CAHIERS DU CENTAL 7

2010

eLexicography in the 21st Century: New Challenges, New Applications

Proceedings of eLex 2009, Louvain-la-Neuve, 22-24 October 2009

Sylviane Granger and Magali Paquot (eds)

UCL PRESSES
UNIVERSITAIRES
DE LOUVAIN

Cahiers du CENTAL

Place Blaise Pascal, 1 – B-1348 Louvain-la-Neuve
tél. : +32 10 47 37 86 – fax : +32 10 47 26 06
email : cental@tedm.ucl.ac.be – http://www.uclouvain.be/cental

Membre de l'Association des Revues Scientifiques et Culturelles – A.R.S.C. (http://www.arsc.be)

Graphisme de la couverture Olivier Vereecken http://aphine.com

Dépôt légal : D/2010/9964/10
ISBN 978-2-87463-211-2
ISSN : 1783-2845

Imprimé en Belgique

Diffusion :
www.i6doc.com l'édition universitaire en ligne

Sur commande en librairie ou à :
Diffusion universitaire CIACO
Grand-Place, 7
1348 Louvain-la-Neuve, Belgique
Tél. +32 10 47 33 78 - Fax +32 10 45 73 50
duc@ciaco.com

Diffusion pour la France :
Librairie Wallonie-Bruxelles
46 rue Quincampoix
75004 Paris
Tél. +33 1 42 71 58 03 - Fax +33 1 42 71 58 09
libwabr@club-internet.fr

Table of Contents

GRANGER, S.; PAQUOT, M. (EDS.). 2010. ELEXICOGRAPHY IN THE 21ST CENTURY: NEW CHALLENGES, NEW APPLICATIONS. PROCEEDINGS OF ELEX2009, LOUVAIN-LA-NEUVE, 22-24 OCTOBER 2009. CAHIERS DU CENTAL 7. LOUVAIN-LA-NEUVE, PRESSES UNIVERSITAIRES DE LOUVAIN, 369-374.

DiCE in the web: An online Spanish collocation dictionary

Margarita Alonso Ramos[1], Alfonso Nishikawa[1], Orsolya Vincze[1]
University of A Coruña

Abstract

DiCE is an online dictionary of Spanish collocations which provides semantic and combinatorial information of lexical units. The dictionary makes use of the typology of lexical functions (Mel'čuk *et al.* 1995), together with natural language glosses to describe the semantic content of collocates. With the aim of showing the different ways in which the database can be exploited, we present the organization of the online interface of the dictionary.

Keywords: collocation, online dictionary, lexical functions, Spanish as a foreign language

1. Introduction

In this paper we present the *Diccionario de Colocaciones del Español* (DiCE), a web-based collocation dictionary of Spanish that is being developed at the University of A Coruña (Alonso Ramos 2005). Collocations in DiCE are idiosyncratic combinations of two lexical units, the *base* and the *collocate*, as defined by Hausmann (1979) and others. DiCE is similar to dictionaries such as the *BBI* (Benson *et al.* 1986), the *LTP* (Hill and Lewis 1997) or the *Oxford Collocations Dictionary* (Crowther *et al.* 2002). However, unlike these English paper dictionaries, it has been conceived from the start as an electronic lexical database. This allows us to provide more information to the user and to implement a flexible means of access to this information[2].

As far as its theoretical framework is concerned, DiCE draws upon the fine-grained typology of *lexical functions* (LFs) introduced in the *Explanatory Combinatorial Lexicology* (Mel'čuk *et al.* 1995). However, users of the dictionary do no have to be familiar with this framework since the semantic content of LFs is paraphrased in natural language glosses.

2. The architecture of the dictionary

DiCE has been conceived as an electronic lexical database, a feature that makes it free from the alphabetical order of conventional dictionaries, given that the architecture of

[1] University of A Coruña, lxalonso@udc.es.

