Parsing technique in computer lexicography (case of Spanish dictionary DLE 23)

Vasyl Lytvyn[1,†], Yevhen Kupriianov[2,*,†], Iryna Ostapova[3,†], and Mykyta Yablochkov[3,†]

[1] *Lviv Polytechnic National University, 12 Stepana Bandera Street, Lviv, 79013, Ukraine*
[2] *National Technical University "Kharkiv Polytechnic Institute", Kyrpychova str. 2, Kharkiv, 61002, Ukraine*
[3] *Ukrainian Lingua-Information Fund of NAS of Ukraine,3, Holosiivskyi avenue, Kyiv, 03039, Ukraine*

Abstract

The paper describes the main methodological and technological solutions for parsing of dictionaries elaborated at the Ukrainian Lingua-Information Fund (ULIF) by the authors. The research is carried out on the basis of the digital text of the Dictionary of the Spanish Language (DLE 23). First of all, explanatory dictionaries of national languages as the most complicated multi-parameter lexicographic systems are of the greatest interest, since they provide the most complete lexicographic description of a language, are created by leading specialists (linguists and IT-engineers) and have wide opportunities to use modern digital technologies to the fullest extent. But the problem of extracting linguistic information for its further use (especially for computerized text processing) has not been solved at present. Therefore, the focus of the research is made on the parsing technology to develop the virtual lexicographic laboratory on the basis of the dictionary text parsed from the online version of DLE 23.

Keywords

digital lexicography, lexicographic system, lexicographic data model, lexicographic data, data analysis, database, user interface, lexicographic data formats.

1. Introduction

One of the main tasks of modern lexicography is to use the potential of the digital environment to meet the information needs of today's advanced users and lexicographers. Currently, both dictionary compilation and its logistics rely on digital technologies. These are primarily corpus technologies CQS (Corpus Query Systems) and digital systems for dictionary compilation and updating DWS (Dictionary Write Systems). A new challenge for lexicography is the direct participation of IT specialists at all stages of creation and use of lexicographic products [1–4, 9, 10].

Despite significant advances in modern digital lexicography, standards for representing lexicographic data that are maximally adapted to the conditions of the digital environment remain open. This also applies to the standards of work with these data.

First of all, explanatory dictionaries of national languages are of the greatest interest, as they provide the most complete lexicographic description of a language, are created by leading specialists and have the opportunity to use modern technologies to the fullest extent [6, 7, 8]. The use of CQS and DWS allows to work non-stop, i.e. the process of dictionary creation and editing is ongoing and the user has the opportunity to access lexicographic information at the current stage of the lexicographic process (Oxford English Dictionary [7], Dictionary of the Ukrainian Language in 20 volumes [8]). However, despite the availability of advanced (as compared to printed versions) user

CLW-2025: Computational Linguistics Workshop at 9th International Conference on Computational Linguistics and Intelligent Systems (CoLInS-2025), May 15–16, 2025, Kharkiv, Ukraine

[*] Corresponding author.

[†] These authors contributed equally.

Vasyl.V.Lytvyn@lpnu.ua (V. Lytvyn); eugeniokuprianov@gmail.com (Y. Kupriianov); irinaostapova@gmail.com (I. Ostapova); gezartos@gmail.com (M. Yablochkov)

0000-0002-9676-0180 (V. Lytvyn); 0000-0002-0801-1789 (Y. Kupriianov); 0000-0001-8221-3277 (I. Ostapova); 0000-0003-1175-1603 (M. Yablochkov)

interfaces, their ability to search, analyze and summarize linguistic information, primarily for professionals, is still limited. Authors traditionally develop not only the structure and content of dictionary entries, but also the search capabilities of the dictionary. As a result, the problem of extracting linguistic information for its further use by experts in their research has not yet been solved. Therefore, the goals of our research work are 1) to develop interface schemes for linguistic research based on the dictionary entries of an explanatory dictionary; 2) to design the structure and implementation of a database to support lexicographic data; and 3) to build an effective research toolkit. Unlike paper dictionaries, this is a feasible task for digital dictionaries [9].

