

Digital Toolkit to Develop Research Potential of Explanatory Dictionary (Case of Spanish Language Dictionary)

Yevhen Kupriianov^a, Volodymyr Shyrovkov^b, Iryna Ostapova^b and Mykyta Yablochkov^b

^a National Technical University “Kharkiv Polytechnic Institute”, Kyrpychova str. 2, Kharkiv, 61002, Ukraine

^b Ukrainian Lingua-Information Fund NAS of Ukraine, Holosiivskiy av. 3, Kyiv, 03039, Ukraine

Abstract

Nowadays linguistic corpora are recognized as a most effective tool to perform linguistic researches in digital environment. However, the dictionaries that actively use corpus technologies for their creation and update remain underestimated in regards to their research potential. Fundamental explanatory dictionaries of national languages are of primary interest for linguistic experts. The dictionaries of this kind are characterized by giving complete well-structured and multi-aspect description of language units, having linguistic theories as a basis for creation and by representing all the linguistic information necessary not only for understanding the meanings of language units in various contexts, but also for their correct use. The present paper describes the project of software toolkit for extracting linguistic information from dictionary text. The authors share their experience gained while creating such kind of research tool and show its advantages for professional linguists. The software project is being carried out for working with Spanish Dictionary “Diccionario de la lengua española. 23^a edición” (DLE 23). The entry texts have been taken from DLE 23 online version (www.dle.rae.es). The dictionary is characterized by detailed description of morphological, stylistic, prosodic, syntactic and combinatorial features of Spanish lexical units. The headword list also includes morphemes, phrases of various types, acronyms and abbreviations. The project in question involves the creation of the virtual lexicographic laboratory (VLL DLE 23) intended for linguistic researches on the basis of DLE 23 text. The theoretical framework of the project consists of the theory of lexicographic systems and theory of semantic states. The examples of applying the current version of VLL as a tool for linguistic research are given.

Keywords 1

Computer lexicography, linguistic information extraction, virtual lexicographic laboratory, explanatory dictionary, digital environment

1. Introduction

One of the present-day tasks of the modern lexicography is to find various ways of using rich potential of digital environment to timely satisfy the information needs of advanced users and modern lexicographers. The up-to-date dictionary making relies on digital linguistic technologies. First of all, we refer to corpus technologies (Corpus Query Systems or CQS) and digital systems to compile and update dictionaries (DWS short for Dictionary Writing Systems). It should be also noted that dictionary-making process involves IT specialists who support and develop digital technologies in linguistics, which is a new challenge for lexicography. Despite major advances in digital technologies,

COLINS-2021: 5th International Conference on Computational Linguistics and Intelligent Systems, April 22–23, 2021, Kharkiv, Ukraine
EMAIL: eugenikupriianov@gmail.com (Y. Kupriianov); vshyrovkov48@gmail.com (V. Shyrovkov); irynaostapova@gmail.com (I. Ostapova); gezartos@gmail.com (M. Yablochkov)
ORCID: 0000-0002-0801-1789 (Y. Kupriianov); 0000-0001-5563-8907 (V. Shyrovkov); 0000-0001-8221-3277 (I. Ostapova); 0000-0003-1175-1603 (M. Yablochkov)



© 2021 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

the lexicographic landscape remains largely heterogeneous. This applies to both the formats of lexicographic data representation, and the standards for working with them [7].

Our interest is focused primarily on comprehensive explanatory dictionaries of national languages. Using CQS and DWS technologies allow non-stop work, i.e. the dictionary-making process is always in progress without completion stage (as in case of Oxford English Dictionary). However, despite the availability of advanced user interfaces, their possibilities for searching, analysis and generalization of linguistic information, primarily for professional linguists are still limited. The authors have been traditionally those who develop not only the structure and content of the entries but the search capabilities of the dictionary. As a result, the problem of extracting linguistic information for its further usage by the experts in their researches is still not resolved. Therefore, the goal of our research work is the development of an interface scheme to conduct linguistic researches on the basis of explanatory dictionary text and the construction of an effective toolkit that implements this scheme. Inspirational is the fact that, unlike paper dictionaries, this is a feasible task for digital lexicographic text [1, 2, 3, 6, 7, 8].

For the purposes of our research we have selected Spanish Language Dictionary entitled “Diccionario de la lengua española. 23ª edición” (shortly DLE 23), which has been published by the Academia Real Española (Spanish Royal Academy). The DLE 23 is the most comprehensive and representative explanatory dictionary of the Spanish language. The 23rd edition was published in October 2014. The year later DLE 23 was made available on CD-ROM and then online at www.dle.rae.es. Now the Academy is working on a 24th edition, which is supposed to be digital only [5].

2. Spanish language dictionary

The DLE 23 is characterized by detailed description of morphological, stylistic, prosodic, syntactic and combinatorial features of Spanish lexical units. The headword list also includes morphemes, phrases of various types, acronyms and abbreviations. The entries contain multi-aspect information which facilitates not only the meaning of a headword in different contexts but also correct usage in communication. The main factor which has determined our choice of the dictionary is the availability of the dictionary text in electronic form in HTML format, which guarantees the authenticity of the text with its paper version and excludes orthographic errors that are typical for OCR. Moreover, the tags allow identification of the information elements of a dictionary entry. Currently a prototype version of VLL DLE 23 which can be accessed at <https://services.ulif.org.ua:44359/>, enlarges research potential of DLE 23 in greater extent.

