Computer-assisted Linguistic Researches on the Basis of Spanish Dictionary Text

Yevhen Kupriianov

National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine

Abstract

The article shows the research potential of the virtual lexicographic laboratory VLL DLE 23 based on the text of the Spanish Explanatory Dictionary (DLE 23). The VLL DLE 23 project has been developed in the Ukrainian Lingua-Information Fund, National Academy of Sciences of Ukraine, with the direct participation of the author. The VLL project has been planned to be carried out in several stages, namely: 1) text extraction (partial) together with HTML markup from available online version of the dictionary; 2) dictionary text analysis to identify meta-language and information elements of the entries; 3) building the model of the lexicographic system (L-system), which in a formal way displays the information elements of DLE 23 and serves as a basis for creating the database and interface. The conceptual basis for elaborating the DLE 23 L-system is the theory of lexicographic systems by the Ukrainian Academician Volodymyr A. Shyrokov. At present the VLL DLE 23 allows studying the Spanish language vocabulary in terms of its composition and classification, as well as some morphological, lexical-semantic and combinational characteristics of the words described in DLE 23. The examples of working with the VLL interface while conducting linguistic experiments, as well as interpretation of the obtained results of experiments are given.

Keywords

Virtual lexicographic laboratory, Dictionary text, Lexicographic systems, Digital dictionary-based researches, Linguistic information analysis

1. Introduction

The fundamental explanatory dictionaries, like Oxford English Dictionary, Duden, Dictionary of the Ukrainian language in 20 volumes, contain the main corpus of national vocabulary and phrase units and offer detailed description of grammar and vocabulary properties of a language. Their size can reach several hundred thousand units. With their large volume, worked out structure and exhaustive lexicographic description, such dictionaries serve as carriers of a huge number of linguistic, cognitive, logical and other relationships (in most cases uncontrolled). But these relationships are represented implicitly in the dictionary, thus complicating to some extent their detection and study. In this regard we find it necessary to develop a methodology which would help to reveal the linguistic information hidden in the dictionary text. According to the authors [1], such "extraction" from the dictionary becomes a real problem due to this implicitness. After all, even a linguistically educated and skilled user of a large dictionary actually "slips" on its surface without having effective tools to penetrate into the depths of its content. However, such penetration is not just expedient or desirable, but it could be an effective source of new linguistic facts about the language system and even language regularities.

The study offered is of current importance due to trends in modern digital lexicography where the efforts are made to implement a wide range of approaches to the comprehensive representation of vocabulary and grammar properties of language units, create integral lexicographic systems

COLINS-2022: 6th International Conference on Computational Linguistics and Intelligent Systems, May 12–13, 2022, Gliwice, Poland
EMAIL: eugeniokuprianov@gmail.com (Y. Kupriianov)
ORCID: 0000-0002-0801-1789 (Y. Kupriianov)

combining different linguistic facts and to build tools for working with dictionary in the digital environment.

The up-to-date information technologies offer easy resolutions of the problems of the dictionary content size, ways of linguistic information presentation and, consequently, help the linguists navigate the dictionary text and find in it not only individual facts about language, but also certain regularities. Among such technologies are virtual lexicographic laboratories (VLLs) designed to manage dictionary text in digital environment. The object of our study is the explanatory dictionary of the Spanish language of the 23rd edition (Diccionario de la lengua española. Edición del Tricentenario; hereinafter DLE 23), published by the Royal Academy in October 2014.

2. Related works

The Ukrainian lingua-information fund, National Academy of Sciences of Ukraine, developed the general theory of lexicographic systems and successfully applied it while creating a number of VLLs for different languages, first of all Ukrainian, Russian, German and Turkish [2]. It made possible to develop both the theoretical framework and technological principles of lexicographic systems, turning them into a reliable tool for modern lexicography. All VLLs that have been created by the Fund are available on a corporate basis at www.ulif.org.ua. The similar projects are being carried out not only in Ukrainian but also foreign lexicography. For example, a virtual research environment for linguists VerbaAlpina providing the researchers with tools for collecting and analyzing linguistic data from linguistic atlases and dictionaries, as well as linguistic-geographical references collected over the past hundred years. One of the main functions of the VerbaAlpina research environment is lexicographic one. This allows researchers to study textual data in onomasiology and semasiology aspects [4].

Theoretical principles of creating virtual environments for language and literature study on the basis of texts are the subject of many researches [2-10]. The virtual environment is usually understood as a digital environment making available a set of research tools to the specialists who may stay in different geographic regions [2]. When building VLL, the following requirements [3] are to be taken into account:

1. *Parsing and indexing:* to get the information from a data collection, the data are to be extracted from the collection and then indexed (as each collection may differ in its approach to data coding, different approaches to data extraction, analysis and indexing should be applied).
2. *Information retrieval*: execution of the queries consisting of one word, collocations and phrases; the queries to be executed for any number of available and selected data collections.
3. *Tools for linguistic analysis*: the tools should be provided to facilitate analyzing linguistic facts from the samples obtained during the information retrieval process.

