



Language as an important means of information encoding

Iryna Denis*

PhD in Historical Sciences, Associate Professor
Department of Psychology and Humanities and Social Studies
Ukrainian National Forestry University
<https://orcid.org/0000-0001-9646-2100>

Abstract. The purpose of the article is to study the significance of language as a means of encoding information, to prove its influence on the process of storing and transferring information. The research methodology consists in the application of various approaches: analytical – to study the literature on the topic of the article, systematic – to study the language as a means of coding information using general scientific methods, first of all systematisation and generalisation, as well as a comparative historical method. The scientific novelty lies in the fact that the paper shows the importance of language as a necessary means of encoding information. Since ancient times, language has been a necessary means of communication in society. It did not have such clearly defined forms, was inarticulate, did not have specific rules and was not standardised, but with the development of society, language becomes more normalised. This played a very important role in the existence of mankind, because with the help of language, our ancestors managed to pass on a huge amount of experience, customs, traditions, cultural heritage and certain norms to their descendants, and today we already consider the concept of language as a tool for encoding the necessary information in various systems. And from this, accordingly, there arises a need to study language as a means of encoding information. Conclusions. Due to the need to encode information, a variety of techniques began to appear to solve this problem. Of course, methods of encoding information with the development of mankind have changed from a simple shift of the alphabet to the emergence of modern digital systems. And as a result, the emergence of the programming language as a special subtype of the artificial language is a consequence of the development of computing technology. This is how special font complexes began to appear to designate characters in electronic computer technology and systems built into it

Keywords: language; information; communication; information encoding

The relevance of the topic of the study

Since ancient times, language has been an important and necessary means of communication in society. It did not have such clearly defined forms, it was inarticulate, it was not normalised, but with the development of society, the language has become more standardised and normalised. This played a significant role in the existence of mankind, because with the help of language, our ancestors managed to transfer a huge amount of experience, customs, traditions, cultural heritage and certain norms to their descendants, and today we already consider the concept of language as a tool for encoding the necessary information in various systems. Hence, there arises the need to study language as a means of

encoding information. Accordingly, the purpose of the article is to clarify the significance of language as an important means of encoding information, to prove its direct influence on the process of storing and transferring information. The scientific novelty is that the paper analyses the importance of language as a necessary means of encoding information.

Analysis of studies and publications

It is generally known that language is one of the main means of transferring information. The well-known researcher V.M. Bryzhko (2017) argued that for a person, an individual piece of “information” is only messages

Suggested Citation:

Denis, I. (2024). Language as an important means of information encoding. *Library Science. Record Studies. Informology*. 20(1), 8-11. doi: 10.63009/lrsi/1.2024.08.

*Corresponding author



Copyright © The Author(s). This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (<https://creativecommons.org/licenses/by/4.0/>)

or data that may reduce (or not reduce) the recipient's ignorance, and its subject connection may indicate (or not indicate) the emergence of an information product (Bryzhko, 2017). The problems of symbolic character of language and its individual levels, the essence of general concepts of system and structure, and the connection of language with society were studied by the linguist O. Melnychuk (1981). The pragmatic approach to the analysis of communicative aspects of language is analysed in the scientific research by L.R. Bezugla (2012). The well-known linguist M.P. Kochergan (2006) investigates the system-forming properties that a language system endows an object with and which it does not have outside the system. The fundamental properties of artificial languages, the features of modern programming languages in view of their connection with natural language are analysed by I.P. Biskub (2015), who came up with the idea of creating a universal artificial language for communication with a computer using the UNL language as an example. The terminological definition of artificial languages, their interpretation in the synchronic and diachronic aspect was carried out by A. Volkov & M. Skab (2001). A detailed description of the principles of development and implementation of modern programming languages was made by M. Gabrielli & S. Martini (2010). R. Sebesta's (2009) work deals with the study of the concept of programming languages, a critical assessment of their current and future state. However, there is still no thorough study on the features of language as a means of encoding information in the scientific literature.

Presentation of the main material

From time immemorial, in every society the main way of encoding information has been a language; furthermore, in each state it can be different or even several languages can be used at the same time. An example with the designation of snow: in the Eskimo language – more than 30 names, in the Ukrainian language – up to half a dozen.

It should be noted that today there are many methods for encoding information, but one of the first was the cryptographic method. Cryptography (Greek: *krypto* – secret; *grafo* – writing; secret writing) is a type of special writing that makes it possible to write down a written text recorded in natural written language in a way that makes its letter-by-letter recognition maximally complicated. Cryptography consists of two opposite processes – encryption and decryption of texts. Encryption is changing the image of characters in the text, their codes or order based on certain predetermined rules. Such rules are called the key. Cryptography is used when there is a need to make the text available (understandable) only to a certain circle of recipients who know the key (for the rest of the recipients, the text must be incomprehensible) (Kuleshov, 2000).