2.1. General characteristics

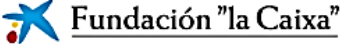
The formal model of a lexicographic system isn't possible to be built without having large and comprehensive dictionary as a basis. In our case we have selected Spanish language dictionary “Diccionario de la lengua española 23ª ed. Edición del tricentenario” (shortly DLE 23) by the Royal Spanish Academy. This dictionary is a fundamental work containing vocabulary to be widely used both in Spain and Latin America. Besides lexical meanings, DLE 23 also provides detailed information on grammar, syntax and usage features of the words composing the headword list.

The headword list of DLE 23 comprises more than 93,000 units representing morphological, lexical and syntactic levels of the Spanish language. The total number of definitions is 195,439. If compared with the previous edition [4], DLE 23 has:


- 21,466 meanings corresponding to different domains,
- 18,712 meanings peculiar to Latin America,
- 435 meanings related to the usage in Spain,
- 333 foreign words not adapted to Spanish,
- 1,637 verbs together with their conjugation models.

2.2. Interface of online version of DLE 23

The current online version of DLE 23 is intended for providing a reference on word semantics, but unfortunately has very limited research potential. The interface consists of a list of filters, a search box and a “Search” button (consultar). The proposed interface allows you to work only with the dictionary register with a few filters: “word form” (por palabras), “lemma” (lema), “contains” (contiene), “exactly” (exacta), “begins with” (empieza por), and “ends with” (termina en).

Consulta posible gracias al compromiso con la cultura de la 

por palabras

agua 

Del lat. *aqua*.

1. f. Líquido transparente, incoloro, inodoro e insípido en estado puro, cuyas moléculas están formadas por dos átomos de hidrógeno y uno de oxígeno, y que constituye el componente más abundante de la superficie terrestre y el mayoritario de todos los organismos vivos. (Fórm. H₂O).
2. f. Líquido que se obtiene por infusión, disolución o emulsión de flores, plantas o frutos, empleado como refresco o en medicina y perfumería. *Agua de azahar, de cebada, de limón*.
3. f. **lluvia** (ll acción de llover). U. t. en pl. con el mismo significado que en sing.
4. f. **lágrimas** (ll gotas de la glándula lagrimal). *Se le llenaron los ojos de agua*. U. t. en pl. con el mismo significado que en sing.
5. f. Vertiente de un tejado. *Una cubierta a dos aguas*.

Figure 1: General view of DLE 23 online version.

Linguistic research requires the access to the entire text of the dictionary, as well as to its separate elements. This requires a theoretical basis for identifying, describing and representing relevant linguistic data from the DLE 23 text.

2.3. Lexicographic analysis

Each fragment of the entry text corresponds to a certain type of linguistic information and can be identified by the format of representation. This format can be well-defined or undefined at all. Let us consider the ways of representing the information in different parts of DLE 23 entry such as headword, headword variants, etymology, morphology, orthography, set of definitions and encyclopedic note.

2.3.1. Entry information elements

In paper version the elements have a linear order and special characters are used to separate them in the text array. In online version each element is located in a separate text line and is highlighted not only with a special marker, but also with color, as shown in Table 1.

Table 1
Information elements and their identification in dictionary text

No	Element	Marker	Type	Color
1	Lemma			
1.1.a	Word	None	Bold / bold italic	Navy blue
1.1.b	“Noun + Adjective” colloc.	None	Bold	Dark brown
1.1.c	Collocation of other type	None	Bold	Light brown
1.2	Homonymy	Upper index	Bold	Navy blue
2	Lemma forms (for words)	None	Bold	Navy blue
2.1	Masculine form	Word form	Bold	Navy blue
2.2	Feminine form	Flection	Bold	Navy blue
3	Lemma variants	“Tb.”	Normal	Light blue
3.1	Variant	None	Bold	Navy blue
3.2	Additional data	None	Normal	Light blue
4	Eymology	“Del” or “Quizá del”	Normal	Green
4.1	Language of etymon	None	Normal	Green
4.2	Etymon	None	Italic for Latin etymons and normal for others	Green
4.3	Additional data	None	Normal	Green
5	Morphology	None	Normal	Light blue
6	Orthography	“Escr.”	Normal	Green
6.1	Orthographic feature	None	Normal	Light blue
6.2	Definition number	“en acep.”	Normal	Light blue
7.	Set of definitions	First definition No “1”	Bold	Black
7.1	Set of labels	None	Normal or italic	Light blue
7.2	Definition	None	Normal	Black
7.2.2	Usage example	None	Italic	Violet
7.2.3	Additional comments	None	Normal	Light blue
7.2.4	Encyclopedic note	“(“ and “)”	Normal	Black

2.3.2. Linguistic information overview

Each fragment of the entry text corresponds to a certain type of linguistic information and can be identified by the format of representation. This format can be well-defined or undefined at all. Let us consider the ways of representing the information in different parts of DLE 23 entry such as headword, headword variants, etymology, morphology, orthography, set of definitions and encyclopedic note.

The entry can be headed not only by a word, but also word-forming elements such as prefixes, suffixes, as well as idiomatic and non-idiomatic collocations. This entry element contains the following linguistic information for the headword:

- Headword structure: morpheme (-**acro**, **andro**-), word (**leche**, **pan**, **yerba**) or collocation (**agua mineromedicinal**, **como agua para chocolate**);
- Headword type: Spanish word (**cama**, **ojo**, **perro**), foreign word (*amateur*, *ballet*), abbreviation (**ADSL**, **ONG**), acronym (**hidrosol**, **laser**);
- Homonymy (**abalear**¹, **abalear**²).