2. Recent developments in dictionary parsing

2.1. Parsing and data analysis in general terms

The process of converting structured text into a particular data structure is called parsing. Any suitable format of the information contained in the source text can be used as a data structure type. In natural language processing the term "parsing" was first used to describe syntax analysis and later to refer to the related analysis, parsing still translated as syntax analysis by Google Translate. This technique is employed in contemporary settings when processing vast volumes of data is required, which can be challenging, if not impossible, to handle manually. Parsing is typically used to organize, process, store, and extract data from websites. Making data available in a machine-readable format is the primary goal of parsing, as most data is often given in a human-readable form. Depending on particular needs and goals, parsing technologies can be used in a wide range of ways. In this regard parsing is becoming more and more crucial for any business as the digital economy grows. An increasingly significant component of research and organizational management is data analytics. Although data comes from a variety of sources, the Web remains its biggest storage. Businesses are seen to be using more complex methods to retrieve information from the internet as big data analytics, artificial intelligence, and machine learning develop.

Parsers carry out the parsing process. Normally, when people refer to "parsers", they mean software applications. Parsers, sometimes known as "bots", are designed to browse websites, download pertinent pages, and extract information that can be used for different purposes. These bots are able to retrieve vast amounts of data in a short amount of time by automating this process. Given that the data is updated and changing frequently, this offers clear benefits. Websites gather all kinds of data, including text, photos, and videos. There are many uses for parsing, particularly in data analysis. In order to analyze consumer sentiment, market research firms utilize parsers to extract data from internet forums or social media. Others research competition by extracting data from vendor websites such as Amazon or eBay. Google frequently analyzes, ranks, and indexes its content via parsing. Contact parsing is another practice used by many businesses when they take contact details off the internet and use them for marketing. Few limitations exist on the applications of parsing. Mostly, everything comes down to goals and creativity.

Nevertheless, it should be noted that parsing data, including bank account information and other personal information, for the sake of fraud and intellectual property theft is a negative aspect of parsing. Though the precise approach may differ based on the program or tools utilized, all parsers adhere to the following three fundamental principles:

Step 1: Sending HTTP request to the server. The parser sends an HTTP request to the target website as its initial action.

Step 2: Extracting and parsing the code from the website. The bot can read and extract a website's HTML or XML code once it has access to the parser. The content of the website is structured according to this code. To be able to recognize and extract predefined elements or objects, the parser then parses the code, which is basically breaking it down into its component parts. These could be identifiers, tags, classes, ratings, particular sentences, or other data.

Step 3: Locally storing the pertinent data. The parser will save the pertinent information locally after receiving, extracting, and parsing the HTML or XML. Typically, the data is saved in a structured format, such as .csv or.xls.

Following the completion of these procedures, the data can be utilized as planned. However, the process is actually carried out numerous times rather than all at once. Numerous problems that require attention are to blame for this. A website may crash, for instance, if poorly developed parsers send an excessive number of HTTP requests. The restrictions on what bots may and cannot accomplish vary from website to website.

2.2. Parsing techniques earlier used in ULIF dictionary-making practice

Monolingual explanatory dictionaries that describe a general-purpose language in its entirety and specialized dictionaries that describe special-purpose languages are the first-order objects for parsing. The list of objects should be restricted to reputable dictionaries that most accurately reflect the range of language units because parsing is a somewhat resource-intensive process. Though they may not be very extensive dictionaries, they can be viewed as an extension of the primary general language dictionary, which includes things like adding new meanings, expanding the list of language units, and using different representations of parameters (as opposed to the standard one).

Parsing a printed dictionary text. This is the very first technology that was developed and created at ULIF to convert the printed Ukrainian language dictionary (SUM 20) into the database format during the creation of a virtual lexicographic laboratory designed to publish and update the dictionary. The parsing process using this technology includes the following steps:

1. Analyzing the printed text and constructing a conceptual model of the lexicographic system.
2. Scanning the text, recognizing and reproducing the text in a digital word processor (with the recovery of structural element markers in linear text).
3. Database schema construction based on the conceptual model.
4. Conversion of marked-up text into a database.
5. Building a computer-based toolkit to provide access to the structural elements of dictionary entries.
6. Development of computer toolkit to build sub-dictionaries on the basis of the main dictionary text.

Parsing a dictionary in pdf format is an improved variant of the first technology, and it is applied to work with dictionaries in the publishing system format (PDF). This parsing technology was applied to the Dictionary of Ukrainian biological terminology. The main stages are:

1. Converting dictionary text into word processor format (.doc).
2. Building a conceptual model of the lexicographic system.
3. Text verification (recovering structural element markers in linear text). Using basic HTML markers for marking up structural elements.
4. Building an XML schema based on a conceptual lexicographic model.
5. Conversion of marked-up text into a structured XML file.
6. Building a database schema based on XML structure.
7. Converting an XML file into a database.
8. Building a WEB-site offering a predefined interface.

