An online collocation dictionary of Spanish

Orsolya Vincze (1), Estela Mosqueira (1) and Margarita Alonso Ramos (1)

(1) Facultade de Filoloxía, Universidade da Coruña
Campus da Zapateira, s/n, 15071 Coruña (Spain)
ovincze@udc.es|estela.mosqueira@udc.es|lxalonso@udc.es

Abstract

DiCE is an online dictionary of Spanish collocations, which provides semantic and combinatorial information on lexical units. It 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. We offer a general presentation of the content of the dictionary, followed by a description of the user's interface and a brief insight into the lexicographer's interface. The main feature of DiCE is that it is conceived as a part of a language-learning environment which combines dictionary, corpus and teaching materials.

Keywords

collocations, lexical functions, online dictionary, learning tools, Spanish as a second language

1 Introduction

In this paper, we present the *Diccionario de Colocaciones del Español* (DiCE), a web-based collocation dictionary of Spanish, which is available online since 2004 with its database being constantly modified and expanded. Although the dictionary has been presented on various occasions (e.g. Alonso Ramos, 2005, 2006, 2008, 2010), hereby, we will focus on some features that have not been described in previous publications.

Similarly to Dicouèbe (Polguère, 2000, Jousse & Polguère, 2005) and DicoInfo (L'Homme, 2009), DiCE constitutes an online implementation of the principles of lexical description introduced in the *Explanatory Combinatorial Lexicology* (Mel'čuk et al., 1995). In addition to providing a theoretically-based description of collocations, DiCE aims to be a useful tool not only for researchers, but also for the general public, that is, native and non-native speakers of Spanish. This is achieved, on the one hand, by adapting the information offered by the dictionary to the needs of general users, for instance, by paraphrasing *lexical functions* with natural language glosses. On the other hand, the web interface is designed to enable flexible

access to the electronic lexical database, satisfying users ranging from researchers through language learners to lexicographers working on DiCE.

In the following sections, we offer a general presentation of the content of the dictionary, followed by a description of the user's interface and a brief insight into the lexicographer's interface.

2 General presentation – the content of DiCE

DiCE is essentially a collocation dictionary: its main objective is to describe restricted lexical co-occurrence, although it also deals with semantic derivatives. More precisely, it concentrates on both paradigmatic and syntagmatic lexical relations controlled by a lexical unit (LU). For now, the list of lemmas treated in the dictionary is limited to the semantic field of emotions, so that DiCE specifies approximately 19500 lexical relations (values of lexical functions).

The lexicographical description of each LU can be divided into semantic and 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 (QSyn) and quasi-antonyms (QAnti) of the LU.

The combinatorial information offered by the dictionary is of two types: syntactic and lexical combinatorics. The syntactic combinatory information of the LU is shown in the Government Pattern (*esquema de régimen)* section, where we specify the projection of the semantic valency structure of the LU onto its syntactic valency structure and, in addition, the subcategorization information associated with the latter. To illustrate this, as shown in Figure 1, in the case of ALEGRÍA 'joy' whose propositional form is *alegría de individuo X por hecho Y* 'person X's joy over fact Y' we specify that, for instance, semantic actant X can be realized as a prepositional phrase headed by the preposition *de,* as a possessive determiner or as an adjective (see examples 3, 10 and 4 respectively). The lexical combinatory information is displayed in the section Collocations. In what follows, we focus on lexical combinatorics.

