AN ONLINE SPANISH LEARNERS' DICTIONARY: THE DAELE PROJECT

Blanca ARIAS-BADIA, Elisenda BERNAL
InfoLex Research Group, Universitat Pompeu Fabra, Barcelona, Spain

Araceli ALONSO
LiCoRN Research Group, Université de Bretagne-Sud, Lorient, France

Arias-Badia, B., Bernal, E., Alonso, A. (2014): An Online Spanish Learners' Dictionary: The DAELE Project. Slovenščina 2.0, 2 (2): 53–71.
URL: http://www.trojina.org/slovenscina2.0/arhiv/2014/2/Slo2.0_2014_2_05.pdf.

Current online dictionaries for learners of Spanish as a second language largely just reproduce their printed counterparts. This report summarizes the advances made in the DAELE, a prototype dictionary for learners of Spanish as a second language that has been designed solely for online publication. The purpose and main features of the macro- and microstructure of the dictionary are briefly described, as well as the methodology whereby the data are collected and the first steps taken to include the most significant collocations by applying a method based on collocational networks.

Keywords: DAELE, Spanish as a foreign language, learners' dictionary

1 INTRODUCTION

The DAELE (*Diccionario de aprendizaje de español como lengua extranjera*, i.e. Spanish learners' dictionary) is an ongoing project which pursues the building of an electronic monolingual dictionary for learners of Spanish as a foreign language at intermediate and advanced language levels. The InfoLex research group from the University Institute for Applied Linguistics (IULA) at the Universitat Pompeu Fabra (Barcelona) has developed the dictionary within the framework of three research projects: the first one devoted to the representation of verbs and adverbs; the second one to nouns and adjectives;

and finally, the third one to lexical combinatorics[1]. What follows is a description of the main features of the DAELE and a brief consideration of the stage of development of this three-part project.

2 COLLECTION OF DATA

In the first instance, the research focused mainly on verbs and nouns, leaving adverbs and adjectives for the second stage. This decision was based on the hypothesis that understanding the behaviour of verbs and nouns and establishing the premises for representing them in a learners' dictionary would facilitate the representation of adverbs and adjectives.

The DAELE is a bottom-up project, that is, it is a 100% corpus-based dictionary that builds up meaning representation from scratch, in contrast with most Spanish dictionaries. For all four categories (verbs and adverbs, nouns and adjectives), the lexicographic corpus of analysis is composed of data from the IULA50, the general-domain texts from the *IULA Corpus Tècnic* (CT-IULA)[2] and texts from the press, the *Spanish Web Corpus* included in the Sketch Engine® [3](SWC), Mark Davies' *Corpus del Español*[4] (CEsp), and the *Corpus de Referencia del Español Actual*[5] (CREA).

The data are analysed by focusing on the following aspects:

---

[1] The three projects involved in the development of the DAELE have been the following: (1) Las categorías verbo y adverbio en el DAELE, funded by the Spanish Ministry of Education and Culture (HUM2006-06982/FILO) and Diccionario de aprendizaje de español como lengua extranjera, funded by Fundación Comillas. Project Leader: Dr. Paz Battaner; (2) Las categorías nombre y adjetivo en el DAELE, funded by the Spanish Ministry of Education and Culture (HUM2006-07898/FILO). Project Leader: Dr. Janet Decesaris; and (3) La representación de la combinatoria léxica en los diccionarios de aprendizaje: nuevos métodos para nuevos diccionarios, funded by the Spanish Ministry of Economy and Competitiveness (FFI2012-37654). Project Leader: Dr. Sergi Torner.

[2] http://bwananet.iula.upf.edu

[3] http://www.sketchengine.co.uk

[4] http://www.corpusdelespanol.org

[5] http://corpus.rae.es/creanet.html

- semantic roles and syntactic order;
- syntactic structure;
- detection of grammatical subcategories;
- variety of meanings of each word;
- phrases.

The software used in the compilation and processing of lexicographic data is TshwaneLex®[6] (TLex), a dictionary-writing system that enables the user to export the database to any format required (MS-Word, HTML, etc.) for its immediate publication. Thus, the website which gives access to a sample of the dictionary, showing the first project results, is easily updated every time a word class is defined. At the moment, a total of 365 verbs are available in the online version. The prototype dictionary is accessible at http://www.daele.eu.