3. Parsing technique for DLE 23

Our interest in parsing the Dictionary of the Spanish Language (DLE 23) [6] is motivated by the factors that are as follows: 1) the international status of the Spanish language; 2) the scientific status of the dictionary; 3) the distinctive school of lexicography. Another, perhaps the most important reason for choosing the dictionary for the research is the availability of a digital version in HTML5 format, which guarantees the authenticity of the dictionary text and transparent structure. DLE 23 is a basic dictionary that includes words that are frequently used in Latin America and Spain. The lexical meanings of language units and an in-depth explanation of their grammatical, syntactic, and pragmatic characteristics are included in every dictionary entry. It should be mentioned that the dictionary's philosophy is derived from the ideas of renowned Spanish lexicographer J. Casares [5]. This idea holds that a dictionary is a tool that gives the user the resources they need to locate the pertinent words and phrases they might need during communication, rather than a collection of entries sorted alphabetically [10]. However, our goal is to develop techniques for parsing the entire text of the digital version of the dictionary and to present the results of the study for use not only for advanced users who use the dictionary as a reference and information system, but also for specialists in the field of computer assisted text processing.

In the context of our research, the explanatory dictionary of a national language is considered as a comprehensive source of information for linguistic research. Due to the large volume, complex structure and completeness of lexicographic description, such dictionaries carry a huge number of implicit linguistic, cognitive, logical and other relations that are difficult or almost impossible to be studied by traditional methods. For a digital dictionary, it is necessary to provide access to any structural element of a dictionary entry and the ability to select a set of entries that meet the user's research interests. In other words, the interface of a digital dictionary should provide the possibility of searching not only by the registered word, as in most digital dictionaries.

The parsing procedure's algorithm, first created and used for DEL 23, is as follows:

1. Scanning the printed text of the dictionary, recognizing and generating a verified list of headwords of the dictionary entries.
2. Developing a program (bot) that reads dictionary entries in HTML5.0 format from the dictionary website.
3. Building a conceptual model of the lexicographic system (based on the structure of the dictionary entry as represented on the dictionary website).
4. Establishing the correspondence of HTML 5.0 markers to the structural elements of the conceptual model (a special software package was developed for this purpose).
5. Building the schema of an XML file.
6. Converting linear text in HTML markup to an XML file.
7. Designing a database.
8. Building software tools to provide the access to the structural elements.

ULIF in its dictionary-making practice makes a distinction between an encoding scheme or database that may replicate a lexicographic system and a formal model, which is a conceptual representation of the system. Regardless of the conditions and/or limitations placed on its ultimate representation, the form and content of lexical information is considered in abstract way. This is particularly important since these possible representations will vary from one application to another; in particular, dictionaries may be encoded not only for publication purposes in print or electronic form, but also to create computational lexicons for use in natural language processing applications. Therefore, it is currently preferable to use the XML format to represent a conceptual model of a lexicographic language, which can later be transformed into many alternative formats.

XML (Extensible Markup Language) is a standard proposed by the World Wide Web Consortium (W3C) for building markup languages of hierarchically structured data for exchange between different applications, in particular via the Internet. It is a simplified subset of the SGML markup language (note that SGML was used to mark up the text of the Oxford Dictionary when it was digitized). An XML document consists of text characters and is human readable. This format is flexible enough to be suitable for use in a variety of industries. In other words, this standard defines a meta-language from which specific, subject-oriented data markup languages are defined by imposing constraints on the structure and content of documents.

At the beginning of its lexicographic activity ULIF used technology to convert files in .rtf markup (publisher format) directly into a database, the schema of which was based on a lexicographic data model. However, with the parsers for the Dictionary of Spanish and the Dictionary of Ukrainian Biological Terminology made use of XML-format, it is possible to reject the format imposed by the printed representation of data and effectively capture the structure of the conceptual model. HTML5.0 (Hyper Text Markup Language), or Hypertext Markup Language, was originally used for visual text markup. Today it is comparable to XML in its capabilities. Therefore, it can be effectively used to convert XML into a format for visual representation in the WEB. Additionally, it serves as the primary format for information presentation on the Internet. All publishing platforms support the communication publishing format known as PDF. It is an electronic version of a collated text that is prepared for printing. Every lexicographic system needs a paper outlining a conceptual data model, regardless of the type of data representation. A document in the doc, docx, or pdf formats should serve as the standard representation of this model.