The VLL DLE 23 project which is the object of the present paper has been carried out in several stages, namely: 1) text extraction (partial) together with HTML markup from available online version of the dictionary; 2) dictionary text analysis to identify meta-language and information elements of the entries; 3) building the model of the lexicographic system (L-system), which in a formal way displays the information elements of DLE 23 and serves as a basis for creating the database and interface.

3. Methods and materials

The construction of the formal model of L-system for DLE 23 requires an appropriate theoretical framework. As such in our study we use the theory of L-systems by the Ukrainian Academician Volodymyr Shyrokov [2]. The main idea of the theory is that the dictionary (D) can be formally described as an L-system, i.e. an abstract linguistic information object containing a set of elementary information units $I^Q(D)$, abbreviated EIU, and a set of descriptions $V(I^Q(D))$ assigned to these units by a subject:

$$S: I^Q(D) \rightarrow V(I^Q(D)) \quad (1)$$

Elementary information units which are the objects of $V(I^Q(D))$ are the result of the lexicographic effect Q. In our study, we understand the lexicographic effect as a process and result of the generation EIU from the finite alphabet:

$$Q: I(D) \rightarrow I^Q(D) \quad (2)$$

The finite alphabet components are: 1) the Spanish alphabet (A ... Z, a ... z) including diacritical characters (ñ, ç; á, é, í, ó, ú, ü); 2) the Greek alphabet (α ... ω); 3) the international Latin alphabet; 4) dictionary metalanguage markers to indicate entry elements (Tb, Del., ♦, ■, (); 5) punctuation (“.”, “,”, “;”, “:”, “ ”), including even symbols ("¡!", "¿?"); 6) typefaces (normal, italic, bold). It should be noted here that EIUs are not only description objects of $V(I^Q(D))$, but also the formants of $V(I^Q(D))$. This corresponds to the condition of lexicographic closure of the dictionary: each unit found in the dictionary text must be also added to the headword list [].

Elementary information units to be grouped with each other within $V(I^Q(D))$ compose the text of the dictionary D in general and the text of dictionary entries $V(x)$ in particular:

$$V(x) = \pi_1 \xi_1^i \pi_2 \xi_2^i \pi_3 \dots \pi_n \xi_n^i, x \in I^Q(D), \quad (3)$$

where ξ_j^i denotes the word contained in the text of the dictionary entry $V(x)$ and included in the word list of the dictionary D; π_i means the words separators, such as punctuation marks, metalanguage markers and symbols. In turn, the entry text is decomposed into information elements that have certain metalanguage markers. This can be shown in a formal way such as:

$$V^{Word}(x) \equiv \{\beta_i(x) \mid i = 1,2, \dots n\}, \quad (4)$$

where $\beta_i(x)$ is a certain information element extracted from the dictionary text. Among $\beta_i(x)$ which form the set of descriptions $V(x)$ are: 1) head word (LEMA); 2) masculine, feminine form (GEN); 3) homonymy (HOM); 4) headword variants (VARLEMA); 5) etymology (ETIM); 6) morphological characteristics (MORF); 7) spelling peculiarities (ORT); 8) definition number (ACEP№); 9) grammar class (GRAM№), 10) set of remarks (NOTA№); 11) definition (DEF№) and 12) encyclopedic reference (ENCICL№). Additionally, there has been introduced INDEF element which indicates an unspecified entry element.

Let us designate by $\Lambda(D)^{Word}$ a set of descriptions covering the formal characteristics of Spanish words, which includes the following elements:

$$\Lambda(D)^{Word} \equiv \{\text{GEN}, \text{HOM}, \text{VARLEMA}, \text{ETIM}, \text{MORF}, \text{ORT}\} \quad (5)$$

Then $P(D)^{Word}$ will indicate a set of descriptions dedicated to the content characteristics of Spanish words, which is comprised by:

$$P(D)^{Word} \equiv \{\text{ACEP№}, \text{GRAM№}, \text{NOTA№}, \text{DEF№}, \text{ENCICL№}\}, \quad (6)$$

By $I^Q(D)^{ESP}$ we shall mean the set of Spanish language units described in (5) and (6). Additionally, we introduce the operators to establish the relationship of each unit with its formal (F) and semantic (C) descriptions. The correspondence between the two description types is provided by the operator H. Therefore, the formal model of the L-system for DLE 23 takes the following form:

$$LS^{ESP} \equiv \{I^Q(D)^{ESP}, \Lambda(D)^{Word}, P(D)^{Word}, F, C, H\} \quad (7)$$

At present, some of the above elements of the formal and semantic description are indexed in the lexicographic database and they are accessible for work through the interface VLL DLE 23. They contain relevant linguistic facts that can be easily revealed using the proposed tools. The next section provides detailed description of the techniques for working with these tools.