3 THE HEADWORD LIST

The first headword list, which covers verbs, was established on the basis of frequency and use in the above mentioned corpora. The defining verbs listed by Soler (2006) were also taken into account. The lexical set of verbs, therefore, includes the most frequent 1,380 verbs in Spanish, which, in many cases, are also polysemous units. Among these verbs, groups have been formed to account for hypernym-hyponym relations, which allow the editors to work with verbs belonging to the same class in a unified, complementary manner and write the subsequent definitions – thus, for example, the group *levantar* ('lift, raise') includes the verbs *gritar* ('shout'), *incorporar* ('include'), *levantar* ('lift, raise'), *madrugar* ('rise early'), *realzar* ('give lift to'), *rebelar* ('rebel, rise up') and *suspender* ('hang, suspend'). For an account of the semantic grouping of verbs, see Renau and Battaner (2010).

Some research related to adverbs has already been developed and criteria for compiling the first headword list of adverbs have been defined following the same criteria applied to verbs; see Torner (2013) for more information.

[6] http://tshwanedje.com/tshwanelex

However, this is a work in progress and a definitive headword list is not yet available.

The procedure for determining which nouns and adjectives are to be headwords has been slightly different to that of verbs. The first list was established by comparing four lists extracted from four different sources: two of them were based on corpora and frequency of use; one was based on a lexical availability index; and the last one was taken from a monolingual Spanish dictionary for native-speaking children: *A Frequency Dictionary of Spanish* [FDS] (Davies 2006); *The corpus PAAU 1992* [C92] (Torner and Battaner 2005); *El Léxico disponible de los estudiantes preuniversitarios de la provincia de Jaén* [LDA] (Ahumada 2006); and *El Diccionario de primaria de la lengua española* [DPVOX] (Battaner 2004).

As a selection criterion, frequency was initially considered. After an analysis of the most frequent words extracted from the two sources based on corpora, it was observed that many nouns and adjectives from several semantic fields were not included. Therefore, the decision was made to cross the nouns and adjectives from this list with those from the other two sources. All four sources were supposed to contain the most relevant words in the Spanish language. While we expected to find a high number of matches, the comparison revealed many differences in the distribution of both word classes. Only 836 lemmata out of a total of 16,176 were found in all four sources, and 9,107 were found in only one source – see Alonso and Estremera (2009) for a detailed description. After considering the lemmata found in up to three of the sources, a total of 7,069 nouns and adjectives were chosen. Nouns are being semantically classified, and the classification system is being created as work proceeds. The ad-hoc system is expected to be useful for descriptors and definition patterns, and the establishment of relationships between lemmata.

4 DICTIONARY ENTRY STRUCTURE

The DAELE was conceived not only as a tool to help the user decode language,

but also as an encoding aid. Different layers of information allow various description levels. These may be displayed depending on the needs of each user. Special attention has been paid to contextual features, as well as to syntactic and pragmatic aspects, in order to facilitate the encoding process for a non-native speaker of Spanish. One of the most innovative features of the dictionary lies in the entry structure: headwords' senses are firstly organized according to semantic criteria by using signposts which specify the various senses of the headword and the semantic relations between senses (Battaner 2010). This is bound to contribute to a better understanding of meanings on the part of the learner.

As explained by Renau and Battaner (2011), the DAELE is being created from scratch in terms of both contents and format, and was inspired by the ideas developed in the COBUILD project by Sinclair and Hanks (1987). The entry structure has been conceived by taking into account the possibilities of Web technology.