Lexicographic systems are well-structured linguistic data that are represented in the digital environment's communicative formats. Although their traditional function as reference and information systems remains relevant, NLP (natural language processing) tasks are becoming more and more important. It's semantics, to begin with. Dictionary interfaces were developed in the "paper environment" and are focused on providing access to a certain dictionary entry using a specific form of a linguistic unit, usually a headword. This strategy has been mostly carried over into digital reproductions. The user's unique linguistic competencies serve as the sole foundation for deeper linguistic information processing. Research of linguistic data of the lexicographic system can be brought to the level of algorithms thanks to the structure of a dictionary entry, which is suitably represented by the language environment. The user has access to arrays of linguistic data chosen based on a typical set of parameters rather than individual entries. Parsing dictionaries and storing information in digital communication formats of the digital environment opens up great possibilities for the integration of lexicographic systems.

4. Experiment

The interface for DLE 23 online version is implemented taking into account the advantages of the digital environment for the user. First of all, it refers to the visualization of dictionary entries, which is shown below by examples.

Example 1. Dictionary entry visualization in a printed format.

abombar^{2} De *bomba.* **1.** tr. Dar forma convexa. ○ intr. **2.** Dar a la bomba. ○ prnl. **3.** Dicho de una cosa: Tomar forma convexa.

Example 2. Dictionary entry visualization in online version.

abombar^{2}

De *bomba.*

1. tr. Dar forma convexa.

2. intr. Dar a la bomba.

3. prnl. Dicho de una cosa: Tomar forma convexa.

Example 3. Dictionary entry visualization with HTML language (The text in bold is the text which can be seen on the screen; the rest being hidden text that manages the visualization of the dictionary entry).

<article id="088zJNJ">

<header title="Definición de abombar" class="f">**abombar**<sup>**2**</sup></header>

<a class="e2" title="Conjugar el verbo abombar2" href="#conjugacioncbD9rJq"></a>

<p class="n2">**De** <em>**bomba**</em></p>

```
<p class="j" id="04CwhyP"><span class="n_acep">1. </span><abbr class="d" title="verbo transitivo">tr.</abbr> <mark data-id="BrtRK35">Dar</mark> <mark data-id="IEvo12v |IFIVvz0" >forma</mark> <mark data-id="AgxvK91">convexa</mark>.</p>
<p class="j2" id="04E67HH"><span class="n_acep">2. </span><abbr class="d" title="verbo intransitivo">intr.</abbr> <mark data-id="BrtRK35">Dar</mark> <mark data-id="002rZ9U |003Ov93">a</mark> <mark data-id="ESraxkH|MiZ5vEt|NWnohQu">la</mark> <mark data-id="5pINrRS|5prGcPu">bomba</mark>.</p>
<p class="j2" id="04EVDwI"><span class="n_acep">3. </span><abbr class="d" title="verbo pronominal">prnl.</abbr> <mark data-id="BxLriBU|DgXmXNM">Dicho</mark> <mark data-id="BtDkacL|BtFYznp">de</mark> <mark data-id="b67JJSq|b6hEWeB|b6iKApr">una</mark> <mark data-id="B3yTydM|B4tWyfU">cosa</mark>: <mark data-id="ZzcN8W0">Tomar</mark> <mark data-id="IEvo12v|IFIVvz0">forma</mark> <mark data-id="AgxvK91"> convexa</mark>.
</p>
</article>
```

As can be seen from these examples, the transparency of the dictionary entry structure for the user is provided by the complexity of the hidden text. Based on the analysis of the texts of the online versions of DLE 23 entries, we identified the following parameters of the left part: *RR* (lemma forms), *DUPL* (regional variant), *ETYM* (etymology), *MORPHO* (inflection), *ORTHO* (spelling), and *UNCRT* (undefined parameter). Each parameter is represented in our model as a text string. The right part consists of elements describing the lexical meaning. The polysemy of a headword is determined by the number of these descriptions. Each description may include several structural elements, namely: *MNGN* (definition number), *REM* (set of marks), *DEF* (definition), *ED* (encyclopedic reference), *COM* (comment), and IL (illustration). The *REM* text string can be split into smaller fragments, each of which contains a label of a specific type: *REM-GR* (grammar); *REM-US* (usage); *REM-ST* (style); *REM-DOM* (domain); *REM-REG* (geographic region). As a rule, a lexical value in the input text is described by a DEF structural element. Additional comments (*COM*) are consistent with the definition. Each definition and comment can be accompanied by its own illustrations (*IL*). An interpretation structure may include several *DEF*s, *COM*s, and *IL*s. The example of the entry text decomposition for the headword **abombar** is shown in the table 1.