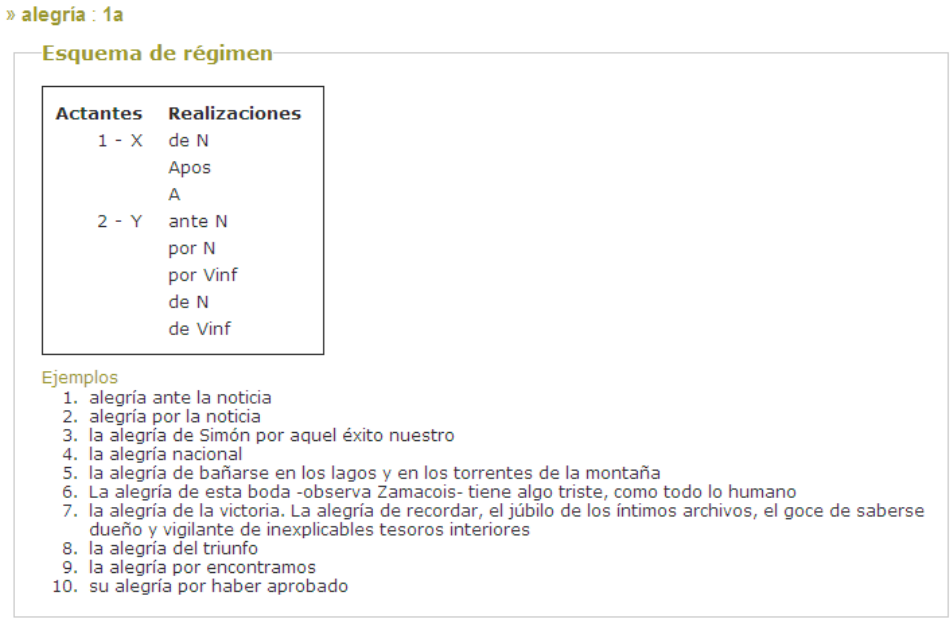

Figure 1: Syntactic combinatorial information on UL *alegría 1a*

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 ALEGRÍA 'joy', the user finds *loco de alegría* 'crazy with joy' or *exultante* 'joyful, exultant', both referring to the participant who is feeling a lot of joy;

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. provocar dolor* '[to] cause pain';

4. *LU + verb*: This section contains verbs that take the LU as the grammatical subject, *e.g. la alegría se desvanece* 'joy disappears';

5. *Noun + de LU*: Here, we find noun collocates that precede the LU, introduced by the preposition *de* 'of'; *e.g. arrebato de celos* 'a fit of jealousy'.

Once a user has entered one of these sections, they will find a list of collocates or semantic derivatives preceded by an LF, and followed by a gloss and 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 *intenso* 'intense' serves to group various adjectives such as *desbordante* 'overwhelming', *enorme* 'enormous', and *indescriptible* 'indescribable', which, in combination with the noun ALEGRÍA 'joy', fulfill the same role, although they do not have strictly the same meaning. Using glosses to describe the meaning of collocations proved to be a very useful feature especially for learners, who may have a problem interpreting or choosing collocations without explicit information on their meaning, generally missing from

conventional collocation dictionaries. For instance, the meanings of the following adjectives used with the noun ODIO 'hatred' are described in the glosses as follows:

(1) *mortal* 'lethal', glossed as *intenso* 'intense' [Magn(*odio*)]

(2) *declarado* 'declared', glossed as *que se manifiesta* 'manifest' [A_1Manif(*odio*)]

(3) *eterno* 'eternal', glossed as *que dura mucho* 'long-lasting' [$Magn_{\_temp}$(*odio*)]

(4) *larvado* 'latent', glossed as *que no se manifiesta* 'that doesn't manifest itself'[A_1nonManif(*odio*)]

3 The user's interface

The user's interface consists of three main components: 1) the dictionary itself, 2) the advanced search component, and 3) the didactic module.

3.1 The dictionary component

This component allows the user to access the dictionary in the more traditional way, through the list of lemmas. Each lemma is associated with a list of lexical units (LUs), where corresponding semantic and combinatorial information can be found (see above).

3.2 The advanced search component

The *Consultas avanzadas* 'advanced search' component serves to carry out specific queries. Rather than looking up the list of collocates of a given LU, it helps us find the answer for specific questions.

