Developing Linguistic Research Tools for Virtual Lexicographic Laboratory of the Spanish Language Explanatory Dictionary

Yevhen Kupriianov[0000-0002-0801-1789], Nunu Akopiants[0000-0002-0709-9926]

National Technical University "Kharkiv Polytechnic Institute", Kharkiv, Ukraine
cuprijanow.eugen@yandex.ua, nunikaosip@gmail.com

Abstract. The present article is devoted to the problems of creating linguistic tools for the virtual lexicographic laboratory of Spanish explanatory dictionary (*DLE 23*). The goal of the research is to consider some issues related to the development of linguistic tools for the virtual lexicographic laboratory. To achieve this goal the dictionary was analyzed to define the peculiarities of linguistic facts representation, its structure and metalanguage. On the basis of the dictionary analysis and the theory of lexicographic systems the formal model of *DLE 23* was developed and its main components, including their relationships, were determined to ensure their availability via linguistic tools for accessing linguistic information. The range of research activities to be performed by using the linguistic tools was outlined.

Keywords: Computer Lexicography, Virtual Lexicographic Laboratory, Digital Environments, Electronic Dictionaries.

1 Introduction

Explanatory Dictionary of the Spanish Language (herein after DLE 23, Diccionario de la lengua española. Edición del tricentenario (23^{a} edición)), like other big explanatory lexicons, contains profound systemic language regularities, which, being almost hidden from the reader, play an important role in identifying linguistically-informative potential of the language.

In general, the properties of the language as a system, as well as their presentation in comprehensive dictionaries of explanatory type, are the object of many researches, for example [1; 2; 3; 4; 5]. In our opinion, such properties have the best representation in monolingual explanatory dictionaries. The article is focused on large, mostly multivolume lexicons, which contain the major part of the national lexicon and phraseology, and that are characterized by a detailed description of lexical-grammatical and lexical -semantic systems of the language. Due to the great amount, elaborated structure and completeness of a lexicographic description such dictionaries are carriers of a huge number of implicitly-defined linguistic, cognitive, logical and other relationships (mostly uncontrolled), making this extensive lexicographical system a kind of "thing-in-itself" [6].

This raises a question of the development of methodology and technology of creation of such lexicographical objects, and also a question of study a variety of effects that explicitly or implicitly operate there. From the beginning, we are talking about the methods of computational linguistics, because, as noted in the book "Computer lexicography" [6], it is physically impossible to perform such studies with a help of traditional methods. So, the first problem here is to create digital analogues of corresponding traditional lexicographic studies or convert them into digital form, followed by the explication of underlying systemic linguistic regularities.

2 Related works

The Ukrainian Lingua-Information Fund (Kyiv city), where studies devoted to this work had been conducted, developed a universal theoretical basis, focused on the construction of lexicographic objects of almost unlimited size and complexity, and on implementation of a profound study of the language systems. We are talking about a theory of lexicographic systems, using the theory the Fund designed and elaborated linguistic software tools to perform researches on the basis of the Ukrainian, Russian and Turkish language explanatory dictionaries [7; 8; 9], as well as the tools for etymological studies of the Ukrainian language [10].

Using these tools (and partly even before they had been created, but with approaches based on the theory of lexicographic systems), a number of studies were conducted and obtained a series of fundamentally new linguistic results for the Ukrainian language. Among these result we should mention a study and establishment of formal structure of headword rows of verb and noun (the Ukrainian language) [9; 10; 11; 12].

However, the developments of the Fund successfully applied to the Ukrainian, Russian and Turkish languages can't be automatically used for the Spanish language for the following reasons:

- typological features of the Spanish language, among them are part-of-speech variation, dependability of lexical meaning on grammatical one, the possibility of acquiring lexical meaning by a word when it functions in the specific grammatical meaning;
- peculiarities of metalanguage and entry structure of *DLE 23* to describe grammatical and lexical properties of Spanish language units.