Table 1
DLE 23 Entry Elements for Headword **abombar**

Entry element	Value
RR	abombar
DUPL	Abombar
ETYM	De *bomba*
MNGN	1
REM-GR	tr.
DEF	Dar forma convexa.
MNGN	2
REM-GR	intr.
DEF	Dar a la bomba.
MNGN	3
REM-GR	prnl.
DEF	Dicho de una cosa: Tomar forma convexa.

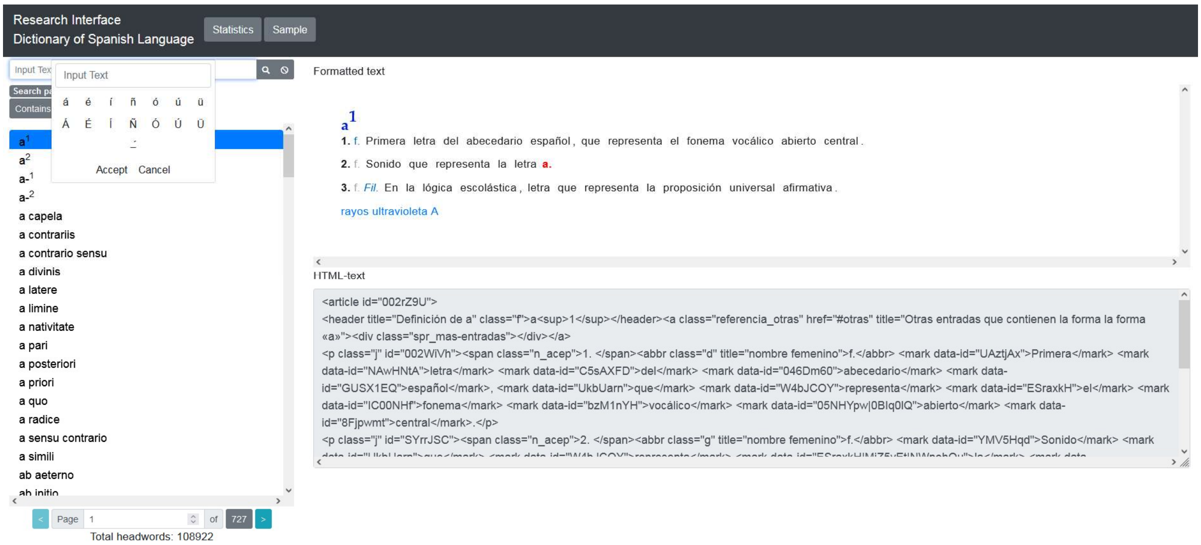
After XML-mapping of DLE 23 dictionary entries by the above-described principle, the next step was to create a lexicographic database. In our experience, relational databases have proven to be inefficient for lexicographic systems. In the case of relational databases, data is stored implicitly as a set of several tables and relationships between them. Working with individual tables as a single object requires the creation of a powerful software infrastructure. In addition, the evolutionary potential of such a digital object is limited by the opacity of the database.

It makes sense to express dictionary entries as classes in object-oriented programming languages with additional processing, editing, and storing in an explicit manner because they are the fundamental components of a lexicographic system with a precisely specified structure. The so-called NoSQL databases (document-oriented databases) offer this capability. A document (object) with a precisely defined structure – in our case, a dictionary entry – is the primary component that is stored and processed in databases of this kind. For our project, NoSQL databases' primary benefit is their capacity to store lexicographic objects explicitly without altering their internal structure. This allows for direct access to every component of the lexicographic object and significantly reduces the likelihood of editing and expanding it. The following criteria served as a guide when selecting a particular NoSQL database: 1) simplicity of usage; 2) transaction mechanism support; 3) parallelism support; 4) scientifically free. Consequently, the LiteDB database (http://www.litedb.org/) was used. It is a free, comparatively basic version of the MongoDB shareware database. Since LiteDB is constructed as a single library file (dll) and a single configuration file (xml), rather than as a whole software package, it also has the benefit of being easy to install and connect.

5. Results

Additional parameterization of dictionary items has been done in order to construct a prototype version of the VLL. Every headword has a set of parameters given to it: 1) headword variations; 2) headword structure; 3) headword type; 4) homonymy; 5) number of meanings; and 6) number of word combinations. The HTML text served as the basis for determining every parameter. The structure of a dictionary entry is clarified by searching for articles using combinations of these characteristics. The creation of an online application to interact with the VLL DLE database was the final task. Based on this, the application was developed. The Net Core 2.1 technology. To make modifying interface elements easier, a collection of HTML, CSS, and Bootstrap JavaScript templates were utilized.