A detailed description of the macrostructure and microstructure of the DAELE is included in Battaner and Renau (2010) and Renau and Battaner (2011), while the list of publications resulting from the project includes reports on several features of the dictionary. Figure 1 shows the structure and components of a verbal entry in the DAELE:

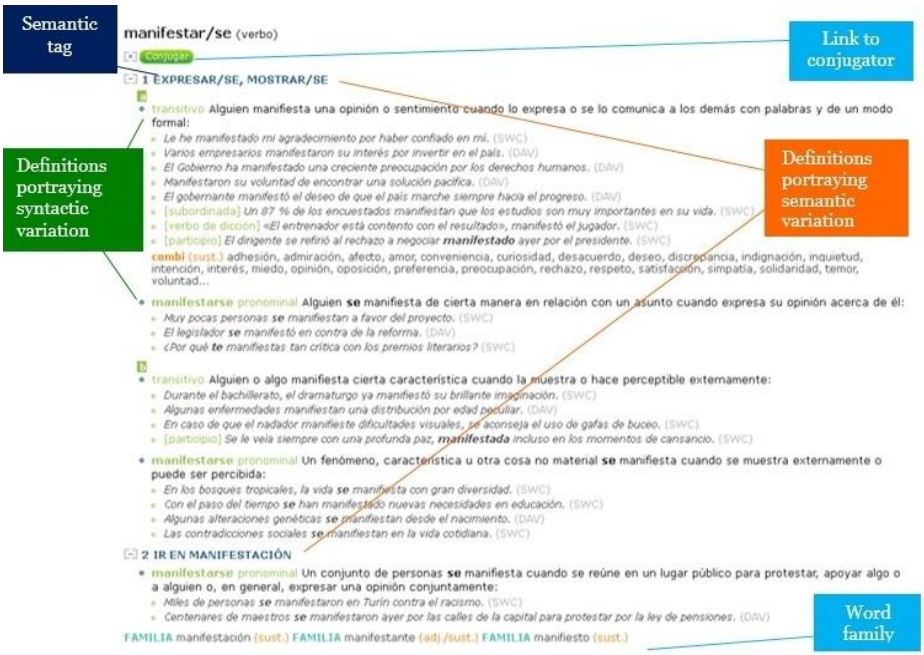


Figure 1: Parts of a verbal entry in the DAELE.

The entry structure for nouns is based on the same principles as the entry designed for verbs. A report on the advances made on nouns and adjectives can be found in Mahecha Mahecha and Decesaris (2011). For adverbs, initial insights are explained in Torner (2013).

The following sections concentrate on three of the most salient features of the DAELE: representation of meaning, examples and collocations.

5 REPRESENTATION OF MEANING

For each entry, two sets of data related to meaning are provided: (i) the semantic tag, whereby the general meaning of the word is expressed and under which the user finds every definition that falls within that word meaning, and (ii) the definition, which covers syntactic aspects (such as argument structure construction(s) and syntactic variation) by using the word as part of a full

sentence definition that may include prepositions, pronouns or any elements that need to be pointed out to the learner. Like in some English dictionaries for learners, such as the *Macmillan English Dictionary for Advanced Learners* (MED) (Rundell 2002), this kind of definition has been selected because it fulfils the decoding function and enables the users to detect syntactic patterns as they learn the meaning of the unit, which is much more difficult with traditional definitions in monolingual dictionaries.

All definitions are corpus-driven in the sense that they are constructed by taking into account the different usage patterns found in the corpora. For the analysis of corpus data and in order to establish the different 'normal' usage patterns of a verb, the method of Corpus Pattern Analysis (CPA) proposed by Patrick Hanks and based on the *Theory of Norms and Exploitations* (TNE) (Hanks 2004, 2013) was applied. CPA has been firstly applied to Spanish for representing specialised lexical units in general dictionaries (Alonso 2009). It was in the framework of the DAELE project that it was fully tested and adapted to represent verbs in Spanish. In the case of the DAELE, it must be noted that CPA was originally conceived for the compilation of a Pattern Dictionary of English Verbs[7] for native speakers of English, and not as a dictionary for learners. Some lexicographic decisions were taken to convert the CPA pattern to a full definition designed to explain not only the meaning of the headword in one particular sense but the use of the word in context.