As researchers note, in the 5th century BC Spartans began to encrypt military messages. To do this, the

senders and recipients of the text were to have sticks of the same diameter. The sender of the text spirally wound a narrow tape around one of the sticks and wrote down the desired text horizontally. The recipient had to wind the tape around the similar stick and read the text horizontally in the same way. A different diameter of the stick made it impossible to read the text (the letters shifted vertically, resulting in a gobbledygook). This method of encrypting texts was called the Skytala cipher after the name of its inventor.

Later in the 1st century BC, the Roman commander Julius Caesar encrypted his messages in a slightly different way, shifting the second (encrypting) tape of the alphabet several positions relative to the first one, presented in the correct sequence. Then, in the text, instead of each letter, he substituted the one that stood in the second line of the alphabet. As a result, the word empire after encryption took on the following form: жймгнжщ (this is if we translate the system of that time into Ukrainian).

Later in the 16th century, Blaise de Vigenere created another original method of encryption, which was based on a table specially formed from the alphabet. To encrypt the text using this table, for each letter of the text it was necessary to indicate its two coordinates: horizontal and vertical. In this case, the same letter received different codes depending on the line from which it was selected. Coordinates could be recorded both with letters and numbers (Kuleshov, 2000).

In contrast to information encryption, the concept of “decryption” is also used. Decryption is the restoration in the text, based on the key, of the primary image of the characters, their codes or order. Decryption is possible without knowledge of the key, but in this case, it requires much more time to reconstruct the text (Kuleshov, 2000).

Usually, cryptography is practically not used in mass media (except for some crossword puzzles). However, sometimes mass media workers need to use cryptography during journalistic investigations or when performing the functions of foreign correspondents. It should be noted that among the cryptography methods, the following are distinguished:

- one-dimensional, two-dimensional, three-dimensional encryption keys – for encryption on paper;
- email encryption via mailboxes;
- file passwording using different methods.

According to S. Kuleshov (2000), the most striking achievement of the past was the deciphering of ancient Persian, Acadian, and Sumerian texts (Grotefend, 1802), as well as ancient Egyptian writing (Cham-pollion, 1822). The basis for deciphering by Grotefend was a comparison of the titles of two kings, and for Champollion, a comparison of the names of the kings (in the first and second cases, the titles and names were presented in different languages, one of which was known).

Another way of encoding traditional writing, which until recently was used in electrical communication channels, is called the Morse code after the surname of its author. Characters in the Morse code are encoded using short ("dots") and long ("dashes") electrical signals. In this case, the dot is considered a unit of duration. The duration of the dash must be equal to the duration of three dots. The pause between signals in a letter (of alphabet) is equal to one dot, and the pause between letters is equal to three dots (Kuleshov, 2000). It is known that the Cyrillic version of this alphabet is based on the external (graphic) similarity of the letters with the Latin ones. In cases where Cyrillic letters do not have graphic analogues, codes are assigned to them arbitrarily. In the Morse code, there are also notations for punctuation marks (period, comma, exclamation mark, question mark, semicolon, colon, parentheses, hyphen, quotation marks) and special signals (section boundary, erasure of previous character, beginning of transferring, readiness to receive, start of action, end of transferring). It should be noted that the total number of "dots" and "dashes" in codes denoting letters depends on the frequency of letters in texts (the more frequently a letter is used, the shorter its designation). Due to the transition to the use of computer technology (respectively, digital codes), the use of the Morse code (analogue signals) has recently been declining. It should be noted that today it is used most often by radio amateurs.

According to scientists, with the advent of computers, in order to organise their communication with people, a task arose – to create special fonts (font sets) for them, in which letters are indicated not by graphic (as in traditional writing), but by digital images (digital codes). A corresponding graphic image of the character was "attached" to each such digital code (Kuleshov, 2000). It should be noted that the font sets supplied with the Windows operating system include Ariel, Courier New, Times New Roman as mandatory ones. In addition, each such font set includes 224 characters, among which there is the Latin alphabet, and some others may also be present.

UNICODE font sets contain approximately 60000 characters. Each such set consists of alphabets and hieroglyphs of most languages of the world. Accordingly, in these font sets:

- codes from 0 to 8191 (hexadecimal codes 0...1FFF) are assigned to the alphabets of all European languages, Hebrew, the alphabets of Arabic and Indian languages, as well as phonetic signs;
- codes 8192...12287 (hexadecimal 2000...2FFF) – for punctuation, mathematical, technical symbols and ornaments;
- codes 12288...16383 (hexadecimal 3000...3FFF) – for phonetic Chinese, Japanese, and Korean writing;
- codes 16384...59391 (hexadecimal 4000...E7FF) – for Chinese, Japanese and Korean characters, as well as Han calligraphy;

- codes 59392...64023 (hexadecimal E800...FDFF) – for private use;

- codes 64024...65535 (hexadecimal FE00...FFFF) – to ensure compatibility with other fonts (Toftul, 2006).