3 Methodology

The method of formal modeling of lexicographic objects like *DLE 23* is based on the theory of lexicographical systems developed by Prof. Volodymyr Shyrokov. According to the theory of lexicographical systems, one of the most important relationships is "subject – object". In this context we will note that the subject (designated by the symbol S) can be a person or a group of persons (lexicographers, linguists, experts etc.) and the object is a set of elementary information units (EIU) marked as $I^{Q}(D)$. In other words, when processing the EIUs in its mental apparatus, the subject acquires a set of their descriptions $V\ (I^{Q}(D))$. It can be formally represented as:

$$S: I^{Q}(D) \rightarrow V(I^{Q}(D)) \tag{1}$$

For each elementary unit $x \in I^{Q}(D)$ there is a description $V(x)$ represented as a dictionary entry. In its turn, $V(x)$ is an element of $V(I^{Q}(D))$. Therefore, we can assert that $V(I^{Q}(D))$ is *DLE 23* comprised by the set of the entries $V(x)$:

$$V\big(I^{Q}(D)\big) = \bigcup\nolimits_{x \in I \langle (D)} V(x) \tag{2}$$

The description of the elementary information units $V(x)$ is represented as a text, a certain sequence of characters, which is called *A*-word. The characters of the *A*-word compose the finite alphabet: $A = \{a_1 \ldots a_n\}$. So, the alphabet of *DLE 23* covers: 1) Spanish alphabet (A ... Z, a ... z) including diacritic characters (ñ, ç; á, é, í, ó, ú, ü); 2) Greek alphabet characters (α ... ω); 3) international Latin alphabet; 4) dictionary metalanguage symbols; 5) punctuation symbols including paired symbols (¡!, ¿?); 6) font patterns.

Within each *A*-word the *A*-subwords formed of the *A*-alphabet characters, are possible to be distinguished. Let us designate the set of *A*-subwords as $B[V(x)]$. In case of *DLE 23* the *A*-subwords are as follows: 1) headword; 2) headword row; 3) headword variants; 4) etymology; 5) irregular forms; 6) orthography; 7) definition block. All of them compose the set $B[V(x)]$:

$$B[V(x)] \equiv \{\beta_i(x) \mid i = 1{,}2, \ldots n\} \tag{3}$$

Here $\beta i(x)$ is an *A*-subword and i is an index number of an *A*-subword in the entry of *DLE 23*. We'd like to note that the identification of $\beta_i(x)$-elements is based on the peculiarities of the dictionary metalanguage that clearly establishes the rules for the representation of a particular element of the dictionary entry. Thus, we have defined the first lexicographic structure to be induced within the set of descriptions $V(I^{Q}(D))$:

$$\beta\big(I^{Q}(D)\big) \equiv \beta = \{\beta_i, 1{,}2, \ldots\} \tag{4}$$

Furthermore, $\beta(x)$ will be the set of the structural elements of the entry $V(x)$ devoted to selected headword x:

$$\beta(x) = \{\beta_i(x), 1{,}2, \ldots\} \tag{5}$$

In its turn $\beta(x)$ structure may be subdivided into smaller elements as a result of σ-operator action:

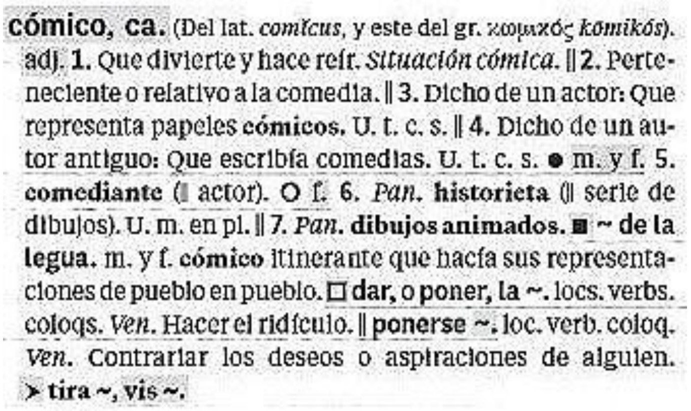
$$\sigma: \beta \rightarrow \sigma[\beta] \tag{6}$$