The focus of CPA is on determining the prototypical syntagmatic patterns in which words in use are associated. CPA facilitates the analysis of corpus data in a systematic way, providing a formalism to systematically describe each of the patterns found in the corpus. It clarifies many aspects of word usage which must be taken into account in the definition. CPA helps to eliminate ambiguity and to develop the dictionary entry. For a full description of the CPA procedure and its adaptation to the Spanish language, see Renau (2012) and Renau and

7 http://www.pdev.org.uk

Battaner (2012). A thorough description of the different steps to construct each definition in the DAELE from CPA can be found in Renau and Alonso (forthcoming). Figure 2 shows an example of the construction of the definition for the verb *alimentar* ('to feed') in the DAELE:

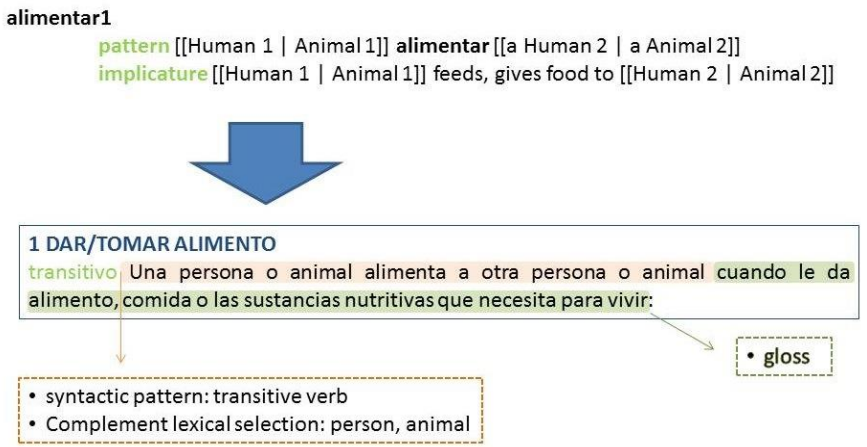


Figure 2: Construction of a definition in the DAELE from CPA.

6 EXAMPLES

With regard to examples, it has been determined that each definition should include a minimum of one example extracted from the corpus. Even though there is not a required maximum of examples, since the amount of examples needed depends on the complexity of each word sense and subsenses, three examples typically suffice. Some examples are introduced by tags (mostly pointing at syntactic features) which describe how the word is used in a specific example, or which provide information about its frequency. After each example and between brackets, there is a reference to the corpus from which the example was taken. Figure 3 shows some of the examples for *manifestar/se*.

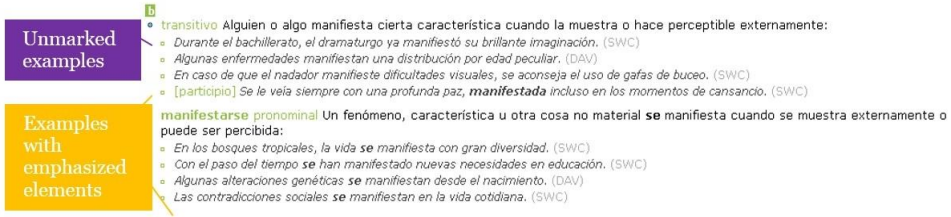


Figure 3: Examples in the DAELE (close-up from Figure 1).

7 COLLOCATION IN THE DAELE

Recently the project has dealt with the phenomenon of collocation, i.e. 'lexical co-occurrence of words' (Sinclair 1991: 170). How words combine in collocations is one of the main difficulties for a learner of the language and questions arise about their use. Collocations have received minimal attention in the Spanish lexicographic tradition. Most Spanish dictionaries do not include information about collocations, or the information provided is so limited that it fails to meet the encoding needs of learners.

A database of the collocates of 570 headwords belonging to four semantic fields – i.e. Family relationships, Social relationships, Information and Communication, and Verbal expression – proposed by the Instituto Cervantes[8] has been compiled and is currently under analysis in order to establish which collocates are most useful for the dictionary user. See Table 1 to find out more about the statistics from this database.

[8] The Instituto Cervantes is a non-profit organisation funded by the Spanish Government. It promotes the study of the Spanish language and culture and is responsible for the design of the Diploma de Español como Lengua Extranjera (DELE), an international Spanish language qualification.

