Exploratory and Text Searching Support in the *Dictionary of the Spanish Language*

Jordi Porta-Zamorano
Centro de Estudios de la Real Academia Española
E-mail: porta@rae.es

Abstract

Online dictionaries try to include search capabilities to meet most users' needs. Although users are not always aware of how to effectively use dictionaries, sometimes it is the interface that does not facilitate a friendly access to the dictionary information. This work aims at lowering the barrier in supporting onomasiological and semasiological advanced searches to the *Diccionario de la lengua española* (DLE) by combining text searches and faceted navigation into a user-friendly dictionary interface, allowing even non-experts to move through the dictionary in a natural and flexible manner. However, since the DLE is an electronic version of a printed dictionary, it contains related and unrelated abbreviations condensing different information that have to be properly converted into the set the facets and values provided by the search system.

Keywords: Dictionary interfaces, dictionary searching, online dictionaries

1 Introduction

The *Diccionario de la lengua española* (DLE) is the 23rd edition of the monolingual general Spanish dictionary produced and published by the *Real Academia Española*[1] (RAE) and the *Asociación de Academias de la Lengua Española*[2] (ASALE). Its basic online edition[3] receives sixty million look-ups per month on average, and provides several search facilities such as lemma and multi-word autocompletion, inflected forms look-up and linguistically-motivated approximate searches implementing an orthophonographic variation model and linking some derivative forms not explicitly registered into the dictionary with its stems. Searches by prefix, suffix, infix, and anagrams are also offered. In addition to these search capabilities, definitions and examples have been lemmatized to provide textual navigation. However, the mentioned search capabilities work in one direction, namely, from words to definitions, supporting only semasiological searches. This paper presents the Advanced DLE, a new interface and search engine to the DLE allowing textual searches and faceted navigation, which can be found within the Enclave RAE Platform[4].

2 Textual Search and Faceted Navigation

Faceted navigation is arguably the most significant innovation in search patterns of the last few decades, which has become nearly ubiquitous in e-commerce (Tunkelang 2009). However, the most common design in dictionary interfaces makes a distinction between basic and advanced searches with a parametric design. In such parametric search interfaces, the user selects all the information in

1 http://www.rae.es
2 http://www.asale.org
3 http://dle.rae.es
4 https://enclave.rae.es

one shot, using *a priori* filters, sometimes combining them using Boolean expressions, and reaching a dead end when selecting unsatisfiable combinations of constraints. By contrast, faceted navigation addresses the weaknesses of conventional parametric approaches by offering a progressive query refinement which also allows users to explore and discover new information.

The Advanced DLE interface allows users to start a query by typing one or more terms into a search box and then narrow the search by iteratively selecting facet values or a textual zone of the microstructure where they want to focus their search. Facet values and text zones give also counts, giving users an overview of the distribution of the selected subset of senses. These counts are dynamically updated with every selection. Figure 1 shows a search where *herramienta* (tool) and *madera* (wood) were typed and noun category was selected. Note that there are nine senses matching these criteria, and that search terms appear only within their definition.

Alternatively, users can start searches by selecting facet values without typing any term. The results area of the interface shows the alphabetically ordered list of all the senses matching the selected criteria. Initially, the results list contains all the dictionary senses. Selected facet values or the textual zone can be deselected at any time widening the results list. Figure 2 shows all verbal senses in the field of information technologies coming from French.

Selections are interpreted using the Boolean model, which is not exposed to users. Selecting values from different facets produces an AND between facets and selecting a textual zone produces an AND across facets.

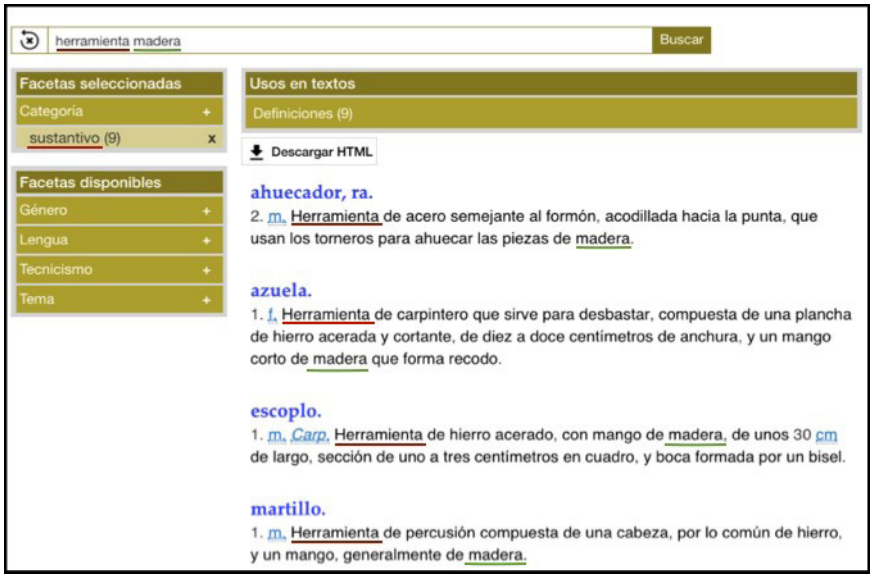

Figure 1: Nouns containing in the definition the terms *herramienta* (tool) and *madera* (wood).

Figure 2: Verbal senses from French in the field of information technologies.

Textual zones can be referred and combined in the search box using a syntax similar to that of Google Advanced Search. Every textual zone is defined as an "advanced operator" and alternative values can be expressed by means of the *O* (OR) operator. Using these operators, one can search words defining plants and fruits, beginning with *al-*, and coming from Arabic, just by typing *definición:planta O fruta O fruto etimología:árabe lema:al** into the search box.