[2] DiCE is maintained in a MySQL database and is implemented in PHP using an Apache Server and the CakePHP environment.

an electronic dictionary is necessarily a network, not a list. Our environment is divided into two zones: the administration zone and the public zone. The first one is handled by the lexicographer; it is dedicated to the edition of lexicographic information contained in the microstructure and the macrostructure of the dictionary (*e.g.* semantic tags or the list of LFs). The public zone can be accessed freely by users. It consists of two main components: the dictionary itself and the advanced search component.

2.1. The dictionary component

We access the dictionary component through the list of lemmas. Each lemma is associated with a list of lexical units (LUs). For each LU, the user can look up the corresponding semantic or combinatorial information. As for the semantic information, the entry of each LU provides: a) a *semantic tag* that represents the generic meaning; b) the *actantial structure* representing the participants of the situation designated by the noun; c) corpus examples, most often derived from the online *Corpus of the Real Academia Española* (CREA); and d) quasi-synonyms and quasi-antonyms of the LU.

As for combinatorial information, we offer two sources of information: 1) the syntactic combinatory information of the LU is shown in the Government Pattern (*esquema de régimen)* section, where we specify the projection of its semantic valency structure onto its syntactic valency structure and, in addition, the subcategorization information associated with the latter, and 2) the lexical combinatory information is displayed in the section Collocations. In what follows, we focus on lexical combinatorics.

Taking a specific LU as the starting point, the user can choose between five different groups of lexical correlates:

> 1) Attributes of the participants: Under this heading, we have grouped those attributes or nouns that refer to the participants of the situation designated by the LU. For example, in the entry for ADMIRACIÓN ‘admiration’, the user finds *digno de admiración* ‘worthy of admiration’ or *admirable* ‘admirable’, both referring to the participant that can compel admiration;
>
> 2) LU + adjective. Here, the user finds adjectives that co-occur with the LU;
>
> 3) Verb + LU: In this section, we have grouped the verbs that take the LU as a direct complement or as a prepositional complement, *e.g. despertar antipatía* ‘[to] arouse dislike’;
>
> 4) LU + verb: This section contains verbs that take the LU as the grammatical subject, *e.g. el enfado se le pasó* ‘his anger subsided’;
>
> 5) Noun *de* LU: Here, we find noun collocates that precede the LU introduced by the preposition *de* ‘of’; *e.g. atisbo de esperanza* ‘a glimmer of hope’.

Once the user has entered one of these sections, he will find a list of collocates or semantic derivates preceded by an LF, a gloss, and followed by one or more examples. In the *gloss* we intend to give a brief indication of the meaning of the collocate in relation to the base. So, the gloss *intensa* ‘intense’ serves to group various adjectives such as *ferviente* ‘burning’, *profunda* ‘profound’, and *enorme* ‘enormous’, which, in combination with the noun ADMIRACIÓN ‘admiration’, fulfill the same role, although they do not have strictly the same meaning. This proved to be a very useful feature especially for learners, who may have a problem choosing correctly between collocations which at first sight might appear to have similar meanings. For instance, the following adjectives used with the noun *admiración* are described in the glosses as follows:

(1) *incondicional*, glossed as *intensa* ‘intense’

(2) *ciega*, glossed as *más intensa de lo conveniente* ‘more intense than convenient’

(3) *general*, glossed as *compartida por muchos* ‘shared by many persons’

(4) *eterna*, glossed as *que dura mucho* ‘long-lasting admiration’