Semantic field (Instituto Cervantes)	***Collocation nodes***
Family relationships	174
Social relationships	88
Information and Communication	123
Verbal expression	185
TOTAL	570

Table 1: Number of collocation nodes for each semantic field in the database.

The database includes information about the collocates found for each of the 570 words in four different sources, consisting of a corpus and three dictionaries – two of them are collocation dictionaries, the third one is a dictionary focused on language use: (a) *Spanish Web Corpus* [SWC]; (b) *Diccionario Combinatorio Práctico* [Práctico]; (c) *Redes*; and (d) *Diccionario de Uso del Español* [DUE]. As suggested above, the next step in the process will be to determine which of the units found should enter the DAELE as useful collocates for learners of Spanish, which semantic and combinatorial features of every lexical unit are lexicographically relevant, and how to incorporate the information in the dictionary.

In order to clarify these aspects, the mechanism of collocational networks is being tested. This method was proposed by Williams (1998) and applied to the DicSci project for the compilation of an *E-Advanced Learner's Dictionary of Verbs in Science* (Williams and Millon 2010, Alonso et al. 2011). As defined by Williams (1998: 156), collocational networks are 'webs of interlocking conceptual clusters realised in the form of words linked through the process of collocation'. Collocational networks are statistically based and corpus-driven. Therefore, the resulting collocational chains depend on the nature of the corpus being used and the statistical techniques applied. Research is currently under way to clarify which statistical measures are best-adapted to calculating the

most significant collocates of the node to be included in a learners' dictionary for upper-intermediate users. Another aspect under scrutiny relates to the criteria for excluding or including units into the network. Figures 4 and 5 illustrate a prototype of the collocational network for *manifestar*:

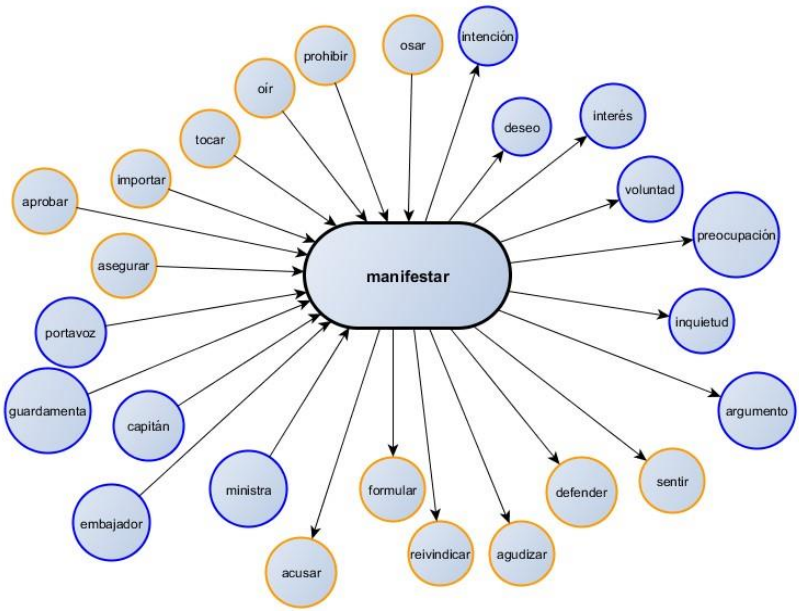


Figure 4: First level of the collocational network for *manifestar*.