Interestingly, researchers identify three stages in creating font sets. At the first stage, each character (including each letter) was encoded with one byte. Since a byte can represent no more than 256 different combinations of binary numbers, such a font set could contain no more than 256 characters. In fact, only 224 codes (starting from the 32nd to the 255th) were used to encode characters in such a font set, and the remaining 32 (from the 0th to the 31st) were used to encode the commands of displays and printers. The graphic representation of letters had only one single letter shape (like on a typewriter). Such type sets were recorded permanently in special chips of displays and printers. Such fonts were used, for example, by the MS DOS operating system (Kuleshov, 2000). At the second stage, just like at the first one, each character continued to be encoded with one byte, however, the graphic representation of the characters could already be varied (for example, in the fonts Courier New, Times New Roman, Ariel). Such fonts were no longer permanently recorded in special display or printer chips, but were used as regular computer information support, which, like any programmes, can be installed and, if necessary, removed. To present information about the graphic design of a character in these font sets, special standards were developed. A very large number of such digitised fonts were developed. They were used with other types of operating systems (these were systems like the Windows family). Such fonts contained: in the area of codes from 32 to 127 – in particular, letters of the Latin alphabet (lower and uppercase); in the area of codes from 128 to 255 – either Latin letters with diacritical marks (for languages based on the Latin alphabet), or letters of other alphabets (Greek, Cyrillic, Hebrew, Arabic, etc.). Thus, the code zone from 128 to 255 was variable, which excluded the possibility of communication between people who used different alphabets (for example, simultaneously Latin with diacritics and Cyrillic, Arabic and Cyrillic) (Kuleshov, 2000). At the third stage – its emergence was caused by the needs of the Internet – in order to ensure the possibility of communication between people exchanging information in different languages (for example, based on the Latin alphabet and hieroglyphs), it became necessary to significantly expand the number of characters (mainly letters, ligatures and hieroglyphs) in these font sets. As a result of the search, font sets were created in which one character was encoded not by one, but by two bytes (two bytes make it possible to encode more than 60000 different combinations of binary numbers, and, therefore, characters). Such fonts are called UNICODE. Like the fonts of the second stage, they are used in operating systems such as the Windows family (Kuleshov, 2000).

Conclusion

Language has made a significant development path from the simplest function of ensuring society's communication to the creation of artificial languages by mankind itself to meet the need for fixing and communicating concepts in certain narrow circles. Thus, certain means of encrypting information began to appear, ranging from cryptography to modern computer systems, without which we can no longer imagine our daily life. In addition, the rapid globalisation of all spheres of human activity has given impetus to the strengthening of international contacts between individuals and entire communities. These processes constantly contributed to the emergence of artificial languages

that would be universal and accessible for performing specific tasks.

Methods of encoding information in the course of human development have changed from the simplest shift of the alphabet to the emergence of modern digital systems – programming language as a special subtype of artificial language, which is a consequence of the continuous development of computing technology.

This is how special font sets began to appear to designate characters in electronic computer technology and systems built into it. And in the course of time and further development of science, these systems in the future will require new approaches to their study and thorough scientific research.

References

- [1] Bezugla, L.R. (2012). [Linguistic pragmatics and discursive analysis](#). *Studia Philologica*, 1, 95-100.
- [2] Bidzilya, Yu. (2021). *Current problems of journalism*. Uzhhorod.
- [3] Biskub, I.P. (2015). [The language of expert systems: Artificial or natural?](#) *Scientific Bulletin of Lesia Ukrainka East European National University. Philological Sciences. Linguistics*, 4, 171-177.
- [4] Brus, M.P. (2004). [Ukrainian business broadcasting](#). Ivano-Frankivsk: Tipovit.
- [5] Bryzhko, V.M. (2017). Information product as an object of property rights. *Information and Law*, 4(23), 5-15. [doi: 10.37750/2616-6798.2017.4\(23\).273091](#).
- [6] Gabrielli, M., & Martini, S. (2010). *Programming languages principles and paradigms*. London; New York.
- [7] Kochergan, M.P. (2006). [General linguistics](#). Kyiv.
- [8] Kuleshov, S.H. (2000). *Documentary studies: History. Theoretical foundations*. Kyiv.
- [9] Kyslenko, U.I. (1997). *Systematic organisation of language*. Kyiv.
- [10] Melnychuk, O.S. (1981). Development of language as a real system. *Linguistics*, 2, 22-34.
- [11] Nyzhnyk, N.R. (2002). *Public administration in Ukraine: Organisational and legal principles*. Kyiv.
- [12] Panchuk, A.M. (2004). [Modelling, information systems and technologies in public administration](#). Kyiv: Millenium.
- [13] Sebesta, R.W. (2009). [Concepts of programming panguages](#). 9th ed. New Jersey.
- [14] Toftul, M.H. (2006). *Logic*. Kyiv: Academia.
- [15] Volkov, A., & Skab, M. (2001). Artificial languages. *Lexicon of General and Comparative Literature*, 625-626. Chernivtsi: Zoloti Lytavry.
- [16] Zagnitko, A.P. (2012). [Dictionary of modern linguistics: Concepts and terms](#) (Vol. 4). Donetsk: DonNU.