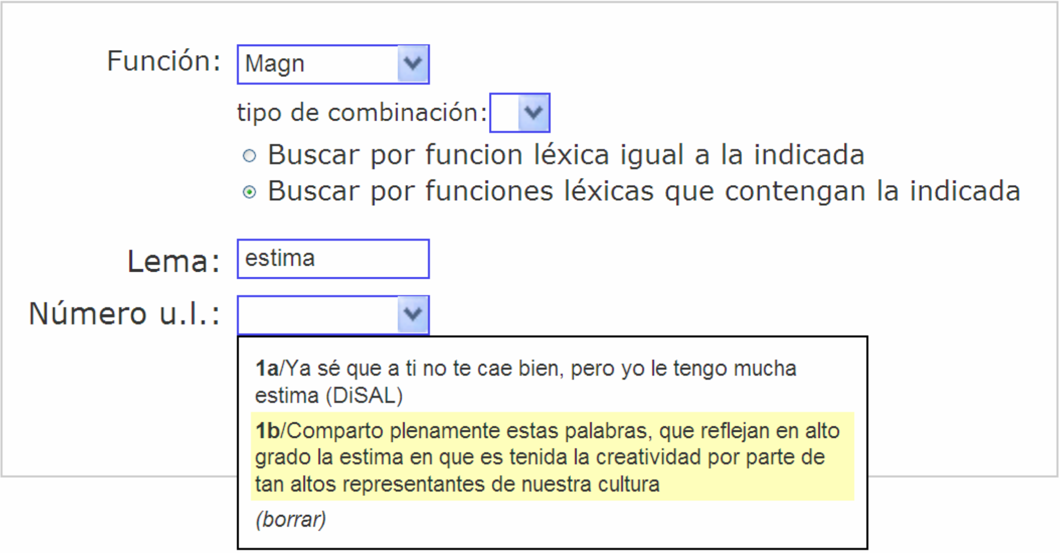
2.2. The advanced search component

The *Consultas avanzadas* ‘advanced search’ component serves principally to carry out specific searches. Rather than making queries for the collocates of a specific LU, it helps us find the answer for particular questions.

We can conduct three types of searches: 1) direct search, 2) inverse search and 3) writing aid.

2.2.1. Direct search

Consultas directas ‘direct search’ allows us to find the collocates of a base described by a given LF. Besides the LF, the user has a further option of specifying the lemma of the base and its lexical unit when carrying out a search (see Figure 1 for an example of a search for the collocates described by the LF Magn of the LU *estima 1b*).



Función: Magn

tipo de combinación:

○ Buscar por funcion léxica igual a la indicada

● Buscar por funciones léxicas que contengan la indicada

Lema: estima

Número u.l.:

1a/Ya sé que a ti no te cae bien, pero yo le tengo mucha estima (DiSAL)

1b/Comparto plenamente estas palabras, que reflejan en alto grado la estima en que es tenida la creatividad por parte de tan altos representantes de nuestra cultura

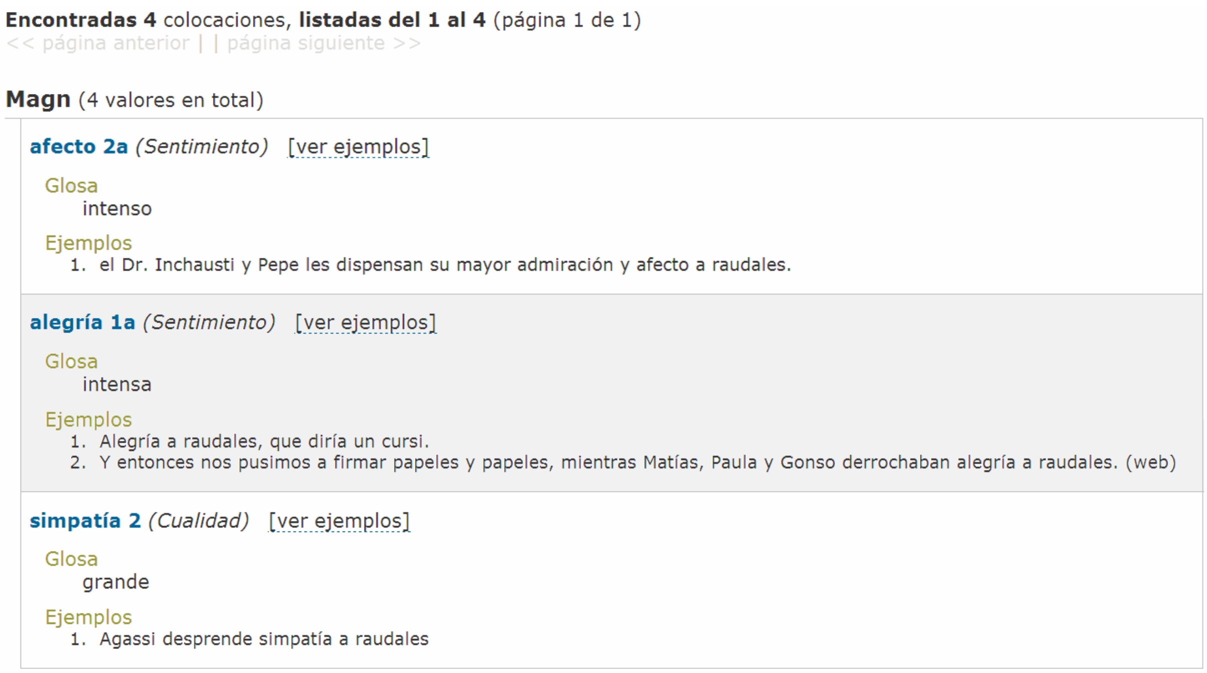
(borrar)