The user is provided with four types of search tools: 1) *direct search*, 2) *inverse search*, 3) *what does it mean?*, and 4) *writing aid*.

3.2.1 Direct search

Consultas directas 'direct search' allows the user to find collocations described by a given LF. As an answer to the query for combination Magn+$Caus_2Oper_1$, the system returns all collocations described by this sequence of LFs stored in the database. To restrict this search, the user has a further option of specifying the lemma of the base and its LU. For instance, as shown in Figure 2, one can restrict the search for collocates described by the LF Magn+$Caus_2Oper_1$ specifying the lemma ALEGRÍA and choosing one of its LUs, in this case *1a*. Note that, in order to facilitate the choice of a specific LU, in each case, an example is visualized together with the numeration used to code LUs in the dictionary entry.

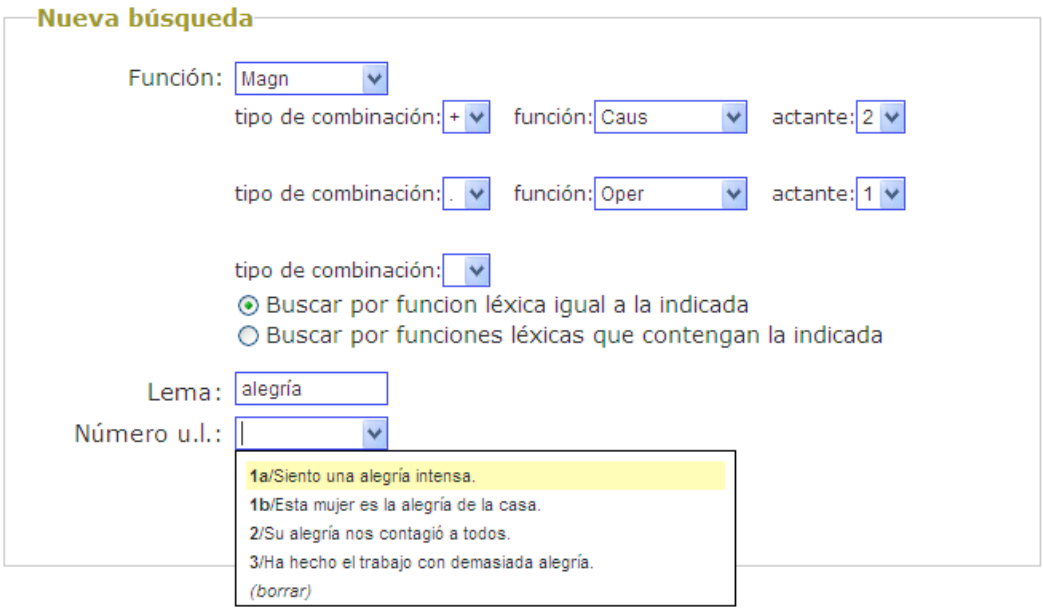

Figure 2: Direct search for *Magn+Caus$_2$Oper$_1$(alegría 1a)*

3.2.2 Inverse search

The option *Consultas inversas* 'inverse search' allows the user to find the base of a collocation starting from the collocate. After having indicated the collocate, the search can be further restricted by specifying the LF associated with the collocation. Figure 3 shows a sample of the results obtained from the search for the collocate *guardar* 'keep'. Through this query the user can learn that the same collocate, in this case the verb *guardar* can be combined with different bases in order to constitute collocations that are codified by different LFs. For instance, *guardar rencor* 'bear a grudge' is described by the LF ContOper$_1$ while *guardar sorpresa* 'have a surprise in store (for sb)' is codified by LF CausFunc$_2$.

