

2010

# eLexicography in the 21<sup>st</sup> 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**

### **Comité scientifique**

Sergio Bolasco	Università di Roma « La Sapienza », IT
Laurence Danlos	Université Paris 7, FR
Guy Deville	Facultés universitaires Notre-Dame de la Paix, BE
Ray Dougherty	University of New York, US
Thierry Dutoit	Faculté Polytechnique de Mons, BE
Louissette Emirkanian	Université du Québec à Montréal, CA
Cédric Fairon	Université catholique de Louvain, BE
Thierry Fontenelle	Centre de traduction de la Communauté européenne, LU
Sylviane Granger	Université catholique de Louvain, BE
Guy Lapalme	Université de Montréal, CA
Eric Laporte	Université de Marne-la-Vallée, FR
Denis Maurel	Université François-Rabelais de Tours, FR
Damon Mayaffre	Université de Nice – Sophia Antipolis, FR
Sébastien Paumier	Université de Marne-la-Vallée, FR
Alain Polguère	Université de Montréal, CA
Antoinette Renouf	University of Central England in Birmingham, UK
Jean Senellart	Systran, FR
Anne-Catherine Simon	Université catholique de Louvain, BE
Agnès Tutin	Université Stendhal – Grenoble 3, FR
Pierre Zweigenbaum	CHU Pitié-Salpêtrière, Université Paris 6, FR

### **Comité de rédaction**

Anne Dister	CENTAL, Université catholique de Louvain, BE
Cédric Fairon (directeur)	CENTAL, Université catholique de Louvain, BE

### **Secrétariat de rédaction**

Bernadette Dehottay	CENTAL, Université catholique de Louvain, BE 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 – <a href="http://www.uclouvain.be/cental">http://www.uclouvain.be/cental</a>
---------------------	---

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

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

© Presses universitaires de Louvain, 2010

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

Imprimé en Belgique

Tous droits de reproduction, d'adaptation ou de traduction, par quelque procédé que ce soit, réservés pour tous pays, sauf autorisation de l'éditeur ou de ses ayants droit.

*Diffusion :*

[www.i6doc.com](http://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](mailto: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  
[libwubr@club-internet.fr](mailto:libwubr@club-internet.fr)

# Table of Contents

Table of Contents .....	v
Acknowledgements .....	ix
<b>Papers</b> .....	<b>1</b>
Andrea ABEL, <i>Towards a systematic classification framework for dictionaries and CALL</i> .....	3
James BREEN, <i>Identification of neologisms in Japanese by corpus analysis</i> .....	13
Annelen BRUNNER, Kathrin STEYER, <i>Wortverbindungsfelder – Fields of multi-word expressions</i> .....	23
Louise-Amélie COUGNON, Richard BEAUFORT, <i>SSLD: a French SMS to Standard Language Dictionary</i> .....	33
Patrick DROUIN, <i>Extracting a bilingual transdisciplinary scientific lexicon</i> .....	43
Isabel DURÁN-MUÑOZ, <i>Specialised lexicographical resources: a survey of translators' needs</i> .....	55
Carolina FLINZ, <i>DIL: a German-Italian online specialized dictionary of linguistics</i> .....	67
Nuria GALA, Véronique REY, <i>Acquiring semantics from structured corpora to enrich an existing lexicon</i> .....	77
Sylviane GRANGER, Magali PAQUOT, <i>Customising a general EAP dictionary to meet learner needs</i> .....	87
Antton GURRUTXAGA, Igor LETURIA, Eli POCIELLO, Xabier SARALEGI, Iñaki SAN VICENTE, <i>Evaluation of an automatic process for specialized web corpora collection and term extraction for Basque</i> .....	97
Patrick HANKS, <i>Elliptical arguments: a problem in relating meaning to use</i> .....	109
Thomas HERBST, Peter UHRIG, <i>Valency information online – research and pedagogic reference tools</i> .....	125
Kristina HMELJAK SANGAWA, Tomaž ERJAVEC, Yoshiko KAWAMURA, <i>Automated collection of Japanese word usage examples from a parallel and a monolingual corpus</i> .....	137
Olga KARPOVA, Mikhail GORBUNOV, <i>Cultural values in a learner's dictionary: in search of a model</i> .....	149
Annette KLOSA, <i>On the combination of automated information and lexicographically interpreted information in two German online dictionaries...</i>	157