The application of the method based on the theory of lexicographic systems for the formal modeling of *DLE 23* will be described in the next section. But the principles of revealing $\beta(x)$-structures and respective $\sigma[\beta(x)]$ elements as well as building-up a formal modeling of *DLE 23* are not possible to be worked out in the preliminary study of the metalanguage and the entry structure of the dictionary.

4 Lexicographic system of *DLE 23*

4.1 Entry structure and metalanguage peculiarities

The entry is subdivided into the left and right parts. The left part is made up of a headword, headword row and information block. The letter contains headword variants, etymological information, orthography and word flection properties. The right part explains the meaning of the headword. The main and obligatory element is the definition. In case of lemmas with part-of-speech variation, the definitions are grouped according to the part of speech (grammatical category). The example of a dictionary entry is shown in Figure 1. Let us deeply analyze the peculiarities of representing linguistic peculiarities of Spanish words in the entry.



cómico, ca. (Del lat. *comĭcus*, y este del gr. κωμικός *kōmikós*). adj. **1.** Que divierte y hace reír. *Situación cómica.* ‖ **2.** Perteneciente o relativo a la comedia. ‖ **3.** Dicho de un actor: Que representa papeles **cómicos.** U. t. c. s. ‖ **4.** Dicho de un autor antiguo: Que escribía comedias. U. t. c. s. ● m. y f. **5. comediante** (‖ actor). ○ f. **6.** *Pan.* **historieta** (‖ serie de dibujos). U. m. en pl. ‖ **7.** *Pan.* **dibujos animados.** ■ **~ de la legua.** m. y f. **cómico** itinerante que hacía sus representaciones de pueblo en pueblo. □ **dar, o poner, la ~.** locs. verbs. coloqs. *Ven.* Hacer el ridículo. ‖ **ponerse ~.** loc. verb. coloq. *Ven.* Contrariar los deseos o aspiraciones de alguien. ➤ **tira ~, vis ~.**

Fig. 1. Entry of the headword *cómico*.

The first zone is the headword or lemma. If it has the forms both for masculine and feminine (in case of nouns and adjectives) the masculine form goes first and then the feminine form, represented only by respective ending, follows after comma. For example, *cómico*, *-ca* should be read as "*cómico* (masculine) and *cómica* (feminine)".

Both forms of the headword compose the second zone of the entry which we'll call a headword row. If a headword has only a masculine or feminine form the headword row is considered to be composed of one form.

The third zone represents the headword variants, additional lemma forms to be commonly used in different regions of Spain or Latin America. In this case respective region and other remarks may be provided in this zone.

The fourth zone contains the etymology of the headword, stating the language of origin, etymon and its original meaning.

The purpose of the fifth zone is to indicate irregular forms of the headword, for example: non-standard comparative and superlative forms of the adjectives, conjugation model of irregular verb etc.

The sixth zone of the entry shows the orthography features (e.g. writing with capital letter) which must be paid to when the headword is used in a particular lexical meaning.

The seventh zone is comprised by the definitions grouped according to each part of speech. Most of the information zones in both parts of the entry can be identified by metalanguage markers the characteristics of which are shown in Table 1.

Table 1. Information zones and their metalanguage markers.