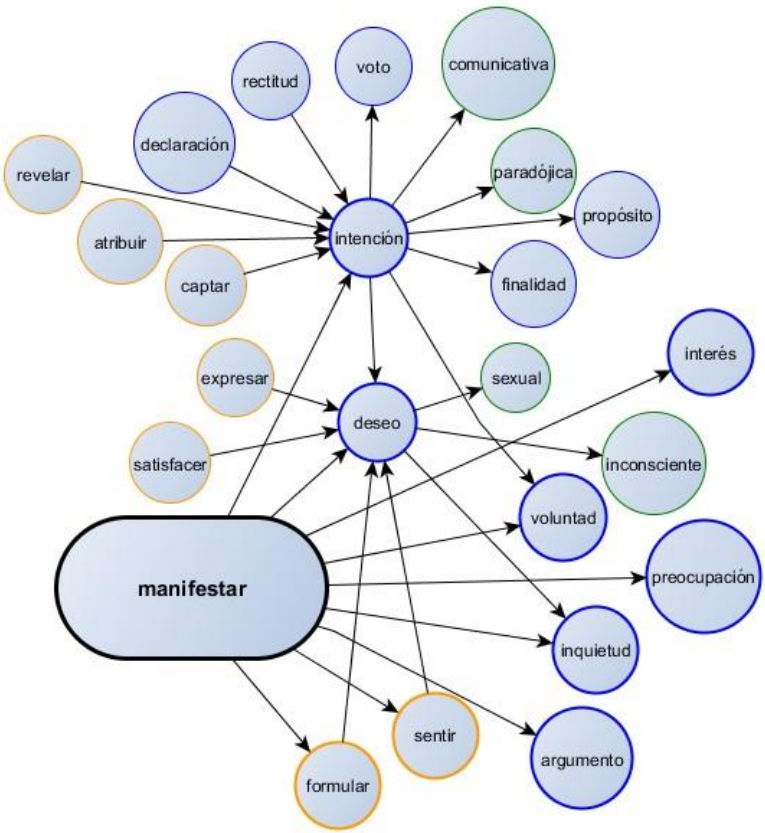


Figure 5: Partial illustration of the first and second level of the collocational network for *manifestar*.

8 FUTURE WORK

The DAELE is an online corpus-driven learners' dictionary prototype whose conception derives from a practical analysis of language in use. In the previous sections, the building method of the dictionary has been presented, describing the main features and components of the dictionary. While many advances have been made on verbs in the DAELE, there are still aspects to be solved related to the other three word classes. As an ongoing project, the DAELE is not a finished work and some improvements in the interface and compilation process are planned for the near future.

The DAELE aims to contribute to the development of e-lexicography in Spanish, clearing a path for building reference tools for non-native speakers of Spanish.

REFERENCES

Dictionaries:

DAELE = *Diccionario de Aprendizaje del Español como Lengua Extranjera*. P. Battaner, S. Torner (eds.). Available at: http://www.daele.eu (20 October 2014).

DPVOX = *Diccionario de primaria de la lengua española* (3rd ed.) (2004). P. Battaner (ed.). Barcelona: Anaya.

DUE = *Diccionario de uso del español* (2nd ed.) (2006). M. Moliner. Madrid: Gredos.

FDS = *A frequency dictionary of Spanish*. (2006). M. Davies. New York/London: Routledge.

LDA = El léxico disponible de los estudiantes preuniversitarios de la provincia de Jaén (2006). I. Ahumada. Jaén: Servicio de Publicaciones de la Universidad de Jaén.

MED = *Macmillan English Dictionary for Advanced Learners* (1st ed.) (2002). M. Rundell (ed.). Oxford: Macmillan Education.

PDEV = *Pattern Dictionary of English Verbs* (in progress). P. Hanks (ed.). Available at: http://www.pdev.org.uk (19 October 2014).

Práctico = *Diccionario combinatorio práctico del español contemporáneo* (2006). I. Bosque (ed.) Madrid: SM.

Redes = *Redes. Diccionario combinatorio del español* (2004). I. Bosque (dir.) Madrid: SM.

Corpora:

C92 = *El corpus PAAU 1992: estudios descriptivos, textos y vocabulario* (2005). S. Torner and P. Battaner (eds.). Barcelona: Universitat Pompeu Fabra. IULA, Documenta Universitaria. Available at: http://www.iula.upf.edu/rec/corpus92 (20 October 2014).

CREA = *Corpus de referencia del español actual*. Real Academia Española. Available at: http://www.rae.es (20 October 2014).

CEsp = *Corpus del Español* (2002). M. Davies. Available at: http://www.corpusdelespanol.org (20 October 2014).

CT-IULA = *Corpus Tècnic de l'IULA*. University Institute for Aplied Linguistics, Universitat Pompeu Fabra. Available at: http://bwananet.iula.upf.edu (20 October 2014).

SWC = *Spanish Web Corpus*. Available at: www.sketchengine.co.uk (20 October 2014).