The headword variants are given for all lexical words, such as nouns, adjectives, adverbs and verbs, including passive participles, and sometimes for grammar words such as articles and interjections. Some variants are provided with other details, namely:

- Geographical area, if the variant usage is limited to particular country or countries;
- Definition number if the headword variant relates only to particular lexical meaning (as it shown in table 2);
- Chronological status indicating that the usage of headword variant is archaic.

The format of this entry part is as follows: (1) headword variant; (2) additional information. The examples of headword variant description are given in Table 2.

Table 2

Headword variant description

No	Headword	Variant	Additional Information
1	sustancia	substancia	-//-
2	jiennense	jienense	giennense
3	enhorabuena	en hora buena	en aceps. 2-4
4	chabola	chavola	p. us.
5	yerbatero	hierbatero	en acep. 2, <i>Col., Ec., Méx. y Perú</i> ; en acep. 4, <i>Chile</i> .

As the table shows the word *sustancia* (substance) has its variant *substancia* and the absence of additional information means that headword and the variant are fully interchangeable. The same can be said about the word *jiennense* (from Jaén city) which can be interchangeable with *jienense* and *giennense*. In some cases there can be usage limits. For example, the usage of *en hora buena* (congratulations!) is limited to lexical meanings described in definitions 2-3. In fourth example the label “p. us.” (from Spanish *poco usado*) shows that *chavola* (cabin) is archaic variant of the headword *chabola*. The fifth example shows the geographical and usage limits for the variant *hierbatero*: in the meaning 2 only in Columbia, Ecuador, Mexico and Peru; and in the meaning 4 only in Chile.

The etymological part of the entry gives brief information about headword origin and is characterized by the following format: (1) the source language; (2) the etymon; (3) and additional information, which may include the semantic changes in etymons, structural changes, as well as the moment from which the word began used in Spanish. The content examples of the etymological part are given in Table 3.

Table 3

Headword etymology description

No	Headword	Language	Etymon	Additional information
1	diccionario	b. lat.	<i>dictionarium</i>	-//-
2	frente	ant.	<i>frunte</i>	y este del lat. <i>frons, frontis</i>
3	cigoto	gr.	ζυγωτός	zygōtós 'uncido, unido', der. de ζυγοῦν zygoûn 'uncir, unir'
4	ubicuidad	lat. tardío	<i>ubiquitas, -ātis</i> ,	y este der. del lat. <i>ubique</i> 'en todas partes'
5	herciano		H. R. <i>Hertz</i> ,	1857-1894, físico alemán
6	bikini	ingl.	<i>bikini</i> ,	y este de <i>Bikini</i> , nombre de un atolón de las Islas Marshall, con infl. de <i>bi-</i> 'bi-', por alus. a las dos piezas

Etymological information can be concise (1), i.e. indicate only the language of origin and etymon, or more detailed (2-4). For example, *ubicuidad* comes from Late Latin word *ubiquitas*, and the letter has been derived from Latin *ubique* “everywhere”. If the word comes from a proper or geographical name (5-6) the information can be of encyclopedic type. In case of *bikini* etymology says that the word has English origin and comes from geographic name Bikini, an atoll of Marshall Isles; morpheme *bi-* having the meaning “composed of two parts”.

The next part of DLE 23 entry is the information about morphological features such as: regular and irregular forms of superlative degree of comparison for adjectives and adverbs; references to conjugation patterns for regular and irregular verbs, as well as irregular passive participles for individual verbs, etc. The examples of morphological information are given in Table 4.

Table 4
Morphological characteristics description

Headword	Morphological characteristics
fuerte	Sup. irreg. fortísimo ; reg. fuertísimo
hacer	V. conjug. en APÉNDICE; part. irreg. hecho
el, la	Neutro lo ♦ Pl. los, las

This part of the entry has neither a special identification marker nor defined format for representing linguistic information. So, the identifier may be its position in the sequence of the entry elements. In any case morphological characteristics go after etymology.

Orthographic information is provided only for headwords, the spelling of which (with a capital or small letter, with or without an accent) can significantly change their lexical meaning. This entry element includes may include the following information: spelling feature and the number of the lexical meaning in the dictionary to which this feature applies (see Table 5).

Table 5
Orthographic features description

Headword	Orthographic features	Lexical meaning to which the feature is applied
inmaculado, da	Escr. con may. inicial	en acep. 2
que	Escr. con acento	en acep. 3
donde	Puede escribirse con acento	en acep. 8

For example, the Spanish word *inmaculada* can have different meanings depending on its initial letter. It means “perfect, faultless” with small initial letter and “Mary, mother of Jesus” with capital letter.

The set of definitions represents the interpretation of the headwords using definitions of different types (standard, contextual, explanatory, by synonym, explanatory and others) and may consist of one or more definitions. Each definition is composed by: 1) introductory part, 2) definition text, 3) usage examples, 4) additional comments on lemma usage and 5) encyclopedic note. The introductory part is used for introducing a definition using keywords corresponding to its type. For example, “Dicho de”, “En” and “Entre” are the keywords for contextual definition, and “U.” for explanatory definition.

There is no introductory part for standard, synonymous and other definitions. The definition text can be a sentence or one word, a phrase, as in the case of a definition by synonym. Usage examples are complementary means of lexical meaning explanation and show headword usage in collocations or in a sentence. The definitions examples are followed by comments to denote additional grammar and usage peculiarities the headword may have in the lexical meaning. Let us give the content examples of lexicographic meaning description for the headword *agua* (water).