$\beta_i(x)$	Information zone	Metalanguage marker of zone beginning	Metalanguage marker of zone end
$\beta_1(x)$	Headword	Normal bold type	Full stop "."
$\beta_2(x)$	Headword row	Normal bold type	Full stop "."
$\beta_3(x)$	Headword variants	"Tb."	Diamond "◆"
$\beta_4(x)$	Etymology	"Del" or "Quizá del"	Diamond "◆"
$\beta_5(x)$	Word-flection properties	None	Diamond "◆"
$\beta_6(x)$	Orthography features	None	Diamond "◆"
$\beta_7(x)$	Definition block	First grammar note	Black square "■"

The information zones corresponding to $\beta_5(x)$ and $\beta_6(x)$ don't have their own metalanguage marker and they can be only identified in the entry by their place in the sequence of information zones. For example, $\beta_5(x)$ always goes after etymology. The definition block, in its turn, can be decomposed in $\beta_7(x)^{GRAM}$ (corresponding to part-of-speech note and / or grammar category), $\beta_7(x)^{PRAGM}$ (a group of notes denoting pragmatic use of the headword, e.g. domain, geographic area, social dialect etc.) and $\beta_7(x)^{SEM}$ (definitions corresponding to $\beta_7(x)^{GRAM}$ and $\beta_7(x)^{PRAGM}$). To group the definitions according to each $\beta_7(x)^{GRAM}$ the following metalanguage means are used in *DLE 23*:

- black circle "●" to identify the group of definitions belonging to the part of speech of the headword;
- white circle "○" to identify the group of definitions belonging to the grammar category of the headword;
- vertical parallel bars "||" to separate definitions within the group marked with black or white circle.

4.2 Model of the lexicographic system of *DLE 23*

In the structure of dictionary entries, we distinguish the set of register (head) lexical units $W = \{x\}$, which serve as the identifiers of the corresponding dictionary entries $V(x)$. The *DLE 23* register includes words and morphemes, certain phrases and abbreviations. Representation of morphemes, phrases and abbreviations as headwords is not inherent for most explanatory dictionaries. For convenience, all language units that act as a headword will be called headwords.

In the structure of each dictionary entry $V(x)$ there is a "left part" $L(x)$, which consists of certain headword parameters, and there is "right part" $P(x)$, in which the lexicographic representation of the semantics of a headword x is given. In the case of

DLE 23, we distinguish two types of language units: lexical level units and collocations (which include the headword). Therefore it is natural to present the structure of the dictionary unit $V(x)$ in the form of a combination of descriptions (dictionary entries) of structural units of both types:

$$V(x) \equiv V^{Lex}(x) \cup \left[\cup_i^{n(x)} \cup_j^{m(i)} V_i^{jFras}(x)\right] \tag{7}$$

Here $V^{Lex}(x)$ is a lexicographic description of the headword x; $V_i^{\,jFras}(x)$ is a description of the j-th phrase of i-th type; $m\,(i)$ is the number of phrases of i-th type, and $n(x)$ is the number of phrase types in the dictionary entry $V(x)$. Each lexicographic description $V^{\,Lex}(x)$ and $V_i^{\,Fras}(x)$ corresponds to the basic structure shown in Figure 2.

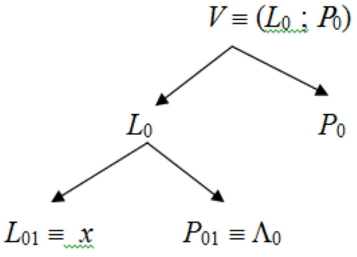


Fig. 2. The structure of lexicographic description in *DLE 23*.

In the case of $V = V^{Lex}(x)$, the headword of the dictionary entry with the corresponding parameters (which we shall now name the parameters of the headword) acts as L_0. For $V_i^{\,Fras}(x)$, L_0 is the phrase in the register dictionary form plus the parameters of the head unit. The structure of the right part P_0 is identical for a lexical unit and phrase. Arrows hereinafter indicate relations of inclusion. The text analysis of the dictionary entries showed that the structures $V(x)$ are almost identical for collocations and headwords.