The dictionary text conversion into database format was performed in four stages: 1) extraction of the text from the online version of DLE 23 to avoid errors typical for optical recognizing the printed version of the dictionary; 2) identification of information elements of formal and semantic descriptions in the extracted text; 3) indexing the information elements of dictionary entries to be accessed through the interface; 4) conversion of dictionary text into database format.

4. Experiment

The main tools of VLL DLE 23 are “Sample” and “Statistics”. The first is used to form samples from dictionary entries, the head words of which correspond to certain linguistic parameters. The function of the second tool “Statistics” is to calculate quantitative parameters for the formed sample

(VLL generates statistics for the entire dictionary by default). The parameters for selecting the dictionary entries (Figure 1) in current version are as follows:

- Head word group: headword in masculine, headword in feminine, both
- Headword structure: word, phrase, morpheme
- Headword type: abbreviation, acronym, word of foreign origin
- Availability of homonymy, polysemy and phrases

Each parameter can be discarded (by ticking the checkbox "Ignore" in the "Sample" dialog box). For homonymy, polysemy and phraseology, quantitative parameters are provided: the number and conditions ("more", "less", "equal", "greater than or equal to", "less than or equal to") are set. The purpose of the full-text search in this case is very specific: to select the headwords by metalanguage elements. The search is performed on the entire dictionary text, marked according to the HTML5.0 standard. The search string of the text can include both text fragments of the dictionary entry and html tags.

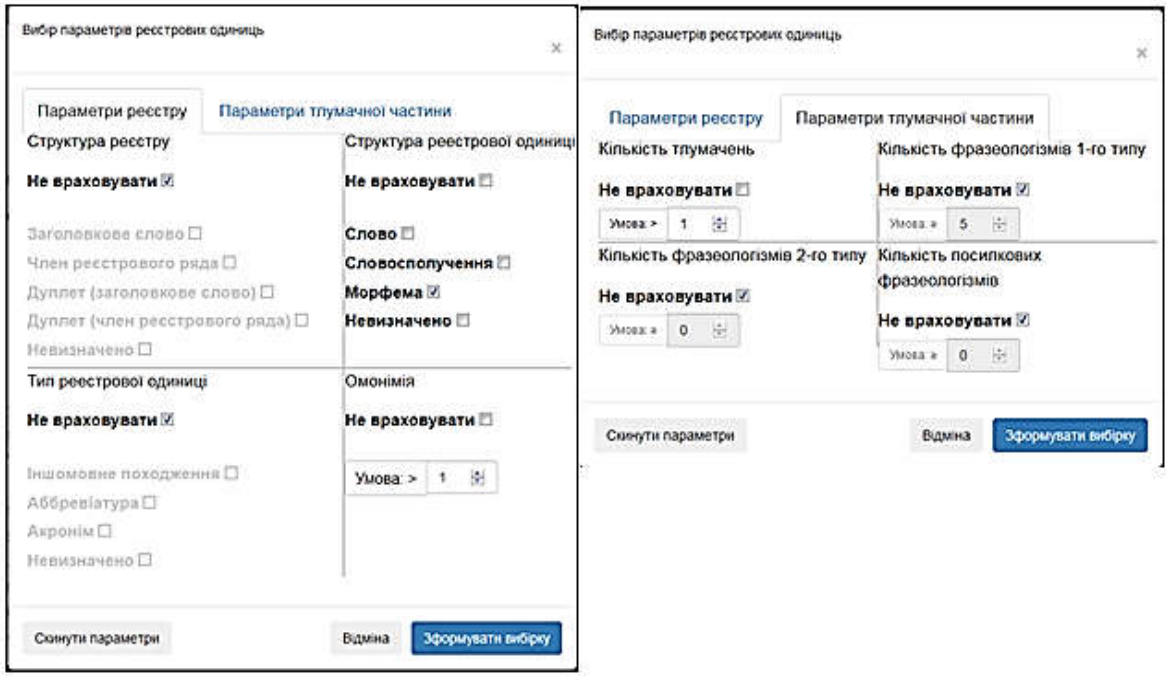

Figure 1: Dialog box "Sample" to select the entries by sample parameters

The interface language in the current version is Ukrainian. In final version of VLL DLE 23, the user will be able to choose between Ukrainian, Spanish and English. The dictionary notes and other metalanguage elements will be translated in the three languages. The VLL DLE 23 interface has been designed to perform linguistic researches based on the dictionary text. Based on these studies, the user can make certain conclusions about the lexical-semantic, etymological, grammatical and usage features of Spanish language units. We will consider below the types of linguistic research that are possible to be carried out on the basis of the dictionary text using the interface of the virtual lexicographic laboratory.