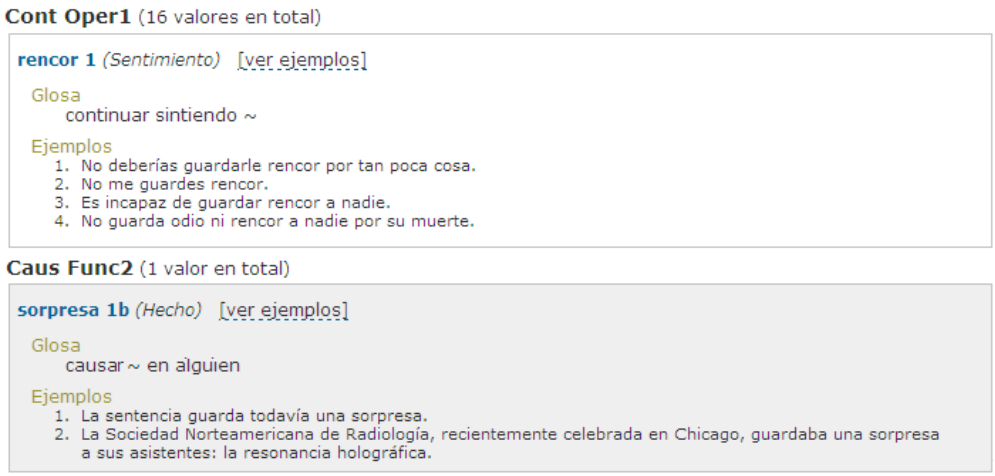

Figure 3: A sample of the results of an inverse search for *guardar* as a collocate

3.2.3 What does it mean?

The module *¿Qué significa?* 'What does it mean?' is oriented towards comprehension. It serves to find which LF – and gloss – codifies the relation between a given base and a collocate. For example, as shown in Figure 4, we can learn that *ligero* expresses the meaning 'small in degree' when combined with the base *arrepentimiento 1* 'repentance'.

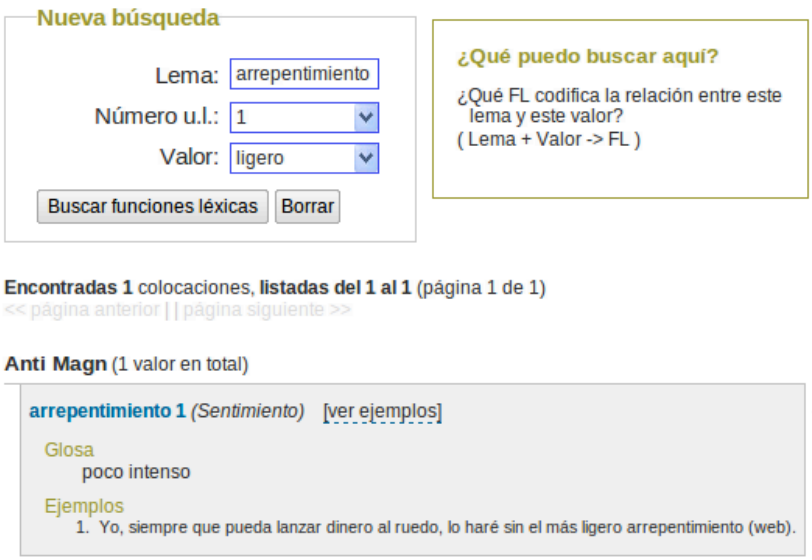

Figure 4: Results of a search for the collocation *arrepentimiento ligero* with the *¿Qué significa?* tool

3.2.4 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. 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. In Figure 5, we show the result of a query for the collocation *arrepentimiento ligero* 'light repentance'.