To build up a formal model of the lexicographic structure of *DLE 23*, taking into account its features mentioned above, we relied on the theory of L-systems by Volodymyr A. Shyrokov [5], according to which any dictionary can be represented as:

$$\{I(D), V\big(I(D)\big), \beta, \sigma[\beta], Red\big[V\big(I(D)\big)\big]\} \tag{8}$$

Here D is *DLE 23*; $I(D) = \{x_i\}$ is a set of headwords; $V(I(D)) = \{V(x_i)\}$ is a set of lexicographic descriptions, namely dictionary entries; β is a set of structures within $V(I(D))$ marked out during dictionary text analysis; σ[β] is a single structure generated by the operator σ on β; The limitation of the operator σ on $V(x_i)$ generates the dictionary entry $\sigma[x_i]$; $Red[V(I(D))]$ is a recursive reduction mechanism which detects more subtle structural elements of the dictionary. In its turn, the set of lexicographic descriptions of each unit $x_i \in I(D)$ can be decomposed in several subsets (Figure 3). The equation (8) will be hereinafter considered as the model of lexicographic system of *DLE 23*.

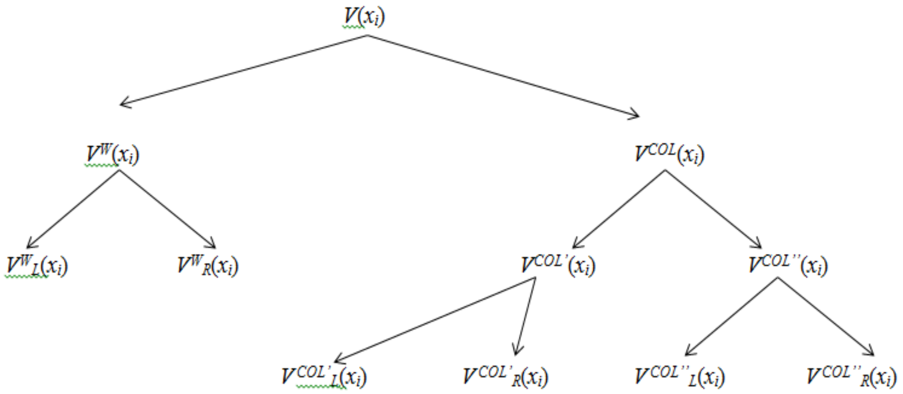


Fig. 3. Set of lexicographic descriptions and its subsets.

By $V^W(x_i)$ and $V^{COL}(x_i)$ we designate the sets of descriptions of headword and derived collocations, respectively; $V^W_L(x_i)$ and $V^W_R(x_i)$ correspond to the “left” and “right” parts. Further, the set $V^{COL}(x_i)$ is divided into the two subsets: a) $V^{COL'}(x_i)$ for collocations of “noun + adjective” type; b) $V^{COL''}(x_i)$ for collocations of other types, e.g. verbal, adverbial, prepositional etc.

The content example of β-structures for V^W(*cómico*) and $V^{COL'}$(*cómico de la legua*) and $V^{COL''}$(*ponerse cómico*) is shown in Table 2, based on the entry text (Fig. 1).

Table 2. Contents of *β*-structures of *V*(*cómico*), $V^{COL'}$(*cómico de la legua*) and $V^{COL''}$(*ponerse cómico*).

***β*-structure**	**V^W(*cómico*)**	**$V^{COL'}$(*cómico de la legua*)**	**$V^{COL''}$(*ponerse cómico*)**
*β*₁(*cómico*)	cómico	cómico de la legua	ponerse cómico
*β*₂(*cómico*)	cómico, cómica	cómico de la legua	ponerse cómico
*β*₃(*cómico*)	∅	∅	∅
*β*₄(*cómico*)	Del lat. comĭcus […].	∅	∅
*β*₅(*cómico*)	∅	∅	∅
*β*₆(*cómico*)	∅	∅	∅
*β*₇(*cómico*)	adj. **1.** Que divierte y hace reír. […]	**cómico** itinerante que […].	loc. verb. coloq. *Ven.* Contrariar […].