Table 6
Elements of lexicographic meaning description

Introductory part	Definition	Usage examples	Additional comment
∅	Líquido que se obtiene [...]	<i>Agua de azahar, de cebada, de limón</i>	∅
∅	lluvia (acción de llover)	∅	U. t. en pl. con el mismo significado que en sing.
∅	lágrimas (gotas de la glándula lagrimal)	<i>Se le llenaron los ojos de agua</i>	U. t. en pl. con el mismo significado que en sing.
U.	para avisar de la presencia de cualquier tipo de autoridad.	∅	∅

The last part of a definition is encyclopedic note, which is provided for the headwords denoting the concepts from natural sciences such as chemistry, physics, and mathematics. This note is a non-verbal way of representing a concept. For example, if the headword denotes chemical substances or elements, then the corresponding formula is shown in parentheses at the end of the definition. When it comes to mathematical or physical quantities, linguistic signs, their symbolic designations are presented. Encyclopedic note in DLE 23 is of two types: 1) “Fórm”, chemical formula, and 2) “Símb”, a symbolic designation of physical or mathematical quantities. The content of encyclopedic note for the headwords *agua*, *hercio* and *kilobyte* and *número pi* is shown in Table 7.

Table 7
Encyclopedic notes

Headword	Type of encyclopedic note	Content of encyclopedic note
agua	Fórm.	H ₂ O
hercio	Símb.	Hz
kilobyte	Símb.	kB
número pi	Símb.	π

As it can be seen from the above, every element of the dictionary entry contains multi-aspect information about Spanish language unit. Describing a language as an established system is illustrative of fundamental dictionaries, especially explanatory ones. It means that these dictionaries, as stated by Prof. V. A. Shirokov, carry a huge number of implicitly given relationships in a language system that cannot be revealed using traditional methods. In this regard, there is a need to create a special software tool with which to reveal these relationships from the text of the dictionary. While working with the tool, the user’s request may vary from an elementary reference about a specific word to generalized grammatical and semantic information related to the entire classes of language units, as well as various relationships developing and functioning in the language system. Elaborating such software tool implies the selection of appropriate theoretical framework. As such, we use the theory of lexicographic systems and the theory of semantic states by V.A. Shyrovok, the main provisions of which are outlined in [9].

3. Method

Developing effective tool with which to extract linguistic information from explanatory dictionary text requires respective theoretical framework. As such we have selected the theory of lexicographic systems and the theory of semantic states by Prof. Shyrovok [9].

According to the theory of lexicographic systems, an explanatory dictionary (like any other dictionary) is considered as a lexicographic system (L-system). And the L-system itself is an information system in which one or several lexicographic effects are induced. The main relations in this system are the relations “subject – object” and “form – content”. Any L-system is defined by the following components:

- D is a fragment of reality, which is the object of lexicographic description;
- S is a subject that makes lexicographic description of D (in our case, we associate it with the authors of the dictionary);
- Q is lexicographic effect observed S by the subject in D and transformed in a set of elementary information units $I^Q(D)$ (in our case, we interpret this component as a set of linguistic units composing a dictionary headword list);
- $V(I^Q(D))$ is a set of descriptions $I^Q(D)$; $S: I^Q(D) \rightarrow V(I^Q(D))$.

In view of the above the following statement will be true for any headword x :

$$I^Q(D) = \{x\}; \forall x \in I^Q(D) S: x \rightarrow V(x); \cup V(x) = V(I^Q(D)) \quad (1)$$

Where $V(x)$ in the dictionary is the text of the dictionary describing a headword x . Hence $V(I^Q(D))$ is a collection of all dictionary entries. On the set of descriptions $V(I^Q(D))$ and, particularly, on each $V(x)$, there can be defined two structures: β and $\sigma[\beta]$. They are the carriers of the linguistic facts and regularities in lexicographic system. At the same time β is set of “very simple” structural elements of

the dictionary such as words, abbreviations, labels, notes, figures, elements of grammar and vocabulary description, etc.). This can be formulated in the following way. For each $x \in I^Q(D)$, a set of structural elements $\beta(x)$ which compose $V(x)$ is determined according to the following principles:

1. $x \in \beta(x)$;
2. Any fragment of the dictionary entry $V(x)$ can be built of the elements $\beta(x)$;
3. The principle of forming the elements $\beta(x)$ is to be common for all $V(x)$, i.e. for all $x \in I^Q(D)$.

It is necessary to indicate importance of the formulated principles of forming β -structures in lexicography. Rule 2 is actually a requirement for the universality of the dictionary metalanguage: any linguistic fact that is fixed in a particular dictionary must be reflected in its metalanguage. Principle 3 implies that all linguistic facts of the same type and phenomena must have a unified representation in lexicographic description. These rules provide objective prerequisites for a formalized definition of the process of linguistic achievement using a lexicographic system.

In their turn, the β elements join into lexicographic structures $\sigma[\beta]$, corresponding to the description of linguistic phenomenon attributed to a headword. So, the whole lexicographic description of the headwords is defined by the elements $(\beta, \sigma[\beta])$. Each dictionary entry of DLE 23 is assigned a basic structure (Fig. 2).

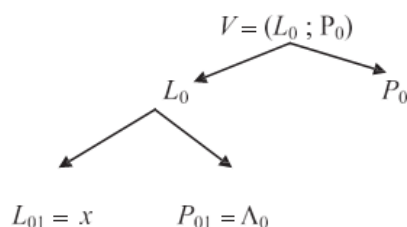


Figure 2: Basic structure of DLE 23 entry

Let us demonstrate the examples of $\sigma[\beta]$ that form lexicographic description of the headword *agua*. The text of the dictionary element is given in a format that preserves the font markup used in online version of the dictionary (Fig. 3).

agua

Del lat. *aqua*.

