dictionaries, diagrams, and tables are built to express paradigmatic relations between indexing terms. A special syntax is developed for the descriptor language.

Lexicographic control means that all keywords are reduced to their normal form, i.e., to the same spelling and complete elimination of synonymy, homonymy, and all kinds of ambiguities with the help of a special normative dictionary. This dictionary lists all keywords in a single morphological form. From the list of keywords, the words that are synonymous are selected. These keywords are grouped into conditional equivalence classes (paradigms). From each such group, one word or phrase is selected that is the semantic dominant of the group, i.e., it most fully defines the meaning of the words of the corresponding group (Сухий О.Л., Міленін В.М., Тарадайнік В.М.: 2005, p. 47).

To sum up, one of the main types of indexing is descriptive language, which is a type of indexing in which the semantic content of a document or query is expressed in many ways by a variety of keywords or descriptors. An information search language designed for coordinate indexing of documents (or parts thereof) and queries using keywords or descriptors is called a descriptor language. Descriptive languages began to be created in the USA in the 1950s. The term 'descriptor' was coined by the mathematician Calvin Muers. He defined a descriptor as 'a verbal symbol used to denote an idea or concept'. In order to find out the key content of a document and translate it into descriptor language, it is necessary to perform an intellectual analysis of the text. Indexing should be done with the help of special dictionaries. In the modern sense, a descriptor is a lexical unit expressed by an informative word or code and is the name of a class of synonymous or similar keywords. A descriptive language is used for coordinate, or as it is also called, 'free' indexing of documents and queries using descriptors or keywords. Descriptive languages are based on an alphabetical list of lexical items. They allow to disclose the content of documents in a sufficiently detailed and multifaceted manner. Descriptors and keywords can be easily supplemented and updated, as any lexical item required for indexing can be included in the alphabetical list. In addition, bibliographic description languages, object and feature-based search languages, and factual search languages are widely used in automated technology.

Further development of indexation is linked to the need to create a reliable system for maintaining and supporting information and linguistic support. The relevance of unifying indexation is even greater if we take into account the prospects for exchanging bibliographic databases, creating electronic consolidated catalogs, and mutual access to information resources. The success of electronic forms of information and their use in libraries depends on strict adherence to the requirements of standards and cataloging methods.

## BIBLIOGRAPHY

1. Giorgi Alessandra. About the Speaker: Towards a Syntax of Indexicality. New York: Oxford University Press, 2010. 229 p.
2. Кушнаренко Н.М. Наукова обробка документів: підручник. Київ: Знання, 2006. 334 с.
3. Лобановська І.Г. Індексування документів ключовими словами. Київ: Нілан-ЛТД, 2011. 32 с.
4. Селіванова О.О. Сучасна лінгвістика: напрями та проблеми. Полтава: Довкілля-К, 2008. 711 с.
5. Сухий О.Л., Міленін В.М., Тарадайнік В.М. Алгоритми пошуку в інформаційних системах. Київ, 2005. 70 с.

## REFERENCES

1. Giorgi Alessandra. About the Speaker: Towards a Syntax of Indexicality. New York: Oxford University Press, 2010. 229 p.
2. Kushnarenko N.M. (2006) Naukova obrobka dokumentiv [Scientific processing of documents]. Kyiv. 334 s. [in Ukrainian]
3. Lobanovska I.H. (2011) Indeksuvannia dokumentiv kliuchovymy slovamy [Indexing of documents by keywords]. Kyiv. 32 s. [in Ukrainian]
4. Selivanova O.O. (2008) Suchasna linhvistyka: napriamy ta problemy [Modern linguistics: directions and problems]. Poltava. 711 s. [in Ukrainian]
5. Sukhyi O.L., Milenin V.M., Taradainik V.M. (2005) Alhorytmy poshuku v informatsiinykh systemakh [Search algorithms in information systems]. Kyiv. 70 s. [in Ukrainian]