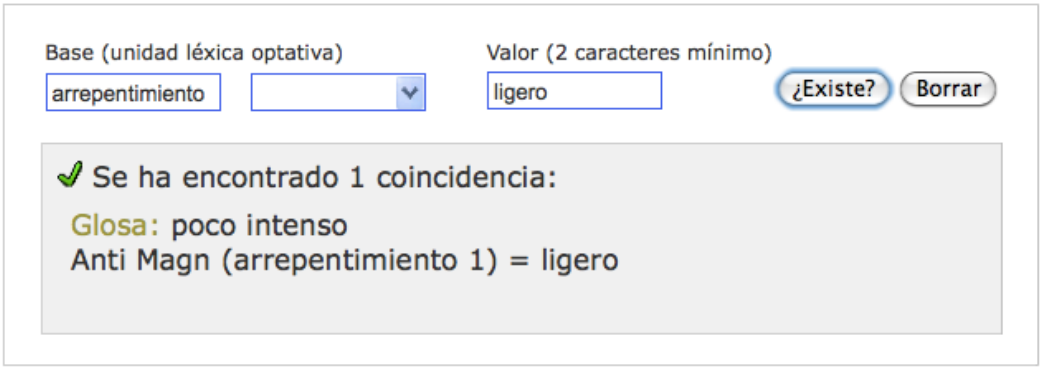

Figure 5: Checking the collocation *arrepentimiento ligero* with the Writing aid tool

2. The second aid enables users to find collocates corresponding to a specific meaning, codified by a gloss, and a syntactic scheme (under "tipo"). Figure 6 shows a search for a collocate adjective of *amor I.1a* 'love' with the meaning 'felt for one another', for which the tool returns the collocation *amor correspondido.*

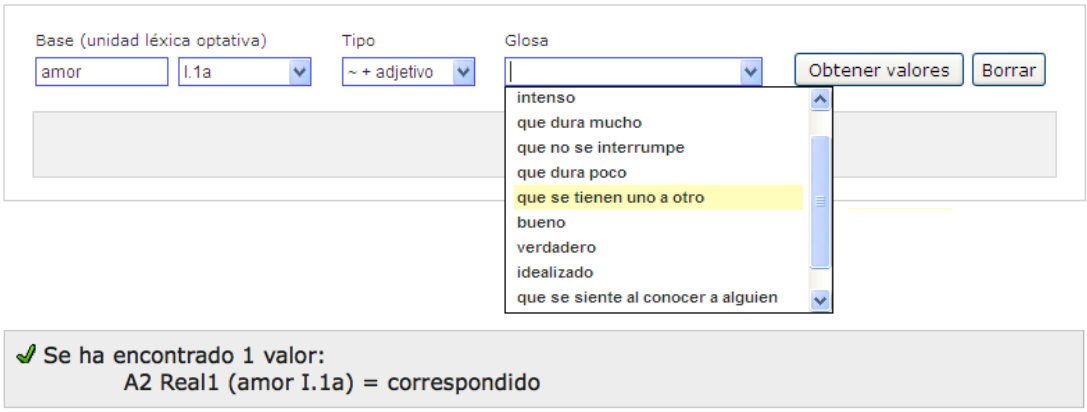

Figure 6: Finding collocates of *amor* I.1a with the meaning 'felt for one another' using the Writing aid tool

3.3 The didactic module

The aim of the exercise module, still in development, is to provide the user with learning material concentrating on collocations. For now, it is limited to a few sections containing exercises related to a particular topic, among others, an introduction to the use of DiCE itself.

Our mid-term goal is to exploit the DiCE database integrating the dictionary with a more complete didactic module, providing an online language learning environment. For further support of the learner, we are planning to offer users the option of creating their own learning space in which they can administrate personal collocation lists, annotations, performance scores and problems identified with respect to specific collocations or collocation types. Furthermore, the DiCE forms part of the COLOCATE Project where we intend to integrate the lexical database with a corpus interface and a checker tool that will provide aid for users on collocations encountered in reading as well as writing tasks[1].

4 The lexicographer's interface

The lexicographer's interface allows the editors of DiCE to carry out instant modifications in the lexical database via the web. Essentially, there are two ways of editing the dictionary: either through viewing the *User's Interface* or through the *Administration Area*.

[1] The COLOCATE Project is being conducted with collaboration of researchers from the Universitat Pompeu Fabra and the Universidade da Coruña. Experiments carried out by Wanner and his colleagues have shown promising results. See Wanner et al. (this volume).