1. f. Líquido transparente, incoloro, inodoro e insípido en estado puro, cuyas moléculas están formadas por dos átomos de hidrógeno y uno de oxígeno, y que constituye el componente más abundante de la superficie terrestre y el mayoritario de todos los organismos vivos. (Fórm. H₂O).
2. f. Líquido que se obtiene por infusión, disolución o emulsión de flores, plantas o frutos, empleado como refresco o en medicina y perfumería. *Agua de azahar, de cebada, de limón.*
3. f. lluvia (l acción de llover). U. t. en pl. con el mismo significado que en sing.
4. f. lágrimas (l gotas de la glándula lagrimal). *Se le llenaron los ojos de agua.* U. t. en pl. con el mismo significado que en sing.
5. f. Vertiente de un tejado. *Una cubierta a dos aguas.*
6. f. Mar. marea (l movimiento periódico de las aguas del mar).
7. f. Mar. Rotura, grieta o agujero por donde entra en la embarcación el agua en que ella flota. *Abrirse, descubrirse un agua.*
8. f. pl. Visos u ondulaciones que tienen algunas telas, plumas, piedras, maderas, etc.
9. f. pl. Visos o destellos de las piedras preciosas.
10. f. pl. Manantial de aguas mineromedicinales.
11. f. pl. agua mineromedicinal. *El balneario es famoso por sus aguas.*
12. f. pl. Zona marítima próxima a la costa de un lugar. *Naufrió en aguas de Cartagena.*
13. f. pl. Mar. Corrientes del mar. *Las aguas tiran o van hacia tal parte.*
14. f. pl. Mar. Estela o camino que ha seguido un buque. *Buscar, ganar, seguir las aguas de un buque.*
15. interj. jerg. U. para avisar de la presencia de cualquier tipo de autoridad
16. interj. Mar. hombre al agua.

Figure 3: DLE 23 dictionary entry for headword *agua*

Based on the text analysis of online version of DLE 23 entries, we distinguish the following parameters for the left part L_0 : RR (lemma forms), DUPL (regional variant), ETYM (etymology), MORPHO (inflection), ORTHO (orthography) and UNCRT (undefined parameter). Each parameter is represented in our model as a text string.

The right part P_0 is composed of the elements of lexical meaning descriptions. The polysemy of the headword is determined by the number of these descriptions. Each description may include several structural elements, namely MNGN (definition No), REM (set of labels), DEF (definition), ED (encyclopedic note), COM (comment), and IL (illustration).

The text line of that DLE 23 entry can be subdivided into smaller fragments, each of them containing a label of specific type: REM-GR (grammar); REM-US (usage); REM-ST (stylistics); REM-DOM (domain); and REM-REG (geographic region). As a rule, the lexical meaning in the entry text is described by the structural element DEF. The comments (COM) are consistent with the definition. Each definition and each comment can be accompanied by its own illustrations (IL). The structure of the interpretation may include several DEF, COM and IL. The splitting of text into structural elements for the heading word *agua* (water) is shown in Table 8. As an example we have taken lexical meaning descriptions 1, 2, 7 and 15.

Table 8

The content of right part elements for the headword *agua*

Element	Content
MNGN	1.
REM-GR	f.
DEF	Líquido transparente, incoloro, inodoro e insípido en estado puro, cuyas moléculas están formadas por dos átomos de hidrógeno y uno de oxígeno, y que constituye el componente más abundante de la superficie terrestre y el mayoritario de todos los organismos vivos
ED	(Fórm. H ₂ O)
MNGN	2
REM-GR	f.
DEF	Líquido que se obtiene por infusión, disolución o emulsión de flores, plantas o frutos, empleado como refresco o en medicina y perfumería
IL	Agua de azahar, de cebada, de limón
MNGN	4.
REM-GR	f.
DEF	lágrimas (gotas de la glándula lagrimal)
IL	Se le llenaron los ojos de agua
COM	U. t. en pl. con el mismo significado que en sing
MNGN	7.
REM-GR	f.
REM-DOM	Mar.
DEF	Rotura, grieta o agujero por donde entra en la embarcación el agua en que ella flota
IL	Abrirse, descubrirse un agua
MNGN	15.
REM-GR	interj.
REM-US	jerg.
DEF	U. para avisar de la presencia de cualquier tipo de autoridad.

According to the theory of semantic states, any linguistic unit, when used in a context, adopts a certain semantic state which represents a sum of grammatical and lexical meanings. In our case, we consider the dictionary as a collection of semantic states of the headwords, the features of which are fixed by the elements $P(x)$.

4. Results and discussions

4.1. Interface of VLL DLE 23 and its research toolkit

As it shown in Fig. 4, the interface of the VLL DLE 23 laboratory consists of four elements: (1) top menu bar containing tools for working with the headword list and the text of DLE 23; (2) headword panel designed to search for words and navigate in the headword; (3) text box to display dictionary entries (the format corresponds to the original online version of DLE 23); and (4) text box to view HTML text of dictionary entries. The top menu bar includes two tools: “Selection” and “Statistics”. The first one contains a group of parameters to make a sample of dictionary entries containing headword linguistic features (type, structure of the register word, homonymy, number of lexical meanings, etc.). The second one generates statistics for a specific sample of the entries or the entire dictionary.