3 From Abbreviations to Facets

Sense facets are an orthogonal set of categories representing the information conveyed mainly by abbreviations and tags of diverse types (etymological, grammatical, geographic, register, semantic, etc.). This information is coded within senses or inherited from its containing entry. There are a total of seventeen facets and 449 values related to senses.

However, because the DLE is an electronic version of a printed dictionary, senses contain related and unrelated abbreviations condensing different information. Unfortunately, abbreviations and facets do not always maintain a straightforward correspondence, and facets extracted from sets of abbreviations must be bundled.

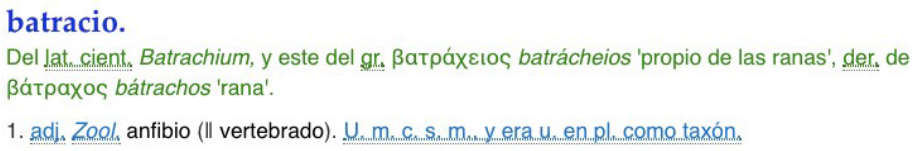

Figure 3: Entry for *batracio* (batrachian).

As an example of these correspondences, a sense like the one in *batracio* (batrachian), shown in Figure 3, contains the following grammatical, domain and usage abbreviations:

- adj. (adjective)
- Zool. (Zoology)
- U. m. c. s. m. y era u. en pl. como taxón (most commonly used as a masculine noun and was used in plural as a taxon)

which generates the following facets bundles:

- {category: adjective, domain: zoology}
- {category: noun, gender: masculine, domain: zoology}
- {category: noun, number: plural, domain: zoology, usage: obsolete}

This bundling scheme avoids retrieving the sense 1 in *batracio* when selecting obsolete adjectives, since this information does not co-occur in the same bundle.

4 Indexing Textual Zones

Textual zones correspond to four different structural parts of a sense in the DLE: headword, etymology, definition, and examples. The first two zones are inherited from the containing entry. Selecting a textual zone restrict the results list to senses containing the search terms in that specific zone. Figure 4 shows how many senses have textual zones containing the term *diccionario* (dictionary): there are twenty-one senses having *diccionario* in *Definiciones* (definitions), three in *Ejemplos* (examples), two in *Lemas* (headwords) and two in *Etimologias* (etymologies).

Figure 4: Textual zones containing the term *diccionario* (dictionary).

Headwords such as *actor*[1]*, triz* are converted to *actor* (actor) and *actriz* (actress), and text in definitions and examples are tokenized before indexing. For etymologies, all the abbreviations are expanded and the information about its language family is inserted into the textual index. By doing so, a user can find senses with lemmas coming from the Latin word *ferrum* or demonyms, just by searching place names. Users can also find senses with lemmas from any Indo-European language in its etymology, even if it is not explicitly mentioned, just searching with wildcards the term *indoeurope** (Indo-European).

5 Search Engine Implementation

The search engine has been implemented on top of SWI-Prolog (Wielemaker et al. 2012). Prolog programs describe relations, defined by means of clauses, and computations are initiated by running a query over these relations. Prolog is particularly well-suited for in-memory databases or declarative knowledge-based applications with inference capabilities.

Facets bundles and inverted indices for textual searches are represented in RDF (Resource Description Framework). SWI-Prolog integrates core packages for efficient main-memory RDF storage and querying (Wielemaker et al. 2003). However, faceted search is computationally demanding, and every time a user makes a selection, facet values, zones and counts have to be recomputed. For this reason, some simple queries involving the selection of facet values with the higher number of senses are precomputed and memoized when the server is started up.

A parser that builds a term representing queries using the advanced operators described at the end of Section 2 has been implemented with a simple definite clause grammar. These query terms are evaluated making intensive use of Prolog higher-order predicates. In addition, SWI-Prolog provides packages for indexing XML and gives support to concurrent HTTP requests and JSON. These packages have been used to provide a RESTful web service access to the dictionary (Wielemaker 2014). The complete backend has been implemented in a compact Prolog program of about six hundred lines of code.

6 Conclusions and Future Work

Preliminary evaluation from a reduced group of language specialized users revealed a high degree of satisfaction with the new search and exploration possibilities offered in the interface of the Advanced DLE. However, for some queries, users could prefer counts or results to be referenced to entries

instead of senses. The possibility to offer both counts and to group senses could satisfy users with different points of view. Other ways of querying the dictionary, using a form instead of advanced operators, seem to be a natural extension to the interface. Finally, some users have suggested new facets for recreational uses, such as the number of letters in the lemma.

References

RAE & ASALE. (2014). *Diccionario de la lengua española*. Espasa, Madrid, 23th ed.

Tunkelang, D. (2009). *Faceted Search*. Systhesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

Wielemaker, J. (2014). SWI-Prolog version 7 extensions. *Proceedings of the Workshop on Implementation of Constraint and Logic Programming Systems and Logic-based Methods in Programming Environments (CICLOPS-WLPE'14)*.

Wielemaker, J., Schreiber, G. & Wielinga, B. (2003). Prolog-based infrastructure for RDF: performance and scalability. *Proceedings of the 12th International Semantic Web Conference (ISWC'03)*, pp. 644-658.

Wielemaker, J., Schrijvers, T., Triska, M. & Lager, T. (2012). SWI-Prolog. *Theory and Practice of Logic Programming* 12(1-2), pp. 67–96.