4.1 Editing DiCE through the User's Interface

Editors have the option of modifying the DiCE database directly from the User's Interface view. This allows quick corrections and modifications of the content while browsing the dictionary. For instance, a lexicographer can access the edition area to correct a mistake in the description of a concrete collocation by simply clicking on the corresponding icon, see Figure 7.

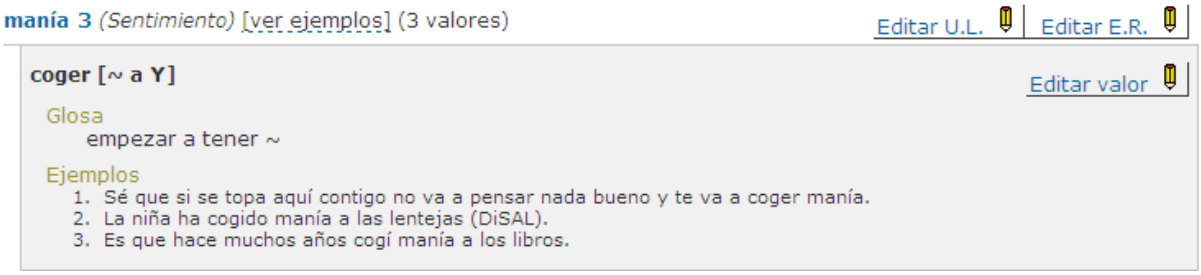

Figure 7: Lexical entry of the collocation *coger manía* 'take a dislike (to sb)' with the "Edit" icons visible in the upper right-hand corner.

4.2 Editing DiCE through the Administration Area

Lexicographers are provided with a more efficient tool to carry out modifications, remove or hide existing data and incorporate new information in the database through the two main options of the Administration Area: 1) *Modification of lexical information* and 2) *Mass update*.

4.2.1 Modification of lexical information

In this section, the lexicographer has access to the database through the list of lemmas to edit both semantic and combinatory information of LUs, such as the semantic tag, the actantial structure and the government pattern of LUs, EuroWordNet IDs[2] and the descriptions of collocations themselves. In Figure 8 we show the lexicographer's interface screen for the lemma AMISTAD.

[2] We provide the ID assigned in Spanish EuroWordNet for each LU. This information comes from the research carried out during a project focusing on linking DiCE with the Spanish EuroWordNet) (see Wanner et al. 2004).

Unidad léxica			EWNET	Esquema de régimen	Colocaciones	Operaciones		
Número de U.L.	Forma proposicional.	Etiqueta semántica	EWNET	Esquema de régimen	Colocaciones	Publicado	Editar	Eliminar
1	amistad de individuo X hacia individuo Y	Sentimiento						
	Ejemplos : 1 - No son novios, entre ellos sólo existe una buena amistad (Clave) 2 - En Sahagún contaba con la amistad y la hospitalidad de Martín y de Zulema							
2a	[individuo Y] es una amistad de individuo X	Individuo						
	Ejemplos : 1 - Me presentó a una amistad de la infancia (larousse) 2 - Tengo algunas amistades en Francia (DUE) 3 - felicitaba el año nuevo a sus amistades							
2b	[individuos Y] son las amistades de individuo X	Individuo						
	Ejemplos : 1 - Tiene amistades en el ministerio que lo apoyarán (DiSAL) 2 - ese negocio se lo debo a las amistades (Lema) 3 - se valió de todas sus amistades para poder concretar contactos en la pequeña comunidad del cine internacional							

Figure 8: Lexicographer's interface screen for editing information on LUs of the lemma AMISTAD

4.2.2 Mass update

This option allows editors of the dictionary to carry out mass modifications of glosses, government patterns, collocate lemmas and lexical functions. So far, editors have found this option especially useful in adding or modifying glosses of large groups of collocations. As we show in Figure 9, glosses of all collocations described by the LF $Oper_1$ and belonging to LUs with the semantic tag 'sentimiento' ('feeling') can be easily changed using this tool.