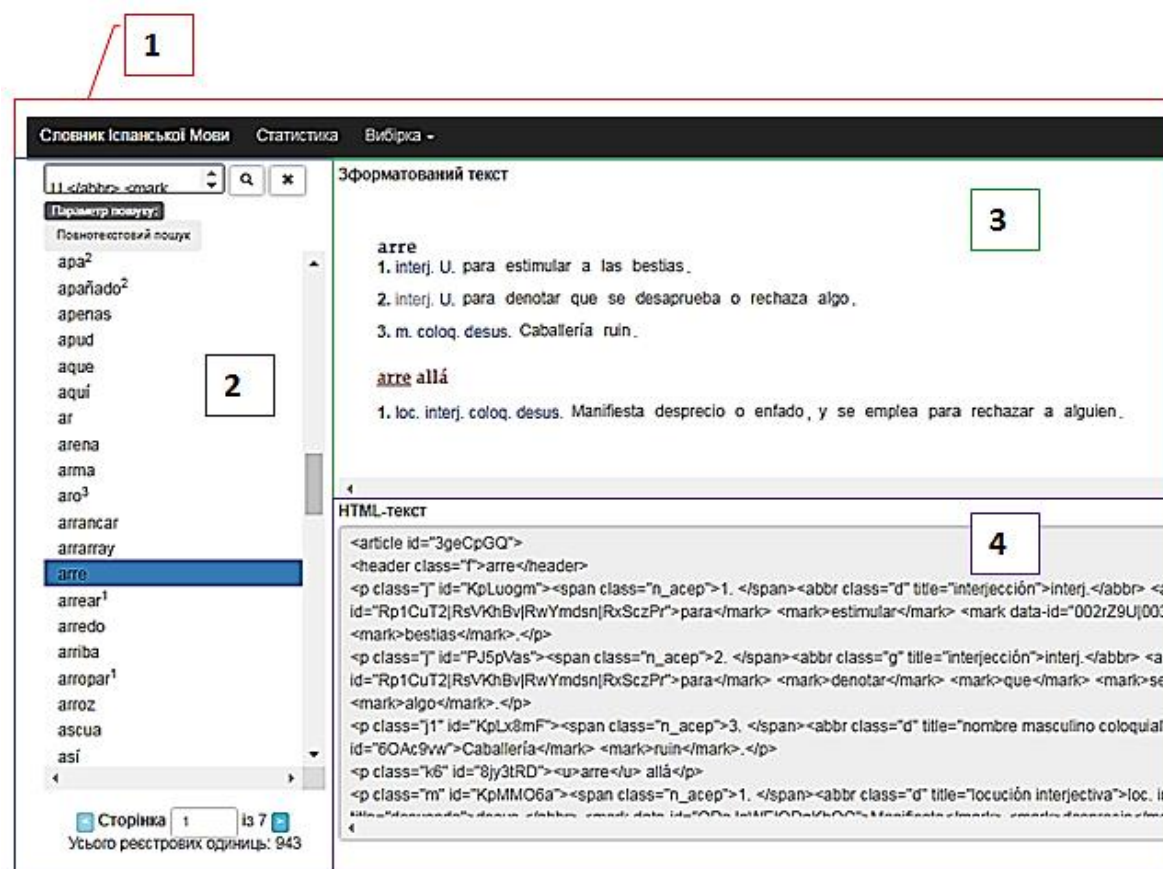


Figure 4: General layout of VLL DLE 23 interface.

With Selection tool it's possible to form an inventory of the Spanish vocabulary reflected in DLE 23. Two tools can be used to select and quantify:

- Cognate words, including homonymous cognate words;
- Spanish vocabulary elements by its origin;
- Words having a specific suffix or prefix, as well as words consisting of a root;
- Language units belonging to a certain linguistic level, for example: morphemes, lexemes, phrases;
- Units of other types such as abbreviations and acronyms.

The VLL DLE 23 tools give the opportunity to the users to select the entries the headwords of which have common grammatical, lexical and other features reflected in the text of the dictionary entries. These properties are displayed in the definition and other elements of the dictionary entry using certain keywords and expressions. In particular, the following linguistic properties of the

headwords can be distinguished from the text of the dictionary entries:

- Participation of the morphemes in forming the words to express particular lexical meaning;
- Tracing the way through which the headword came from another language to Spanish (directly or through intermediary language);
- Lexical meaning development of the headwords of foreign or native origin;
- Semantic structure of the headwords, including diminutives, augmentatives etc.;
- Availability or absence of Spanish equivalents to the words coming from foreign languages;
- The etymology of the headwords belonging to different languages of origin;
- Ability of the words both native and foreign origin to form collocations, for example “Noun + Adjective” and other types (adjectival, verbal prepositional etc.);
- Words belonging to a particular semantic field headed by a broader word.

4.2. Examples of VLL DLE 23 application

The current version of VLL DLE 23, the developers of which are the authors of this article, is intended for making an inventory of language units and conducting linguistic researches with statistical calculations. Let us consider some examples of applying the VLL.