The main window (Fig. 1) consists of the following interface elements: a panel, a register search panel, a register units window, a window for the formatted text of a dictionary entry, and a window for plain text with HTML markers. The search panel displays all the headwords of the dictionary in alphabetical order (corresponding to the selection conditions). The list of headwords is divided into groups of 150 items. The user can use the “Forward” and “Back” buttons or enter the page number in the text field to go to the corresponding part of the general list. The formatted dictionary entry is displayed in the same way as in the original online version of the dictionary. The HTML text of the dictionary entry is used for full-text search (the search string may contain HTML markers and meta-language elements). The DLE 23 VLL interface provides the following modes of working with the dictionary: a) headword list; b) dictionary entries; c) full-text search.



a^1

1. f. Primera letra del abecedario español, que representa el fonema vocálico abierto central.

2. f. Sonido que representa la letra **a.**

3. f. *Fil.* En la lógica escolástica, letra que representa la proposición universal afirmativa.

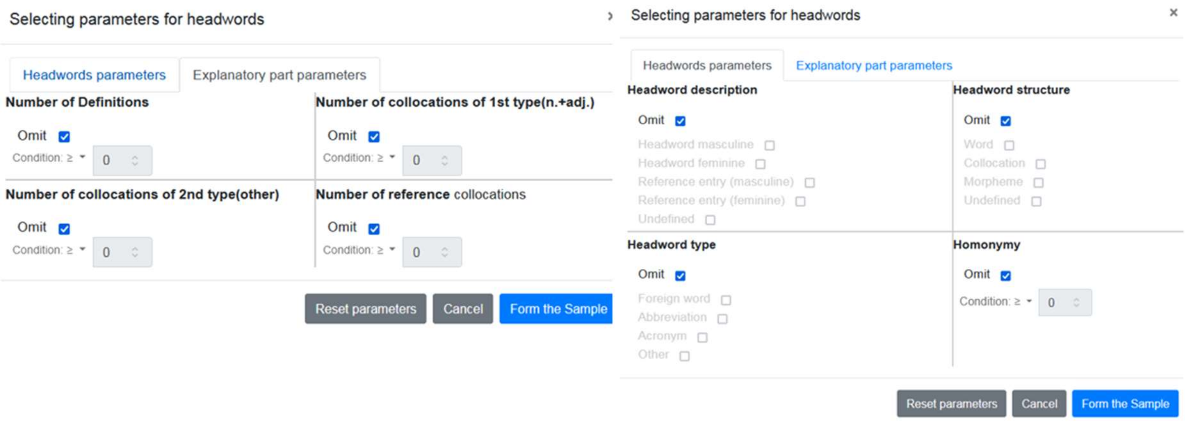
rayos ultravioleta A

```
<article id="002rZ9U">
<header title="Definición de a" class="f">a<sup>1</sup></header><a class="referencia_otras" href="#otras" title="Otras entradas que contienen la forma la forma
«a»"><div class="spr_mas-entradas"></div></a>
<p class="j" id="002WiVh"><span class="n_acep">1. </span><abbr class="d" title="nombre femenino">f.</abbr> <mark data-id="UAztjAx">Primera</mark> <mark
data-id="NAwHNtA">letra</mark> <mark data-id="C5sAXFD">del</mark> <mark data-id="046Dm60">abecedario</mark> <mark data-
id="GUSX1EQ">español</mark>, <mark data-id="UkbUarn">que</mark> <mark data-id="W4bJCOY">representa</mark> <mark data-id="ESraxkH">el</mark> <mark
data-id="IC00NHf">fonema</mark> <mark data-id="bzM1nYH">vocálico</mark> <mark data-id="05NHYpw|0BIq0IQ">abierto</mark> <mark data-
id="8Fjpwmt">central</mark>.</p>
<p class="j" id="SYrrJSC"><span class="n_acep">2. </span><abbr class="g" title="nombre femenino">f.</abbr> <mark data-id="YMV5Hqd">Sonido</mark> <mark
```

Figure 1: VLL DLE 23 main window.

Headword list search filters, which include entering a string of characters that either start or end with the search phrase, or you can choose a search word by clicking on it in the register list. When you don't know how to spell something, search filters help you discover it more quickly. Diacritical characters can be entered more easily with the help of a virtual keyboard. The headword list is composed of all word forms of the headwords (only a case for nouns and adjectives). Unlike the original DLE 23, the VLL DLE 23 list is supplemented with feminine forms and regional variants. Regardless of the operating mode, information about the number of dictionary entries is always displayed. The current version works with 106323 entries.