Actualizar la glosa a...

Nuevo valor de Glosa:

sentir ~

Actualizar Limpiar

650 colocaciones, **listadas de la 1 a la 20** (página 1 de 33)

<< página anterior | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | página siguiente >>

Lema	Unidad léxica	Función léxica	Glosa	Valor	Régimen
abatimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sentir	
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	tener	[~]
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sentir	[~]
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	experimentar	[~]
aborrecimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	nutrir	[ART ~]
aburrimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sentir	[~]
aburrimiento (Nombre masculino)	1 (Sentimiento)	Oper1	sentir ~	sufrir	[de ~]

Figure 8: Mass update of natural language glosses of collocations with the LF $Oper_1$

Similarly, it is possible to edit LFs of a large amount of collocations sharing the same collocate or base (for example, to change the combination of LFs that encodes collocations

with the collocate *enfermizo* 'unhealthy' from Magn+AntiBon to Magn+AntiVer), or to change the government pattern associated with collocations described by a given LF, etc. All these options are quite useful in making necessary changes or correcting errors in the database in a largely efficient way.

4.3 Exploiting DiCE as a corpus

We think that the examples found in the lexical entries of the dictionary can constitute valuable data for research. This is why we have included a tool to extract corpus examples from the dictionary in the Administration Area. Introducing a specific LF or combination and, optionally, a lemma, the lexicographer can download a .txt file with the corresponding example sentences. For this reason, we can say that the DiCE contains a corpus of collocations (Alonso Ramos, 2009). To illustrate this point, all examples containing collocations described by the LF $Func_1$ constitute a corpus of 3814 words; those that contain collocations described by $Oper_2$ amount to 3236 words; and, those illustrating cases of $IncepOper_1$, contain 3887 words, and so on. Collocation corpora obtained from the DiCE can be tagged and parsed in order to obtain a collocation Treebank for further investigation needs.

5 Conclusion and future work

A dictionary, as we conceive of it, is necessarily a project that is constantly in the course of development and indefinitely undergoing changes. In this way, the DiCE already has a version 1.0 (which was available from 2004) and the current version undergoing constant modifications since last year. It is changing not only in its content, but also in its interface and ways of access to information.

With respect to the content of the DiCE, future changes concern information on the frequency of use of collocations. After a long process of semantic disambiguation of corpus samples, we managed to assign frequency scores to the LUs contained in the dictionary, that is, the bases of collocations. As the next step, we will proceed to assign frequency information to collocations as a whole. Another piece of information we aim at adding shortly concerns assigning collocations to particular levels of L2 Spanish: this means specifying which collocations should be taught to learners of different levels (elementary, intermediate and advanced), with a view to extracting leveled teaching materials. With respect to the access to information, our goal is to develop a semantic typology of LFs (similar to the one proposed by Jousse, et al. 2008) that would allow the user to look up collocations with a semantic focus. For instance, if a user is searching for how to verbalize the meaning related to the phase of the starting fear, it would be convenient to find verb+object collocacions like *coger miedo* 'take fear of sg' as well as subject+verb collocations like *entrarle miedo* 'fear enters sb', *asaltarle miedo* 'fear assaults sb', or *invadirle el miedo* 'fear invades sb'. At this moment, these cannot be found in one single search, given that collocations are currently classified according to their syntactic structure.

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, but this is not enough. In line with the proposals put forward by Verlinde et al. (2009), the concept of dictionary is changing towards being a more flexible and more dynamic tool, which is more oriented to the users' needs; a tool that should be considered as a *leximat* (Tarp,

2008). Jousse et al. (2008) also prefer referring to this new concept as a *lexical site*, instead of a *dictionary*, due to the connotations of a linear vision carried by this latter term, while the first one proves to be a better model of lexical knowledge, as a constantly evolving network. Independently of the term we use to refer to these new lexical resources, the fact is that they have ceased to be stand-alone products, and they are necessarily integrated with other resources such as corpus and other dictionaries and glossaries. This is exactly the course of evolution we intend DiCE to take within the framework of the COLOCATE Project (see above).