4.3 Lexicographic objects σ[β$_j$(*x$_i$*)]

As it was stated above the $\beta_j(x_i)$ are possible to be split into smaller objects $\sigma[\beta_j(x_i)]$ generated by the σ-operator. In *DLE 23* there are $\sigma[\beta_j(x_i)]$ objects corresponding to:

1. a certain information element of the entry, e.g. only etymology, headword variants, definition block etc. (σ_0);
2. linguistic information distribution in the entry, e.g. distribution of lexical meanings by grammatical classes or categories of a headword (σ_1);

3. information provided by the dictionary metalanguage, e.g. definition types (σ_2), origin, etymon language, number of definitions etc.
4. linguistic information concerning the interaction of different language levels, e.g. morphology – semantics; (σ_3).

The full list of $\sigma[\beta_j(x_i)]$ objects as a result of application of σ-operator to $\beta_j(x_i)$ structures, is given in Table 3.

Table 3. Lexicographic objects $\sigma[\beta]$ and their content.

$\beta_j(x_i)$	σ-type	**Linguistic facts to be derived from σ[β]**
$\beta_1(x_i)$	σ_0	Headword or lemma
	σ_2	Headword type (word, prefix, suffix)
	σ_2	Origin (indigenous, borrowed)
	σ_2	Homonymy
$\beta_2(x_i)$	σ_0	Availability of headword row
$\beta_3(x_i)$	σ_0	Availability of headword variants
	σ_1	Headword variants having some restrictions for their use
	σ_1	Headword variants without restrictions for their use
$\beta_4(x_i)$	σ_0	Etymon form
	σ_2	Language the etymon comes from
	σ_2	Etymon language
$\beta_5(x_i)$	σ_0	Availability of irregular forms
$\beta_6(x_i)$	σ_0	Orthography
$\beta_7(x_i)$	σ_2	Part of speech
	σ_2	Grammatical category
	σ_2	Part of speech variation
	σ_0	Peculiarities in pragmatic use
	σ_2	Pragmatic characteristics
	σ_1	Dependability of lexical meaning on grammatical meaning
	σ_1	Dependability of lexical meaning on pragmatic use
	σ_2	Definition type
	σ_3	Ability to form collocations
	σ_3	Semantics of derived words
	σ_2	Headword type (monosemantic, polysemantic)
	σ_2	Number of definitions

4.4 Interface

To conduct linguistic researches on the basis of *DLE 23* the interface of the virtual lexicographic laboratory allows:

- the access to the subsystems $V^{LEX}(x)$, $V^{COL'}(x_i)$ and $V^{COL''}$ of the Spanish language dictionary (vocabulary and collocations);
- selection of $\sigma[\beta_j(x_i)]$ structures within the subsystems to get linguistic facts of any kind (for example: etymology and semantics of a Spanish word);
- logical operations on $\sigma[\beta_j(x_i)]$ structures to reveal a group of Spanish language units with common linguistic peculiarities.

Applying logical operations “AND”, “OR” and “NOT” to different $\sigma[\beta_{\mathrm{j}}(x_i)]$ objects through the interface, it is possible to make a sample of Spanish language units having common linguistic characteristics. For example, hereinafter we form a sample of words the suffix of which denotes the process and result. In this case the set of $\sigma[\beta_{\mathrm{j}}(x_i)]$ is as follows: $\sigma_0[\beta_1(x_i)]$ = «aje» AND $\sigma_2[\beta_1(x_i)]$ = «m.» AND $\sigma_3[\beta_1(x_i), \beta_7(x_i)]$ = «Acción y efecto de + X». As a result, we’ll get the words with suffixes: *-ado* (*lavado*, *peinado*), *-aje* (*etiquetaje*, *embalaje*), *-ión* (*cubrición*, *gestión*).