Spanish language composition. The user has the opportunity to explore the types and number of the words composing the dictionary list: morphemes (prefixes and affixes), words and phrases. In this case, the user can choose a certain type and structure of headwords by clicking appropriate checkbox. There's also additional parameter to include homonymic words in the sample.

Experiment 1. Let us make a selection of the dictionary entries containing homonymic morphemes (Figure 2). To obtain the sample, you must follow these steps:

1. Open the dialog box "Selection" from the main window of the VLL DLE 23
2. Tick the checkbox "Morpheme" in the tab "Headword parameters"
3. Set the homonymy parameter for "≥ 1" in the panel "Homonymy".

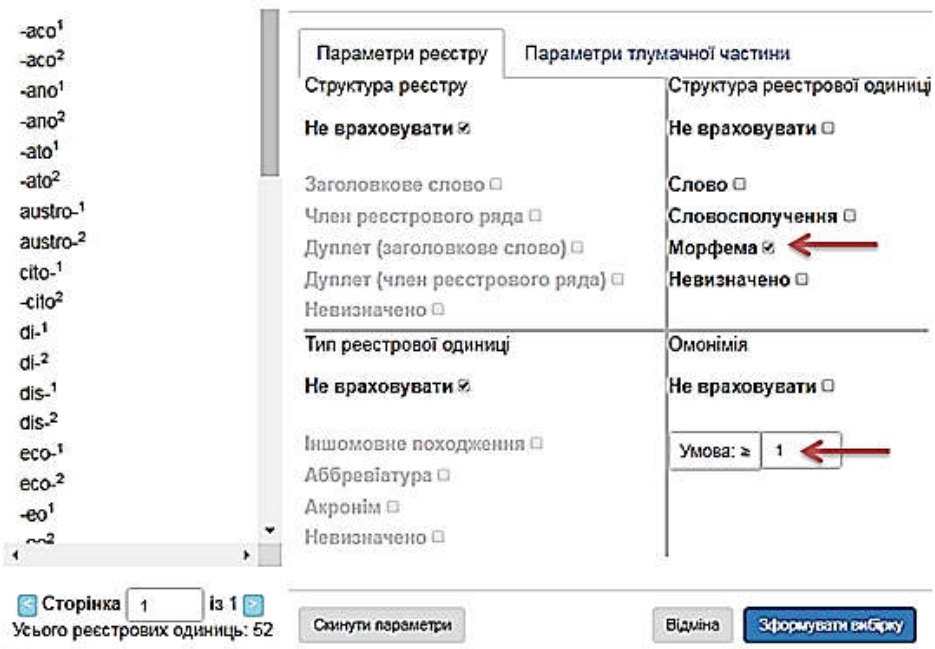

Figure 2: Sample of headwords (left) and chosen parameters (right)

Experiment 2. The next example of using VLL DLE 23 shows the possibility of selecting dictionary entries with Spanish words that have masculine and feminine forms (*anciano* "old man" (m.) – *anciana* "old woman" (f.), *duque* "duke" (m.) – *duqueza* "dutchess" (f.), *regulador* (m.) – *reguladora* (f.)). The sample of these words has been obtained by the following procedure:

1. Open the dialog box "Sample" from the VLL main window and activate the tab "Headword list options"
2. Select the options "Headword in feminine" and "Headword in masculine" in the panel "Headword group"
3. Select the option "Word" in the panel "Headword structure"

The sample of the entries corresponding to the parameters is shown in Figure 3. The sample is represented by the following parts of speech: nouns (*abuelo*, *-la*; *duque*, *-eza*; *perro*, *-ra*), adjectives and adjectival nouns (*ácido*, *-da*, *rojo*, *-ja*), articles (*un*, *una*; *el*, *la*) and pronouns (*él*, *ella*; *nosotros*, *-as*).

Figure 3: Sample of headwords (left) and chosen parameters (right)

Morphology. The study of morphological features of the Spanish language involves activation of dictionary information elements both in the tabs “Headword list parameters” and “Explanatory part”. As an example, let us investigate the peculiarities of forming of derived words lexical meaning of which is inherited from another word they originate from. The entries describing these words in DLE 23 have specific definition patterns like: 1) aum. de + X (aum. from X); 2) dim. de + X (dim. from X); 3) adj. sup. de + X (highest degree from approx. X); 4) Acción de + X (action from X); 5) Action and effect of + X (action and result from X); 6) De manera + X (X + way, for example: “cunning way”, “fast way”); 7) Con (de modo) + X (C + X, for example: “with honor”, “with ease”); 8) “Que + X” (which + X); 9) “Persona que + X” (person who + X); 10) Que siente + X (“who has an emotional state”); 11) Que tiene + X (“which contains something”); 12) Ponerse + X (“acquire characteristics X”).