Jette Hedegaard KRISTOFFERSEN, Thomas TROELSGÅRD, <i>Making a dictionary without words: lemmatization problems in a sign language dictionary</i> .....	165
Vincent LANNOY, <i>Free online dictionaries: why and how?</i> .....	173
Godelieve LAUREYS, <i>The Hub&amp;Spoke Model put into practice. Report on the semi-automatic extraction of a pre-version of a Finnish-Danish dictionary from a multilingual interlinkable database</i> .....	183
Robert LEW, Patryk TOKAREK, <i>Entry menus in bilingual electronic dictionaries...</i>	193
Marie-Claude L'HOMME, <i>Designing specialized dictionaries with natural language processing. Examples of applications in the fields of computing and climate change</i> .....	203
Hanhong LI, <i>Word frequency distribution for electronic learner's dictionaries</i> ...	217
Marc LUDER, <i>Building an OLIF-based lexical database for representing constructions</i> .....	229
Cédric MESSIANT, Thierry POIBEAU, <i>Automatic lexical acquisition from corpora: some limitations and tentative solutions</i> .....	241
Mojca PECMAN, Claudie JULLIARD, Natalie KÜBLER, Alexandra VOLANSCHI, <i>Processing collocations in a terminological database based on a cross-disciplinary study of scientific texts</i> .....	249
Bálint SASS, Júlia PAJZS, <i>FDVC – Creating a corpus-driven frequency dictionary of verb phrase constructions for Hungarian</i> .....	263
Stefania SPINA, <i>The Dici project: towards a dictionary of Italian collocations integrated with an online language learning platform</i> .....	273
Daniela TISCORNIA, <i>An ontology-based approach to the multilingual complexity of law</i> .....	283
Sabine TITTEL, <i>Dynamic access to a static dictionary: a lexicographical “cathedral” lives to see the twenty-first century – the Dictionnaire étymologique de l'ancien français</i> .....	295
Lars TRAP-JENSEN, <i>Access to multiple lexical resources at a stroke: integrating dictionary, corpus and Wordnet data</i> .....	303
Agnès TUTIN, <i>Showing phraseology in context: onomasiological access to lexico-grammatical patterns in corpora of French scientific writings</i> .....	313
Isabel VERDAGUER, Elisabet COMELLES, Natàlia J. LASO, Eva GIMÉNEZ, Danica SALAZAR, <i>SciE-Lex : an electronic lexical database for the Spanish medical community</i> .....	325
Serge VERLINDE, <i>The Base lexicale du français: a multi-purpose lexicographic tool</i> .....	335
Alexandra VOLANSCHI, Natalie KÜBLER, <i>Building an electronic combinatory dictionary as a writing aid tool for researchers in biology</i> .....	343
Michael ZOCK, Tonio WANDMACHER, Ekaterina OVCHINNIKOVA, <i>Are vector-based approaches a feasible solution to the “tip-of-the-tongue” problem?</i> .....	355

<b>Posters and software demonstrations</b> .....	367
Margarita ALONSO RAMOS, Alfonso NISHIKAWA, Orsolya VINCZE, <i>DiCE in the web. An online Spanish collocation dictionary</i> .....	369
Margarita ALONSO RAMOS, Leo WANNER, Nancy VAZQUEZ VEIGA, Orsolya VINCZE, Estela MOSQUEIRA SUAREZ, Sabela PRIETO GONZALEZ, <i>Tagging collocations for learners</i> .....	375
James BREEN, <i>WWWJDIC – A feature-rich WWW-based Japanese dictionary</i> .....	381
Elisa CORINO, Cristina ONESTI, <i>Have I got the wrong definition of ...? How to write simple technical definitions on the basis of examples taken from Newsgroup discussions</i> .....	387
Nathalie GASIGLIA, <i>Some editorial orientations for a multi-tier electronic monolingual school dictionary</i> .....	393
Fadila HADOUCHE, Marie-Claude L'HOMME, Guy LAPALME, <i>Automatic annotation of actants in specialized corpora</i> .....	399
Jakob HALSKOV, Pia JARVAD, <i>Automated extraction of neologisms for lexicography</i> .....	405
Adam KILGARRIFF, Vojtěch KOVAR, Pavel RYCHLÝ, <i>Tickbox Lexicography</i> .....	411
Vera KUZMINA, Anna RYLOVA, <i>ABBY Lingvo electronic dictionary platform and Lingvo Content dictionary writing system</i> .....	419
Margit LANGEMETS, Andres LOOPMANN, Ülle VIKS, <i>Dictionary management system for bilingual dictionaries</i> .....	425
Héctor MARTÍNEZ, Marta VILLEGAS, Núria BEL, Santiago BEL, Francesca ALEMANY, <i>Lexicography in the grid environment</i> .....	431
Carolín MÜLLER-SPITZER, Christine MÖHRS, <i>The “Online Bibliography of Electronic Lexicography” (OBELEX)</i> .....	439
Cornelia TSCHICHOLD, <i>From lexical database to intelligent vocabulary trainers</i> .....	445
Eveline WANDL-VOGT, <i>Multiple access routes. The dictionary of Bavarian dialects in Austria / Wörterbuch der bairischen Mundarten in Österreich (WBÖ)</i> .....	451

# DiCE in the web: An online Spanish collocation dictionary

Margarita Alonso Ramos<sup>1</sup>, Alfonso Nishikawa<sup>1</sup>, Orsolya Vincze<sup>1</sup>  
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<sup>2</sup>.

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 not 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

---

<sup>1</sup> University of A Coruña, lxalonso@udc.es.

<sup>2</sup> 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 derivatives 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 *admira*ción 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:

tipo de combinación:

Buscar por función léxica igual a la indicada

Buscar por funciones léxicas que contengan la indicada

Lema:

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)

<p><b>afecto 2a</b> (<i>Sentimiento</i>) [<a href="#">ver ejemplos</a>]</p> <p>Glosa intenso</p> <p>Ejemplos 1. el Dr. Inchausti y Pepe les dispensan su mayor admiración y afecto a raudales.</p>
<p><b>alegría 1a</b> (<i>Sentimiento</i>) [<a href="#">ver ejemplos</a>]</p> <p>Glosa intensa</p> <p>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)</p>
<p><b>simpatía 2</b> (<i>Cualidad</i>) [<a href="#">ver ejemplos</a>]</p> <p>Glosa grande</p> <p>Ejemplos 1. Agassi desprende simpatía a raudales</p>

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)    
 Valor (2 caracteres mínimo)

✔ 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)    
 Tipo    
 Glosa

✔ 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<sup>3</sup>.

## 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.

---

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