*Figure 1. Direct search for Magn(*estima *1b)*

2.2.2 Inverse search

We can conduct two types of searches using the option *Consultas inversas* ‘inverse search’:

1) The first type of search allows us to find the base of a collocation starting from the collocate. After having indicated the collocate, we also have the option of specifying the LF associated with it. Figure 2 shows the results obtained from the search for the collocate *a raudales* ‘in abundance’.



Encontradas 4 colocaciones, **listadas del 1 al 4** (página 1 de 1)
<< página anterior | | página siguiente >>

Magn (4 valores en total)

afecto 2a *(Sentimiento)* [ver ejemplos]

Glosa
intenso

Ejemplos
1. el Dr. Inchausti y Pepe les dispensan su mayor admiración y afecto a raudales.

alegría 1a *(Sentimiento)* [ver ejemplos]

Glosa
intensa

Ejemplos
1. Alegría a raudales, que diría un cursi.
2. Y entonces nos pusimos a firmar papeles y papeles, mientras Matías, Paula y Gonso derrochaban alegría a raudales. (web)

simpatía 2 *(Cualidad)* [ver ejemplos]

Glosa
grande

Ejemplos
1. Agassi desprende simpatía a raudales

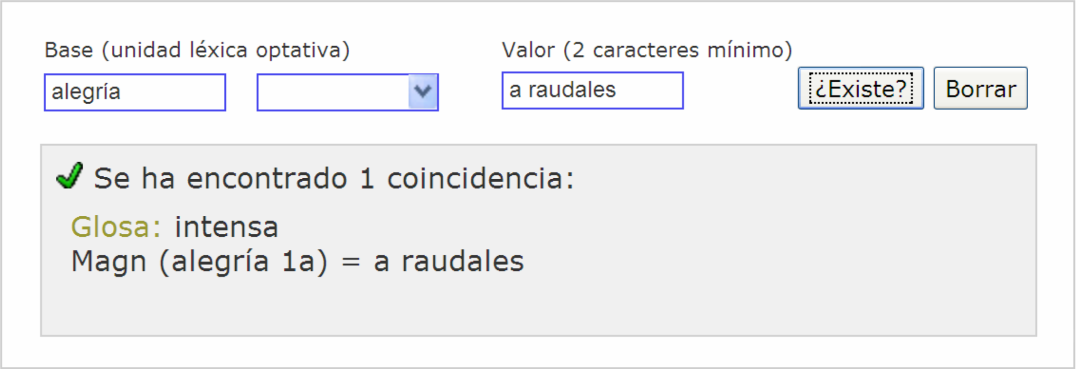
Figure 2. Results of an inverse search for a raudales *as a collocate*

2) The second type of search is more oriented towards comprehension. Here we can find out which LF – and gloss – codifies the relation between a given base and a collocate. For example, we can find that *a raudales* adds the meaning ‘intense’ to the base *alegría*.

2.2.3 Writing aid

The option *Ayuda a la redacción* ‘writing aid’ is intended to resolve questions concerning lexical combinatorics raised by any speaker of Spanish, including learners and native speakers. It helps us verify whether a certain combination of words is correct. At this moment, we offer the following two types of aid:

1) The first kind of aid allows the user to check whether a given base can co-occur with a given collocate (cf. Figure 3).