Experiment 3. Let’s analyze the use of suffixes with which to form the verbal nouns denoting an action and its result. Indicative of these words are monosemic and retain the lexical meaning of the original word. In DLE 23 such nouns have the definition pattern *Acción y efecto de + verb*. To get a sample of these words, the followings steps are to be made:

1. Open the dialog box “Sample” from the main window of the VLL and open the tab “Headword list options”
2. Set the option “Word” in the group “Headword structure”
3. Open the tab “Explanatory part” and set the condition “=” and the value “1” for the parameter “Number of interpretations”
4. Select “Full-text search” on the search bar in the main window of VLL DLE 23
5. Insert the text fragment (along with HTML tags) into the search field: “Acción </mark><mark data-id="c8HoARq|c8HrfrV|c8IFPyp">y</mark><mark data-id="EOoHYxJ">efecto</mark> <mark data-id="BtDkacL|BtFYznp">de</mark>”

The result of VLL DLE 23, as well as the sample parameters selected (explanatory part) are shown in the figure 4.

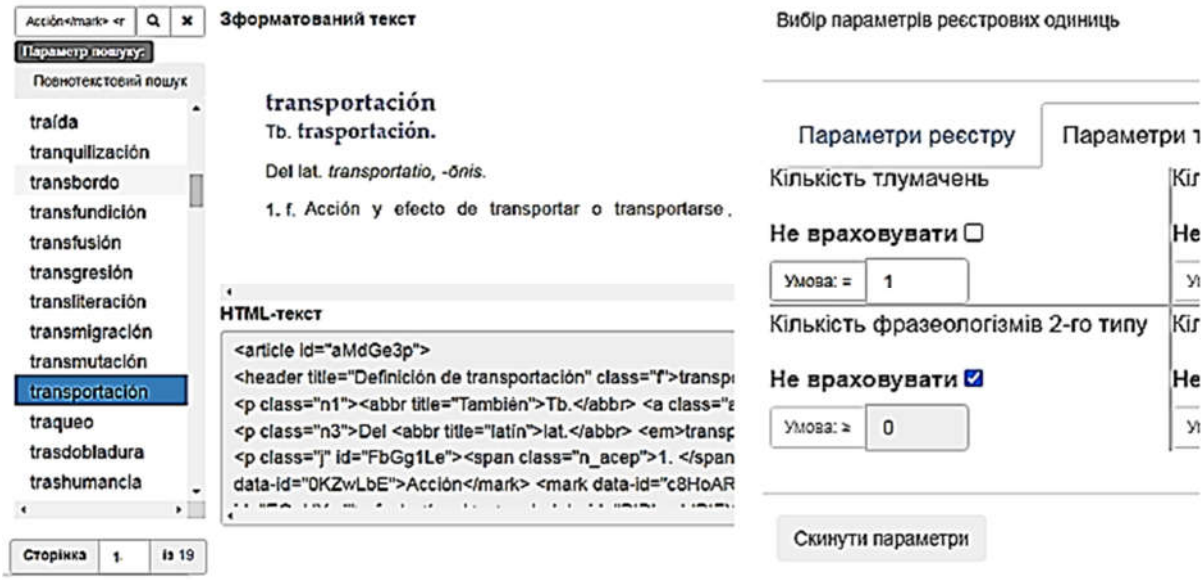

Figure 4: Sample of headwords (left) and chosen parameters (right)

There has been a sample of 2708 dictionary entries describing verb nouns. The suffixes used to form these nouns are given in the Table 1.

Table 1

Suffixes used to form verbal nouns with lexical meanings “*Acción y efecto de + verb*”

Suffix	Verbal nouns	Verbs
-ado	*ajardinado, alisado, trefilado*	*ajardinar, aliszar, trefilar*
-anza	*antojanza, habitanza*	*antojarse, habitar*
-aje	*cronometraje, desmontaje*	*cronometrar, desmontar*

-ción	*abstracción, abolición, actualizacón*	*abstraer, abstraerse, actualizr*
-dura	*carenadura, trastornadura*	*carenar, trastornar*
-eo	*abejoneo, baboseo, cañoneo*	*abejonear, abosear, cañonear*
-miento	*abarcamiento, acotamiento*	*abarcar, acotar*
-que	*trastrueque*	*trastrocar*

Lexical semantics. In this regard, the current version of VLL DLE 23 makes it possible to form the samples of dictionary entries the head words of which have common meaning. These words are characterized by common seme attributable to a particular lexical-semantic class or thematic group, but at the same time one or more differential semes by which the words differ from each other. Differential semes are identified during component analysis of headword definitions after sampling.