The **dictionary entry mode** in the current version of DLE 23 is intended for selecting dictionary entries that meet the parameters of the structural elements of a dictionary entry. The mode is activated by clicking the “Selection” menu, after which a dialog box appears. The dialog box has two tabs: “Headword parameters” and “Explanatory part parameters” (Fig. 2).


Selecting parameters for headwords
Headwords parameters
Explanatory part parameters
Reset parameters
Cancel
Form the Sample
Selecting parameters for headwords
×
Headwords parameters
Explanatory part parameters
Reset parameters
Cancel
Form the Sample

Number of Definitions
Omit ☑
Condition: ≥ 0

Number of collocations of 1st type(n.+adj.)
Omit ☑
Condition: ≥ 0

Number of collocations of 2nd type(other)
Omit ☑
Condition: ≥ 0

Number of reference collocations
Omit ☑
Condition: ≥ 0

Headword description
Omit ☑
Headword masculine □
Headword feminine □
Reference entry (masculine) □
Reference entry (feminine) □
Undefined □

Headword structure
Omit ☑
Word □
Collocation □
Morpheme □
Undefined □

Headword type
Omit ☑
Foreign word □
Abbreviation □
Acronym □
Other □

Homonymy
Omit ☑
Condition: ≥ 0

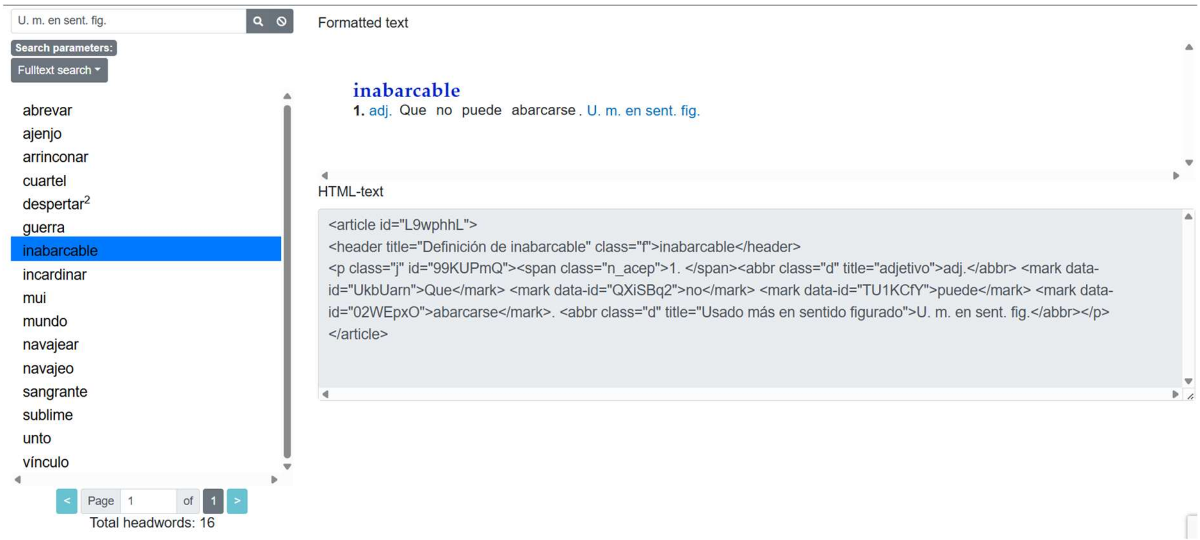
Figure 2: Dialog box to select dictionary entries with specific parameters.

Samples in the VLL DLE 23 can be considered as sub-dictionaries. The available tools allow you to create the following dictionaries:

- Morpheme dictionary;
- Homonym dictionary;
- Latin expression dictionary;
- Acronym and abbreviation dictionaries;

- Phrase dictionaries (of various types);
- Dictionaries of foreign words and foreign phrases;
- Common gender dictionary;
- Dictionary of unambiguous words;
- Dictionary of polysemous words.

The **full-text search mode** is effective when you need to select dictionary entries by certain meta-language elements of DLE 23: various labels, symbols that make up additional comments ("U. t. c. s."), meta-language markers ("Orth.", "Conjug. c."), etc. In addition, the search text string can contain both the text of a dictionary entry and HTML code elements ("<abbr title="Usado solo en sentido figurativo">") from the text field. In the current version of VLL, full-text search combined with the sampling tool is a very powerful tool for linguistic research. The figure 3 shows the example of the entry samples, the headwords of which are mostly used in figurative meaning.