Books and articles:

Alonso, A. (2009): *Características del léxico del medio ambiente en español y pautas de representación en el diccionario general*. PhD dissertation. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Alonso, A. and Estremera, E. (2009): Corpora and headword lists for learner's dictionaries: the case of the Spanish Learners' Dictionary DAELE (*Diccionario de aprendizaje del español como lenguja extranjera*). Working paper from research project HUM2006-07898/FILO. IULA, Universitat Pompeu Fabra.

Alonso, A., Millon, C. and Williams, G. (2011): Collocational networks and their application to an E-Advanced Learner's Dictionary of Verbs in Science (DicSci). In I. Kosem and K. Kosem: *Electronic Lexicography*

in the 21st century: New Applications for New Users. Proceedings of eLex 2011, Bled, 10-12 November 2011. 12-22. Ljubljana: Trojína, Institute for Applied Slovene Studies.

Battaner, P. (2010): El uso de etiquetas semánticas en los artículos lexicográficos de verbos en DAELE". In *Quaderns de filologia. Estudis lingüístics* 15. 139-158.

Battaner, P. and Renau, I. (2010). El proyecto DAELE verbos: un diccionario de aprendizaje de español como lengua extranjera. In *II Encuentros (ELE) Comillas: el profesor de ELE: metodología, técnicas y recursos para el aula*, Comillas, 18-20 November 2010. 142-150. Comillas, Santander: Fundación Comillas.

Hanks, P. (2004): The Syntagmatics of Metaphor and Idiom. *International Journal of Lexicography* 17 (3). 245-274.

Hanks, P. (2013): *Lexical Analysis. Norms and Exploitations*. Cambridge, MA: MIT Press.

Mahecha Mahecha, V. and DeCesaris, J. (2011): Representing nouns in the "Diccionario de aprendizaje del español como lengua extranjera" (DAELE). In I. Kosem and K. Kosem: *Electronic Lexicography in the 21st century: New Applications for New Users. Proceedings of eLex* 2011, Bled, 10-12 November 2011. 180-186. Ljubljana: Trojína, Institute for Applied Slovene Studies.

Renau, I. (2012): *Gramática y diccionario: Las construcciones con* se *en las entradas verbales del diccionario de español como lengua extranjera*. PhD dissertation. Barcelona: Institut Universitari de Lingüística Aplicada. Universitat Pompeu Fabra.

Renau, I. and Alonso, A. (forthcoming): Systematising corpus-based definitions in second language dictionaries.

Renau, I. and Battaner, P. (2010): Agrupación semántica de verbos de un

DAELE a través de sus relaciones hiperonímicas. In M. A. Castillo Carballo and J. M. García Platero (coords.). *La lexicografía en su dimensión teórica*. 543-560 Málaga: Servicio de Publicaciones de la Universidad de Málaga.

Renau, I. and Battaner, P. (2011): The Spanish Learner's Dictionary DAELE on the Panorama of the Spanish E-lexicography. In I. Kosem and K. Kosem (eds.): *Electronic Lexicography in the 21st century: New Applications for New Users. Proceedings of eLex* 2011, Bled, 10-12 November 2011. 221-226. Ljubljana: Trojína, Institute for Applied Slovene Studies.

Renau, I. and Battaner, P. (2012): Using CPA to represent Spanish pronominal verbs in a learner's dictionary. In R. Vadvedt and M. Torjusen (eds.): *Proceedings of the 15th Euralex International Congress*, Oslo, 7-11 August 2012. 350-361. Oslo: University of Oslo.

Sinclair, J. (ed.) (1987): *Looking up: an account of the COBUILD Project in lexical computing and the development of the Collins COBUILD English Language Dictionary*. London: Collins.

Sinclair, J. (1991): *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Soler, J. (2006): *Definició lexicogràfica i estructura del diccionari*. Barcelona: Institut d'Estudis Catalans.

Torner, S. (2013): Suffix *-mente* adverbs in DAELE, a Spanish Learners' dictionary. *International Journal of Lexicography* 26 (4). 469-497.