Experiment 4. For example, we have to make a sample of dictionary entries the head words of which in Spanish denote the types of ships. Such words in DLE 23 are defined through generic definitions beginning with the word *embarcación*. To select them:

1. Select “Full-text search” on the search bar in the main window of VLL DLE 23
2. Insert the text fragment (along with HTML tags) into the search field: “<mark data-id="EajnWJX">Embarcación</mark>”

Figure 5: Sample of headwords composing lexical-semantic class *embarcación* (left) and headword description (right)

As a result, there has been formed a sample of 371 head words, of which only 115 meet the specified query. The result of the query execution is shown in the Figure 5. The obtained words to denote the types of ships / ships, and their meanings are reflected by the following semes:

- Purpose: *para transporte* (for transport); *para pescar* (for fishing), *costear o atravesar los ríos* (to coast and cross the rivers); *para servicios auxiliares* (for auxiliary service); *destinada a explorar las profundidades del mar* (designed to explore the sea depths); *puesquera* (fishing);
- Components: *con cubierta y un solo palo* (with deck and single mast); *sin cofas y de dos palos* (without tops and with two masts);
- Geometric characteristics: *de estructura cóncava* (concave structure); *en forma de artesa* (in the form of a trough); *de fondo chato* (with flat bottom); *de fondo plano* (with flat bottom);
- Area of use: *usada en Filipinas* (used in the Philippines); *usada en el Mediterráneo* (used in the Mediterranean); *se usaba en las Indias orientales* (used in the East Indies);
- Material: *hecha de haces de totora* (made of bundles of totora);

- Features of functioning / use: *hinchable* (inflatable); *sumergible* (submersible);
- Time period of use: *usada en la Edad Media* (used in Middle Ages);
- Method of actuation: *cuatro remos por banda* (four oars per band); *velas* (sails); *vela y remo* (sail and oar);
- Dimensions: *grande* (big); *pequeño* (small); *menor* (smaller); *muy pequeña* (very small).

It should be noted that the sample also contains irrelevant headwords denoting: 1) actions related to ships (*abarrancar, abatir, abordar, embarcar*); 2) auxiliary elements to operate ships (*amarra, ancla*); 3) service personnel (*amarrador, arraez, velero*); 4) parts of ships (*proa, remo, vela*), etc. This is due to certain limitations in the functionality of the current version of VLL DLE 23.

Syntax (*collocations*). The subject of DLE 23 description also covers the collocations, idioms and clichés. Their lexicographic description is based on the same parameters as one-component units. There are two types of collocations in the dictionary:

- "Noun + adjective" (as an adjective can be a prepositional construction with *de*)
- Phrases formed with other parts of speech (verbal, prepositional, adverbial, etc.)

Experiment 5. Let's make a sample of the entries the head words of which form the collocations with the prepositional construction "de crédito". In the dialog box "Sample":

1. Choose the option "Word" in the panel "Head word structure"
2. Set the condition "≥" and the value "1" in the tab "Explanatory part", panel "Number of collocations *Noun + Adjective*"
3. Choose the search mode for "Full-text search" on the search bar in the main window VLL DLE 23
4. Insert the query in the search field as a piece of text (along with HTML tags): "de crédito"

As can be seen in the Figure 6, such words are *carta* (*carta de crédito*), *cuenta* (*cuenta de crédito*), *riesgo* (*riesgo de crédito*), *título* (*títulos de crédito*), *transferencia* (*transferencia de crédito*). However, the possibilities to study combination characteristics of Spanish words by VLL are quite limited due to the lack of appropriate means to clarify the user query (for example, search only in the area of collocations). Therefore, the sample may include irrelevant head words.