The linguistic tools for conducting researches of the Spanish language are grouped in respective tabs devoted to each *DLE 23* information element, for example “Headword”, “Headword variants”, “Etymology” etc. Each tab corresponds to specific $\sigma[\beta_j(x_i)]$-structure and contains checkboxes to set σ-links (“Origin”, “Homonymy”, “Headword type” etc.).

In our opinion, the above mentioned features of the interface distinguish the virtual lexicographic laboratory from other electronic (online) lexicographic resources (www.dle.rae.es, www.oed.com, dictionary.cambridge.org etc.).

5 Conclusions

One of the main tasks of modern computer lexicography is updating and supporting fundamental lexicons – large paper explanatory dictionaries – in digital environment. Computer lexicography successfully solves this task by combining many years of traditional lexicography experience with the latest computer technologies. Virtual lexicographic laboratories (VLLs) are the result of such combination, i.e. systems that enable both the operation of dictionary material and the conducting series of linguistic studies.

The virtual lexicographic laboratory provides the users with linguistic tools with a wide range of opportunities for the study of grammatical, semantic, pragmatic, and other features of the Spanish linguistic units. Unlike digital dictionaries and dictionary writing systems the VLL offers a software interface for implementation of:

- access administration functions: users authorization and identification; new users adding and removing; access control (read only, reading and editing of the dictionary);
- lexicographic works: creation of a number of derivative dictionaries on the basis of explanatory dictionary; representation of dictionary entries in any format;
- research work: research at a certain language level, presented in the explanatory dictionary (grammar, including derivation; lexis, including semantics; pragmatics); research at the junction of language levels: grammar and semantics, word forming and semantics, semantics and pragmatics etc.

Unabridged monolingual dictionaries, among them *DLE 23*, in digital format are found to be powerful research environment facilitating the navigation and access to their structural elements and integration of language facts in one object. This can be achieved by formalizing the structure of dictionary text in a form of β-structures and σ-links.

In prospect, it is planned to develop the theory of lexicographic systems by Prof. Volodymyr Shyrokov for creating the virtual lexicographic laboratory for the dictionary of Spanish language inflection. To elaborate this laboratory it is necessary to work out the principles of word flection formal modeling, especially for the Spanish language.

6 References

1. Saussure, F.: Cours de linguistique générale, Paris (1997)
2. Mulder, J. W. F., Hervey, S. G. J.: Language as a System of Systems. In: La Linguistique. 11(2), pp. 3-22 (1975)
3. Trubetzkoy, N.: Principles of phonology, Berkeley, University of California Press (1969)
4. Mel'čuk, I. A.: Explanatory combinatorial dictionary. In: Open Problems in Linguistics and Lexicography. Polimetrica, Monza, pp. 222–355 (2006)
5. Hjemslev, L.: Prolegómenos a una teoría del lenguaje, Madrid (1971)
6. Shyrokov, V.: Computer lexicography, Kyiv (2011)
7. Ukrainian language dictionary in 20 volumes: Rusanivsky V. (ed.), Kyiv (2010)
8. Ozhegov S.: Tolkovyiy slovar russkogo yazyika, Moscow (1997)
9. Shyrokov, K., Shyrokov, V.: Zastosuvannia formalizmu nechitkykh mnozhyn dlia vyznachennia hramatychnykh staniv turetskykh sliv. In: Movoznavstvo, pp. 51-56 (2005)
10. Etymological dictionary of the Ukrainian language in 7 volumes: Melnychuk (ed.), Kyiv (1989)
11. Pogribna, O., Chumak, V., Shyrokov, V., Shevchenko, I.: Linhvistychna klasyfikatsiia ukrainskoho imennyka u svitli teorii leksykohrafichnykh system. In: Movoznavstvo, pp. 62-82 (2004)
12. Rabulets, O., Sukharina, N., Shyrokov, V., Yakymenko, K.: Diieslovo v leksykohrafichnii systemi, Kyiv (2004)