U. m. en sent. fig.

Search parameters:

Fulltext search

abrevar
ajenjo
arrinconar
cuartel
despertar²
guerra
inabarcable
incardinar
mui
mundo
navajear
navajeo
sangrante
sublime
unto
vínculo

Page 1 of 1

Total headwords: 16

Formatted text

inabarcable
1. adj. Que no puede abarcarse . U. m. en sent. fig.

HTML-text

```
<article id="L9wphhL">
<header title="Definición de inabarcable" class="f">inabarcable</header>
<p class="j" id="99KUPmQ"><span class="n_acep">1. </span><abbr class="d" title="adjetivo">adj.</abbr> <mark data-
id="UkbUarn">Que</mark> <mark data-id="QXiSBq2">no</mark> <mark data-id="TU1KCfY">puede</mark> <mark data-
id="02WEpxO">abarcarse</mark>. <abbr class="d" title="Usado más en sentido figurado">U. m. en sent. fig.</abbr></p>
</article>
```

Figure 3: Entry samples, the headwords of which are mostly used in figurative meaning.

6. Conclusions

In today's world, information is becoming one of the most valuable resources. In this focus we consider linguistic information as well. The most authoritative sources of such information are explanatory dictionaries of national languages. The linguistic paradigm shift brings these dictionaries into priority sources for extracting linguistic knowledge. Traditional formats of lexicographic data representation are not suitable for deep analysis. In addition, the problem of data visualization remains open: presenting the results of data analysis in an understandable and visual way.

An explanatory dictionary is one of the most complex lexicographic products. Our research is based on a lexicographic data model. It performs the following functions: it sets the algorithm for parsing the text of a dictionary entry, it is used to form both the XML schema and the database schema. We believe that the presentation of lexicographic information in XML and text-oriented database formats will be effective for the further development of data analysis. Such a database is the basis for the VLL, which in our case serves as a tool for data analysis and visualization. Today, the VLL performs the following functions:

- inventory of registered words that meet the set parameters (specific word, foreign word, morpheme, abbreviation, phrase, homonymy, polysemy);

- studying the linguistic features of register words in the text of a dictionary entry. This makes it possible to identify regularities in the Spanish language that are implicit in the dictionary;
- statistical studies that demonstrate the frequencies of the studied linguistic phenomena (for example, the ratio of native and borrowed vocabulary, etc.).

Based on these studies, the user can draw certain conclusions about the lexical-semantic, etymological, grammatical and pragmatic features of Spanish language units. It is planned to expand the toolkit to provide access to any structural element of a dictionary entry by various parameters and to provide the ability to output lexicographic data in XML format.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. John, Kelleher and Brendan Tierney. Data science, The MIT Press, Cambridge, MA, 2018.

[2] L. Cox, The Model Thinker: What You Need to Know to Make Data Work for You by Scott E. Page, Basic Books, New York, NY, 2019.

[3] R. Mitkov, The Oxford Handbook of Computational Linguistics, 2nd ed, Oxford Handbooks, Oxford, 2022.

[4] V. Shyrokov, Language. Information. System, Akademperiodyka, Kyiv, 2021. doi:10.15407/academperiodyka.451.160.

[5] J. Casares, Nuevo concepto del diccionario, Editorial CSIC, Madrid, 1992.

[6] Diccionario de la lengua española. URL: https://dle.rae.es/.

[7] Oxford English Dictionary (OED). URL: https://www.oed.com/.

[8] Dictionary of Ukrainian in 20 volumes (SUM-20), Volumes 1–15, Ukrainian Lingua Information Fund NAS of Ukraine, Kyiv. URL: https://sum20ua.com/.

[9] L. Trap-Jensen, Lexicography between NLP and linguistics: aspect of theory and practice, in: J. Čibej, V. Gorjanc, I. Kosem, S. Simon Krek (Eds.): Lexicography in Global Contexts, Proceedings of the 18th EURALEX International Congress 2018, 17-21 July Ljubljana, 2018, pp. 25–38.

[10] F. Zahra Belkadi, R. Esbai, A Model-Driven Engineering: From Relational Database to Document-oriented Database in Big Data Context, in: Proceedings of the 16th International Conference on Software Technologies, ICSOFT 2021, pp. 653-659. doi: 10.5220/0010604906530659.