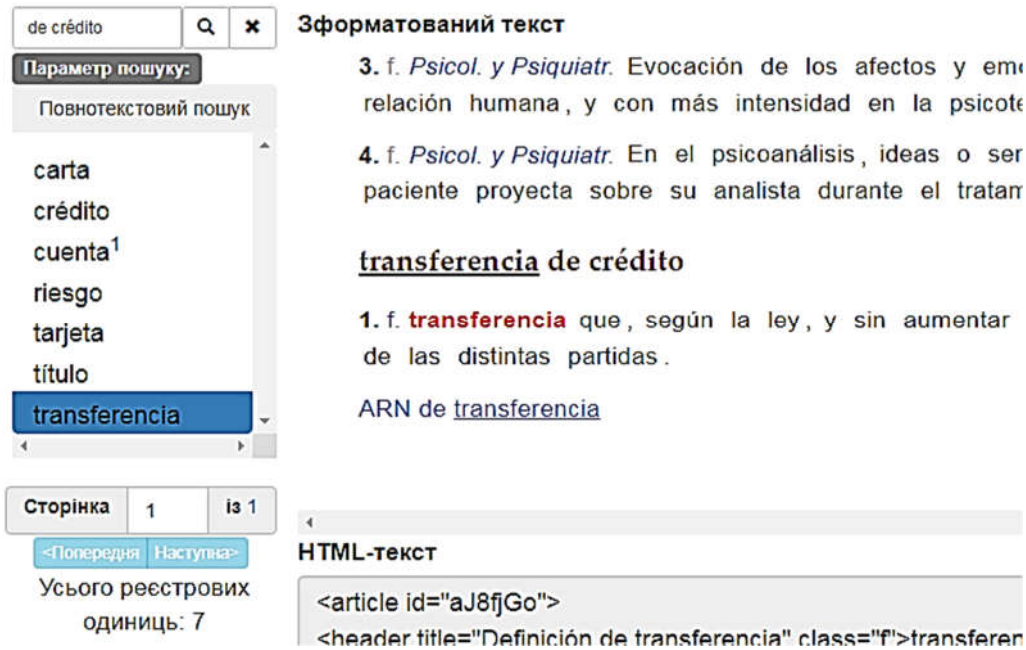

Figure 6: Sample of headwords forming the collocations *noun + de crédito*

5. Results

At the current elaboration stage, VLL allows studying the vocabulary of the Spanish language in terms of its content and classification, as well as some morphological, lexical-semantic and combinatorial characteristics of the words described in DLE 23. Additionally, VLL tools can be used

while performing statistical research, creating sub-dictionaries (for example, of foreign words, of morphemes, of collocations etc.). To compare VLL DLE 23 research potential with similar resources we would like to consider also the functionality of the online Oxford Dictionary (www.oed.com), which provides the ability to form a sample of English words by the following metalanguage parameters:

1. Grammar class (noun, adjective, verb, etc.)
2. Categories: subject area (art, law, language, philosophy, etc.), use (archaic, ironic, regional, poetic, etc.), region (Africa, India, Britain, etc.)
3. Languages of origin (African languages, English, European languages)
4. Chronology, i.e. when the head word was first fixed in the OED at different times
5. Source (a table of the most popular authors and works, citations of which are in the OED).

The comparative interface characteristics of the VLE DLE 23, online version of DLE 23 and the online Oxford English Dictionary are shown in Table 2.

Table 2
Comparative characteristics of different interfaces

Characteristics	DLE 23 online	VLL DLE 23	OED online
Inventory of language units	Yes	Yes	Yes
Sampling by thematic categories	No	No	Yes
Sampling by part of speech	No	No	Yes
Sampling by keyword in the definition	No	Partially	Yes
Sampling by language of origin	No	No	Yes
Sampling by stylistic characteristics	No	No	Yes
Sampling by type of headword unit	No	Yes	No
Sampling by language unit structure	No	Yes	No
Statistical calculations	No	Yes	No
Generating sub-dictionaries	No	Yes	Yes

The table shows that, despite its current capabilities, VLL DLE 23 tools need further improvement. In particular, this applies to indexing other information elements of the dictionary (Table 3), which have been identified in course of lexicographic analysis. There is also a need to provide other parameters to form the samples, such as “geographical region”, “attributed to the style of language”, “subject area” and others.

Table 3
Dictionary information elements

Information element	Information	Access available
LEMA	Head word	Yes
GENERO	Head word in masculine and feminine	Yes
HOM	Homonymy	Yes
VARLEMA	Headword variants	No
ETIM	Etymology reference	No
MORFO	Word flection peculiarities	No
ORTOGR	Spelling characteristics	No
INDEF	Undefined element	Yes
ACEPN№	Definition number	Yes
GRAMN№	Grammar class corresponding to lexical meaning	No
NOTAN№	Category notes	No
DEFN№	Definition	No
ENCICLN№	Encyclopedic note	No
FRASNA	Collocations “Noun +Adjective”	Yes
FRASOT	Collocations of other types	Yes

6. Discussions

The works on expanding the research potential of VLL DLE 23 are underway. The rest of the entry information elements listed above in the Table 3 are planned to be indexed soon. To work with the dictionary text in digital environment, it is necessary all its information elements that may be of interest to the linguist during the research need being made accessible. This will ensure executing queries more accurate and getting the samples without irrelevant head words. In particular, we are talking about samples formed by the tool "Full-text search". In addition, a full-text search should be not only applied for the entire dictionary text, but also for its individual text elements. Using the parameters listed in the Table 4, it will be possible to study the variability of the of head word forms, the origin of the head words, their semantic structure, etc.

Table 4

Sample parameters to be provided in VLL DLE 23

Information element	Sample parameters	Parameter value
VARLEMA	Regional variant	Variant form
	Availability of detailed information about regional variant	Yes / No
	Text fragment	Specified by the user
ETIM	Language of origin	Language name
	Availability of additional information about etymon	Yes / No
	Text fragment	Specified by the user
GRAMN№	Grammar class including grammar category	Grammar note
NOTAN№	Domain, region, language style, chronological status	Category note
DEFN№	Definition type	Meta-language marker
	Availability of a key word in definition	Specified by the user
	Text fragment	Specified by the user
ENCICLN№	Availability of encyclopedic information	Yes / No
	Type of encyclopedic information	Formula / Symbol

Thus, the expanded version of VLL DLE 23 interface will contribute to a wider range of linguistic researches based on dictionary text. On their basis the user will be able to draw certain conclusions about lexical, semantic, etymological, grammatical and usage features of Spanish language units. The possibility of using full-text search to work with these information elements individually will be also provided for.