Acknowledgements

This work has been supported by the Spanish Ministry of Science and Innovation and the FEDER Funds of the European Commission under the contract number FFI2008-06479-C02-01. We would also like to thank the anonymous reviewers for their valuable remarks and comments.

Bibliography

Alonso Ramos, M. 2005. Semantic Description of Collocations in a Lexical Database. In Kiefer, F. et al. (eds.). *Papers in Computational Lexicography COMPLEX 2005*, 17-27. Budapest: Linguistics Institute and Hungarian Academy of Sciences.

Alonso Ramos, M. 2006. Towards a Dynamic Way to Learn Collocations in a Second Language. In Corino, E., C. Marello & C. Onesti (eds.) *Proceedings of the Twelfth EURALEX International Congress*, 909-923. Torino: Accademia della Crusca, Università di Torino, Edizioni dell'Orso Alessandria.

Alonso Ramos, M. 2008. Papel de los diccionarios de colocaciones en la enseñanza de español como L2. In Bernal, E. & J. DeCesaris (eds.) *Proceedings of the XIII EURALEX International Congress*, 1215-1230. Barcelona: IULA, Documenta Universitaria.

Alonso Ramos, M. 2009. Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario? In Cantos Gómez, P. & A. Sánchez Pérez (eds.). *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*, 1191-1207. Murcia: AELINCO.

Alonso_Ramos, M. 2010. No importa si la llamas o no colocación, descríbela. Mellado, C. et al. (eds.), *Nuevas perspectivas de la fraseología del siglo XXI*, 55-80. Berlin: Frank & Timme.

Jousse, A. L. & A. Polguère. 2005. *Le DiCo et sa versión DiCouèbe. Document descriptif et manuel d'utilisation*. Versión du rapport 1.0 – 19 avril 2005, Montréal: Observatoire de linguistique Sens-Texte (OLST).

Jousse, A. L., A. Polguère & O. Tremblay. 2008. Du dictionnaire au site lexical pour l'enseignement/apprentissage du vocabulaire. In Grossmann, F. & S. Plane (eds). *Les apprentissages lexicaux. Lexique et production verbale*, 141–157. Villeneuve d'Ascq: Presses univeritaires du Septentrion.

L'Homme, M. C. 2009. *DiCoInfo: Le dictionnaire fondamental de l'informatique et l'Internet. Document descriptif et manuel d'utilisation*, Montréal: Observatoire de linguistique Sens-Texte (OLST).

Mel'čuk, I., A. Clas & A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire,* Louvain-la-Neuve: Duculot.

Polguère, A. 2000. Towards a theoretically-motivated general public dictionary of semantic derivations and collocations for French. In Heid, U., S. Evert, E. Lehmann & C. Rohrer (eds.) *Proceedings of the Ninth EURALEX International Congress*, 517-527. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-Knowledge*. Tübingen: Niemeyer.

Verlinde, S., P. Leroyer & J. Binon. 2009. Search and you will Find. From Stand-alone Lexicographic Tools to User Driven Task and Problem-Oriented Multifunctional Leximats, *International Journal of Lexicography*, 23(1):1–17.

Wanner, L., M. Alonso & M. A. Martí. 2004. Enriching the Spanish *EuroWordNet* by *Collocations*. In *Proceedings of LREC* 2004, Vol. 4, 1087-1091, Lisbon: ELRA.

Wanner, L, G. Ferraro & R. Nazar (this volume). Collocations: A Challenge in Computer Assisted Language Learning. *Proceedings of the Fifth International Conference on Meaning-Text Theory* (MTT '11).