Base (unidad léxica optativa): alegría

Valor (2 caracteres mínimo): a raudales

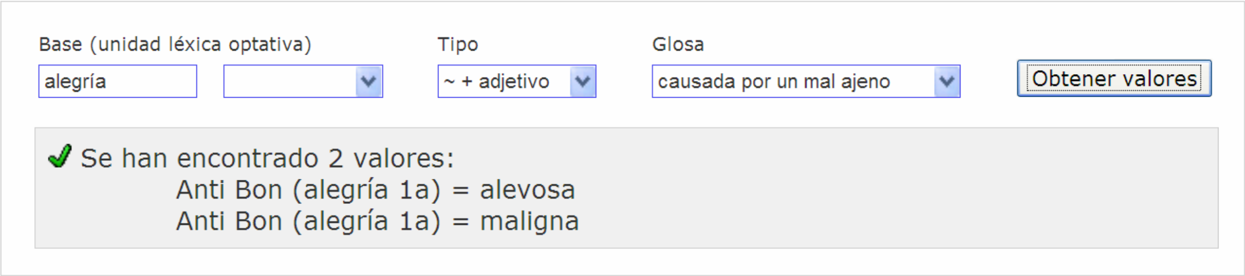
✓ Se ha encontrado 1 coincidencia:

Glosa: intensa

Magn (alegría 1a) = a raudales

Figure 3. Checking collocations with the Writing aid tool

2) The second aid provides as search results collocates corresponding to a meaning, codified by a gloss, and a syntactic scheme (under “tipo”). Figure 4 shows a search for collocate adjectives of *alegría* ‘joy’, meaning ‘caused by the misfortune of another person’.



Base (unidad léxica optativa): alegría

Tipo: ~ + adjetivo

Glosa: causada por un mal ajeno

✓ Se han encontrado 2 valores:

Anti Bon (alegría 1a) = alevosa

Anti Bon (alegría 1a) = maligna

Figure 4. Finding collocates with the Writing aid tool

3. Conclusion and future work

As we have shown, the electronic format of DiCE and the codification of collocations through LFs and glosses turn out to be a clear advantage over conventional collocation dictionaries. The use of LFs allows for the efficient systematization of the representation of collocations. Such a systematic representation is an important aid for both the lexicographer and the users of the dictionary. DiCE, thus, provides structured information on Spanish collocations, which can be exploited in various ways by users through the different search options.

From a lexicographical point of view, along with a gradual expansion of the dictionary itself, one of our primary concerns is to look more closely into the possibilities of standardising the glosses and generalizing them in accordance with the meaning of bases.

In the mid-term future we also aim at exploiting the database integrating the dictionary with an exercise module, providing an online language learning environment. For further support of the learner, the next release of DiCE will furthermore offer each user the option to create his/her own learning space in which he/she can administrate personal collocation lists, annotations, performance scores and identified problems with respect to specific collocations or collocation types[3].

References

ALONSO RAMOS, M. (2005). Semantic Description of Collocations in a Lexical Database. In F. Kiefer G. Kiss and J. Pajzs (eds.) *Papers in Computational Lexicography COMPLEX 2005.* Budapest: Linguistics Institute and Hungarian Academy of Sciences: 17-27.

BENSON, M., BENSON, E. and ILSON, R. (1998). *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam and Philadelphia: John Benjamins.

CROWTHER, J., DIGNEN, S. and LEA, D. (eds) (2002). *Oxford Collocations Dictionary for Students of English.* Oxford: Oxford University Press.

HAUSMANN, F.J. (1979). Un dictionnaire des collocations est-il possible? *Travaux de littérature et de linguistique de l'Université de Strasbourg*, 17/1: 187-195.

HILL, J. and LEWIS, M. (eds) (1997). *LTP Dictionary of Selected Collocations*. London: LTP.

MEL'ČUK, I., CLAS, A. and POLGUÈRE, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

[3] This paper was written within the framework of a research project: FFI2008-06479-C02-01 (Ministerio de Ciencia), and partially funded by FEDER.