7. Conclusions

The virtual lexicographic laboratory VLL DLE 23 which has been elaborated in at Ukrainian Language Information Fund enhances in effective way linguistic researches based on the dictionary text. In this context, the dictionary text can't be regarded as a means providing the reference about specific word or collocation but a tool assisted basis for researching the regularities in the language system. This mainly concerns fundamental explanatory dictionaries providing detailed and exhaustive description of language. At the current stage of development, VLL DLE 23 makes possible to derive linguistic data from the DLE 23 text while conducting linguistic researches of the Spanish language.

With current stage of VLL development, allows to conduct research on the vocabulary of the Spanish language in terms of its content and classification, as well as some morphological and lexical-semantic, and combinatorial features of words that are the subject of DLE 23. The list of VLL-assisted linguistic researches include:

1. *Spanish language vocabulary*: types of language units, their homonymy, the correlation between nationally-specific and loan vocabulary, structural types of language units (morphemes, words, phrases, abbreviations). In future, the user will be able to create samples of register words

based on their use in a particular field, geographical area of use, language style and chronological status

2. *Studying morphological features* of the Spanish language, in particular the participation of affixes in building certain word forms (femininities, names of professions, derived words, etc.), derivational structure of the words, models of verb inflection, etc.

3. *Analysis of the words that have common lexical meaning*, making a sample representing the words of a certain lexical-semantic class. The current VLL version still has limited capacity to analyze lexical meanings reflected in interpretations, as a result of which irrelevant words may be included in the sample

4. *Combinatorial features of head words*, i.e. the ability of Spanish words to form collocations. Sampling inaccuracies arise due to the lack of appropriate means to clarify the user query (for example, search only in the area of phraseology). At the moment, the project of VLL DLE 23 is underway to improve its functionality.

8. References

[1] V. Shyrokov, Computer linguistic studies: Proceedings of the Ukrainian Lingua-Information Fund NAS of Ukraine, vol. 1: Research paradigm and basic language information structures, Ukrainian Lingua-Information Fund, Kyiv, 2018.

[2] R. Stanković, R. Stijović, V., Duško, C., Krstev, O. Sabo, The Dictionary of the Serbian Academy: from the Text to the Lexical Database, in: S. Krek, C. Jaka, V. Gorjanc (Eds.), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 2018, pp. 941–949.

[3] M. S. Sarwar, T. Doherty, J. Watt, R. O. Sinnott, Towards a virtual research environment for language and literature researchers, Future Generation Computer Systems, 29(2) (2013) 549–559. doi: 10.1016/j.future.2012.03.015.

[4] C. Mutter, A. Wiatr, The Virtual Research Environment of VerbaAlpina and its Lexicographic Function, in: S. Krek, C. Jaka, V. Gorjanc (Eds.), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 2018, pp. 775–785.

[5] Ch. Ore, O. Grønvik, Studying language change through indexed and interlinked dictionaries, in: Z. Gavriilidou, M. Mitsiaki, A. Fliatouras (Eds.), Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion, volume 1, Democritus University of Thrace, Greece, 2020, pp. 321-330.

[6] V. Giouli, N. Sidiropoulos, Making Dictionaries Visible, Accessible, and Reusable: The Case of the Greek Conceptual Dictionary API, in: Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion, volume 1, Democritus University of Thrace, Greece, 2020, pp. 91–100.

[7] C. Marini, E. Jezek, CROATPAS: A Resource of Corpus-derived Typed Predicate Argument Structures for Croatian, in: Proceedings of the 6th Italian Conference on Computational Linguistics (CLiC-it), Bari, Italy, 2019. URL: http://ceur-ws.org/Vol-2481/paper42.pdf

[8] A. Tavast, M. Langemets, J. Kallas, K. Koppel, Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX, in: S. Krek, C. Jaka, V. Gorjanc (Eds.), Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, 2018, pp. 749–761.

[9] M. José Domínguez Vázquez, D. Bardanca Outeiriño, A. Simões, Automatic Lexicographic Content Creation: Automating Multilingual Resources Development for Lexicographers, in: Post-Editing Lexicography – Elex 2021, Brno, Czech, 2021, pp. 269–287.

[10] G. D. S. Anderson, A. Luisa Daigneault, Living Dictionaries: An Electronic Lexicography Tool for Community Activist, in: Post-Editing Lexicography – Elex 2021, Brno, Czech, 2021, pp. 339–360.