4.2.1. Formation of lexicographic types

This function consists in the selection of dictionary entries, the headwords of which can be attributed to a certain class by their common linguistic properties. With VLL DLE 23 the classes of Spanish words, united by common linguistic (grammatical, semantic, usage) properties, are possible to be visualized. Such classes of the words described in the dictionary are called lexicographic types. Let us form, for example, a lexicographic type composed of the verbs, the conjugation of which is similar to that of the verb *agradecer* (to thank). The verbs in question are conjugated using a set of inflections {-zco, -ces, -ce, -cemos, -céis, -cén, etc.}. The figure 5 shows the result of VLL DLE 23 work on the selection of the verbs representing such lexicographic type. On the left, a list of verbs included in the lexicographic type is shown. At the top of the figure is the dictionary entry of a verb. Below is the “Statistics” window, shown that the formed lexicographic type includes 218 verbs, of which 5 are homonymous. In similar way the user can get any other lexicographic types taking into account various linguistic properties. For example, we can get the verbs denoting a movement from one point to another. In this case, lexicographic type will cover the words such as *abandonar* (to walk leaning on a stick), *ambлар* (to amble, to stroll), *caminar* (to walk), *callejear* (to wander), *correr* (run) etc.

The screenshot displays the VLL DLE 23 interface. On the left, a scrollable list of verbs includes: enrudecer, enruinecer, ensandecer, ensamecer, ensilvecerse, ensoberbecer, ensombrecer, ensordecer, entallecer, entenebrecer, enternecer, entestecer, entigrecerse, entontecer, entorpecer, entreparecerse, entullecer, entumecer, envagüecer, and envanecer. The main area shows the dictionary entry for 'abastecer', including its conjugation and a definition: '1. tr. Proveer a alguien o a algo de bastimentos, víveres u otras'. Below this is a 'Статистика' (Statistics) window titled '(остання вибірка)' (last selection). The statistics are as follows:

Статистика (остання вибірка)	
Кількість статей	Кількість фразеологізмів
Усього: 218	1-го типу(ім.+прик.): 0
Відсилкові: 0	2-го типу(інш.): 7
З омонімією: 5	
Морфем: 0	
З зарумбовою зоною: 4	
Без зарумбової зони: 214	

At the bottom, there is a page indicator 'Старінка 1 із 2' and a 'Закрити' (Close) button.

Figure 5: Lexicographic type of the verbs conjugated the same way as *agradecer*.

4.2.2. Researching language regularities

One of the examples of linguistic research to be conducted by means of VLL DLE 23 is the way of forming verbal nouns denoting the action and the result of this action. Such words are described in DLE 23 using the definition: *Acción y efecto de + verb*. This definition pattern serves as a search query. The results obtained are shown in Fig. 6.

The screenshot shows a search interface with a search bar containing the text "acortamiento". Below the search bar is a list of suggested headwords, with "acortamiento" selected. To the right, the definition for "acortamiento" is displayed, consisting of three numbered items. Below the definition is a section labeled "HTML-текст" containing the HTML-escaped version of the definition. At the bottom of the interface, there is a pagination control showing "Сторінка 1 із 28" and a total count of "Усього реєстрових одиниць: 4182".

Параметр пошуку:
Повнотекстовий пошук

acortamiento

acoso
acotamiento
acotejo
acrecentamiento
acrecimiento
acreditación
acribadura
acristalamiento
activación
actuación
acuartelamiento
acuatizaje
acuchillado
acuchillamiento
acuerdo
aculturación
acumulación
acumulamiento
acúmulo

acortamiento

1. m. Acción y efecto de acortar o acortarse .
2. m. *Astron.* Diferencia entre la distancia real de un planeta al So
3. m. *Ling.* Palabra resultante de la reducción de la parte final o i

autobús y bacteriófago , respectivamente .

HTML-текст

```
data-id="DjUR5ks|DjcR8zI">Diferencia</mark> <mark data-id="FkLKqOW|FKI</mark> distancia</mark> <mark>real</mark> <mark data-id="BtDkacLj|BtFYzr</mark> al</mark> <mark data-id="YFFnS1v|YFFxu1L|YFGdevUj|Y</mark> Tierra</mark> , <ma</mark> distancia</mark> <mark>proyectada</mark> <mark data-id="Y5fMg1</mark> de</mark> <mark data-id="ESraxkH|MIZ5vEtj|NWnohQu">la</mark> <mark data-id="ESraxkH|MIZ5vEtj|NWnohQ</mark> <p class="j" id="Q7j7HfH"><span class="n_acep">3. </span><abbr class="g" data-id="RUI938s">Palabra</mark> <mark data-id="WFMntQj|WFOpceQ">re</mark> <mark data-id="ESraxkH|MIZ5vEtj|NWnohQu">la</mark> <mark>reducción</mark> <mark> parte</mark> <mark> fin</mark> <mark data-id="ESraxkH|MIZ5vEtj|NWnohQu">la</mark> <mark> parte</mark> <mark> fin</mark> <mark data-id="BtDkacLj|BtFYznp">de</mark> <mark data-id="RLQXxGn">otra</mark> </mark> <mark> cine</mark> , <mark> bici</mark> , <mark> bus</mark></i> <mark> y</n</mark> <mark> cinematógrafo</mark> , <mark> bicicleta</mark> , <mark> autobús</ma
```

< Сторінка 1 із 28 >
Усього реєстрових одиниць:
4182

Figure 6: Selection of the headwords denoting the action and the action result.

On the basis of these results the researcher can make certain conclusions regarding the use of the suffixes to form such kind of nouns, e.g.:

- *-ada* if the noun is derived from the verbs denoting blows or similar actions: *bofetada* (slap), *puñalada* (blow) etc.;
- *-azo* if the noun is derived from the verbs denoting blows with something: *botellazo* (blow with bottle), *culatazo* (blow with a rifle butt) etc.;
- *-ido* if the noun is derived from the verbs denoting sounds or noises: *chillido* (scream), *ladrido* (barking) etc.;
- *-ón* if the noun is derived from the verbs denoting energetic or quick actions *empujón* (push), *resbalón* (slip) etc.