Williams, G. (1998): Collocational Networks: Interlocking Patterns of Lexis in a Corpus of Plant Biology Research Articles. In *International Journal of Corpus Linguistics* 3 (1). 151-171.

Williams, G. and Millon, C. (2010): Going organic: Building an Experimental Bottom-up Dictionary of Verbs in Science. In A. Dykstra and T.

Schoonheim (eds.), *Proceedings of the 14th Euralex International Congress*. 1251-1257. Leeuwarden: Fryske Akademy.

Other publications resulting from the project:

Estremera, E. (2008): Etiqueta semántica (Proyecto nombres y adjetivos). Working paper from research project HUM2006-07898/FILO. IULA, Universitat Pompeu Fabra.

Battaner, P. and Renau, I. (2012): Verbo y diccionario de español para extranjeros. Las prestaciones en la codificación de verbos para el Diccionario de aprendizaje de español como lengua extranjera (DAELE). In H. E. Lombardini and M. E. Pérez Vázquez (coords.). *Núcleos: estudios sobre el verbo en español e italiano*. 77-97. Bern: Peter Lang.

Bernal, E. (2013): Los nombres del verbo: reflexiones acerca de las categorías gramaticales en los diccionarios. Propuesta de representación para el DAELE. In S. Torner and E. Bernal (eds.). *Los verbos en el diccionario*. 13-32. A Coruña: Universidade da Coruña.

Bueno, R. (2011a): El *Diccionario de aprendizaje de español como lengua extranjera* (DAELE): una nueva herramienta en el aprendizaje de LE. Una mirada hacia la representación de la categoría del verbo. In *Domínios de Lingu@gem* 5 (2). 202-216. Uberlândia: Universidade Federal de Uberlândia.

Bueno, R. (2011b): El diccionario de aprendizaje del español como lengua extranjera (DAELE): una herramienta para la enseñanza del verbo. In A. Ferrari et al. (coords.). *Horizontes de Brasil: escenarios, intercambios y diversidad*. 1948-1958. Barcelona: Ediciones APEC (Asociação de Pesquisadores Brasileiros da Catalunha).

Bueno, R. (2012): La representación lexicográfica de la subordinación verbal sustantiva en un DAELE: selección preliminar de verbos para su

estudio. In A. Nomdedeu, E. Forgas and M. Bargalló (eds.). *Avances de lexicografía hispánica: tomo II*. 191-204. Tarragona: Universitat Rovira i Virgili.

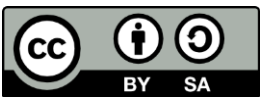
Renau, I. and Bernal, E. (2012): El Diccionario de aprendizaje de español como lengua extranjera (DAELE): primera prueba de usuarios. In A. Nomdedeu, E. Forgas and M. Bargalló (eds.). *Avances de lexicografía hispánica: tomo II*. 245-259. Tarragona: Universitat Rovira i Virgili.

SPLETNI SLOVAR ŠPANŠČINE ZA NEMATERNE GOVORCE: PROJEKT DAELE

Obstoječi spletni slovarji za tiste, ki se španščine učijo kot drugega jezika, so povečini zgolj tiskane različice, prenesene v elektronski medij. Prispevek prikazuje napredek pri projektu DAELE, katerega cilj je izdelava prototipa slovarja za nematerne govorce španščine, zasnovanega (zgolj) za spletno objavo. Opisani so ciljni uporabniki slovarja in razlogi za izdelavo, predstavljene so glavne značilnosti slovarske makrostrukture ter mikrostrukture. Poročilo predstavlja tudi način zbiranja gradiva in njegovo analizo, posebna pozornost pa je posvečena opisu vključitve najpomembnejših kolokacij v slovarska gesla z uporabo metode, ki temelji na kolokacijskih mrežah.

Ključne besede: DAELE, španščina kot tuji jezik, slovar za nematerne govorce

To delo je ponujeno pod licenco Creative Commons: Priznanje avtorstva-Deljenje pod enakimi pogoji 2.5 Slovenija.

This work is licensed under the Creative Commons Attribution ShareAlike 2.5 License Slovenia.

http://creativecommons.org/licenses/by-sa/2.5/si/