This information can be used not only for linguistic research, but also for the preparation of teaching materials on Spanish grammar.

4.2.3. Statistics generation

In addition to linguistic researches VLL DLE 23 is designed to generate statistics, both for the entire dictionary and for a separate sample. For example, you need to count how many words in

Spanish have different forms for masculine and feminine gender. The result of the work is shown in Fig. 7. The statistics obtained are as follows:

- 19,011 headwords out of which
- 840 are homonyms;
- 111 are morphemes;
- 2257 form collocations;
- 16754 don't form collocations.

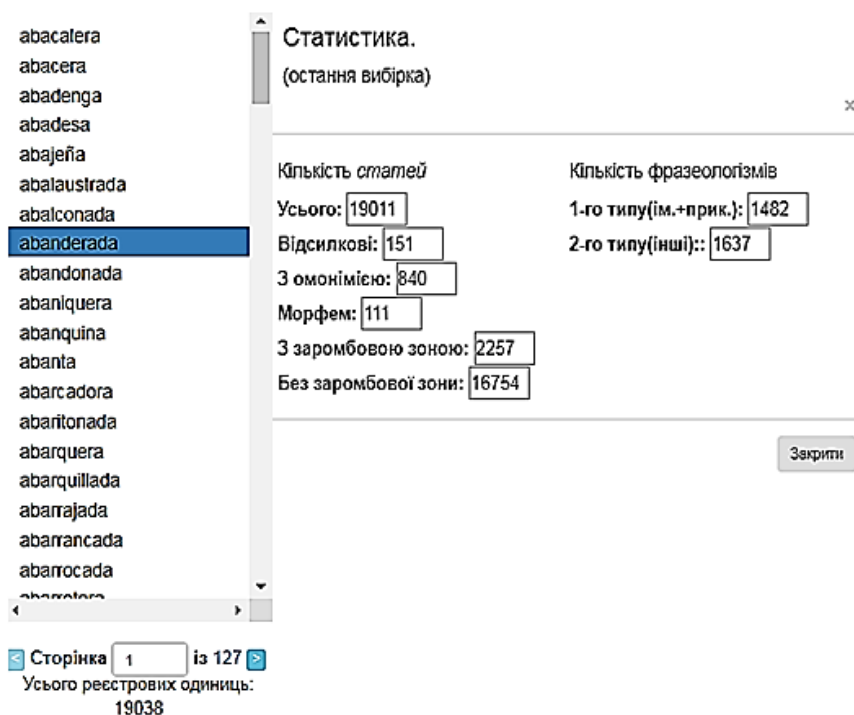


Figure 7: Statistics showing a number of headwords with different gender forms

5. Conclusions and future works

Currently the developed virtual lexicographic laboratory gives a user the opportunity to analyze the text of the explanatory Spanish dictionary and perform on its basis:

- An inventory of the headwords satisfying the specified parameters (native word, foreign word; morpheme, abbreviation, word, collocation etc.);
- Extraction of linguistic characteristics of headwords from the text. This makes it possible to identify regularities in the Spanish language, which are presented in the implicit form in the dictionary;
- Statistical studies that show the frequency of the considered linguistic phenomena (for example, the ratio of national and borrowed vocabulary).

In future the current version of VLL DLE 23 will be provided with an expanded toolkit to work separately with each dictionary entry element, determining not only its presence or absence, but also its specific content.

6. References

- [1] A. Wills, E. Jóhannsson, Reengineering an Online Historical Dictionary for Readers of Specific Texts, in: I. Kosem, T. Z. Kuhn (Eds.), Electronic lexicography in the 21st century: Smart lexicography, Proceedings of eLex 2019 conference, Sintra, Portugal, 2019, pp. 116–129.

- [2] M. Alipour, B. Robichaud, M.-C. L'Homme, Towards an Electronic Specialized Dictionary for Learners, in: I. Kosem, M. Jakubiček, J. Kallas, S. Krek (Eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age*, Proceedings of eLex 2015 conference, Herstmonceux Castle, United Kingdom, 2015, pp. 51–69.
- [3] R. Lew, Online dictionary skills, in: I. Kosem, J.Kallas (Eds.), *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of eLex 2013 conference, 2013, Tallinn, Estonia, pp. 16–31.
- [4] Sobre la 23.^a edición del Diccionario de la lengua española, 2014. URL: https://www.rae.es/sites/default/files/Cifras_23.a_edicion_del_Diccionario.pdf
- [5] El nuevo diccionario académico será digital y más panhispánico, 2017. URL: <https://www.rae.es/noticias/el-nuevo-diccionario-academico-sera-digital-y-mas-panhispanico>.
- [6] T. Roth, Going Online with a German Collocations Dictionary, in: I. Kosem, J.Kallas (Eds.), *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of eLex 2013 conference, Tallinn, Estonia, 2013, pp. 152–163.
- [7] D. Deksne, I. Skadiņa, A. Vasiljevs, The modern electronic dictionary that always provides an answer, in: I. Kosem, J.Kallas (Eds.), *Electronic lexicography in the 21st century: thinking outside the paper*. Proceedings of eLex 2013 conference, 2013, Tallinn, Estonia, pp. 421–434.
- [8] V. Apresjan, N. Mikulin, Dictionary as an Instrument of Linguistic Research, in: Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, Tbilisi, Tbilisi State University, 2016, pp. 224–231.
- [9] V. Shyrokov, *Computer lexicography*, Kyiv, 2011.