Discovering hidden collocations in a bilingual Spanish–English dictionary

Margarita Alonso Ramos

Universidade da Coruña, Campus de Zapateira s/n, 15071 CORUÑA (SPAIN)
E-mail: lxalonso@udc.es

Abstract

This paper addresses the problem of how to exploit the collocational information included in an online Spanish–English dictionary. Even though collocations are not identified as such in this dictionary, abundant collocational information is used as a means of distinguishing senses. Given that this information is structured in XML markup, the conversion into a bilingual collocation database seems viable in order to obtain the germ of a first Spanish–English collocation dictionary. The concept of collocation used here comes from the Explanatory and Combinatorial Lexicology (Mel'čuk, 2012). In this framework, collocations are understood as recurrent phrases composed of two lexical units, one of which, the *base*, is selected according to its meaning, while the selection of the other, the *collocate*, is determined by the base. The methodology I propose consists of reorganizing the links between words in such a way that the bilingual collocational correspondence is included in the entry for the base. The lexical tool obtained as a result of this reorganization could be exploited for different applications in natural language processing, ranging from machine translation to computer assisted language learning systems.

Keywords: collocations; bilingual dictionary; reusability of lexical resources

1 Introduction

Collocations are usually not especially well treated in bilingual dictionaries, irrespective of the language pair concerned[1]. This can be attributed to the fact that bilingual dictionaries tend to put more emphasis on comprehension than on language production, whereas collocations are mainly *idioms of encoding* (Makkai, 1972). Such is the case of the online bilingual *Oxford Spanish–English Dictionary* (OSED http://www.oxforddictionaries.com/es/traducir/espanol-ingles/). This dictionary provides answers for an L2 Spanish user who wants to understand the meaning of a word, but gives a more complicated access to an L1 Spanish user aiming to produce a collocation in English. For instance, an L1 Spanish user who wants to know how to say *coger una enfermedad* 'to catch an illness' in English will not find the answer in the entry for the noun, but in the entry for the verb, after scrolling through a rather long article in order to find the translation *to catch an illness*. However, if the information

[1] For an overview of the treatment of collocations in the French–Spanish Larousse dictionary, see Alonso Ramos (2001). As far as the collocations in Spanish–English electronic dictionaries, see Corpas Pastor (in press).

is included under the entry for the noun *enfermedad*, access would be easier, because this is the point of departure: the user wants to speak about an illness, the *base* of the collocation.

The concept of collocation used here comes from the Explanatory and Combinatorial Lexicology (Mel'čuk, 2012). This concept does not differ substantially from that used in the *Oxford Collocations Dictionary* (OCD). In this framework, collocations are understood as recurrent phrases composed of two lexical units one of which, the *base*, is selected according to its meaning, while the selection of other, the *collocate***,** is determined by the base; in the above example, the collocate *coger* is lexically context dependent on the base *enfermedad*. Both elements of the collocation are selected in different ways. The lexical selection of the base is semantically driven, whereas the selection of the collocate is lexically driven. For instance, if a speaker wants to name the meteorological phenomenon consisting of water falling onto Earth in drops, the selection in English of the noun *rain* is semantically driven, whereas the selection of *heavy* to express that the rain is intense is lexically driven. In contrast, in Spanish or in French, it is not possible to translate *heavy rain* as *lluvia pesada* (Sp.) or *pluie lourde* (Fr.) with the literal translation of *heavy*. The correct translations are *fuerte lluvia* and *forte pluie* lit. 'strong rain'. In English you can say *a strong wind*, but not *a strong rain*, in contrast to Spanish and French, which use the adjective *fuerte* or *fort* in both cases. In this case we have three collocations where the base is a N and the collocate is an Adj.

The grammatical patterns displayed by collocations include also relations between: 1) V and N, the N being the subject or the object of the V; 2) V and Adv; and 3) N and N. See the following table:

Language	Base	Collocate	Gram.Pattern
En.	***rain***	*heavy*	N-Adj
Es.	*lluvia*	*fuerte*	N-Adj
Fr.	*pluie*	*forte*	N-Adj
En.	***to rain***	*cats and dogs*	V-Adv
Sp.	*llover*	*a cántaros*	V-Adv
Fr	*pleuvoir*	*des cordes*	V-Adv
En.	***walk***	*take*	V-Obj
Es.	*paseo*	*dar*	V-Obj
Fr.	*promenade*	*faire*	V-Obj
En.	***secret***	*lies in*	V-Subj
Es.	*secreto*	*estriba en*	V-Subj
Fr.	*secret*	*resides dans*	V-Subj
En.	***chocolate***	*square*	N-N
Es.	*chocolate*	*onza*	N-N
Fr.	*chocolate*	*carré*	N-N

Table 1: Collocational equivalences following different grammatical patterns

Collocations are especially problematic for production, but not so much for comprehension. If a user of the dictionary needs an adjective expressing 'intense' when speaking about *rain*, he needs to find that *rain* combines with *heavy* in the entry for

rain. This is the normal procedure in collocation dictionaries such as the OCD: to provide the information under the entry of the base; i.e. the noun entry in the case of verbal, nominal and adjectival collocates (*rain, secret, chocolate*), and the verb entry in the case of adverbial collocates (*to rain*). However, in bilingual dictionaries, even if collocations are included, they are not identified as such, but are presented as a means of distinguishing senses, as I will show in the next section[2].

This poor arrangement of collocational information can be found in printed bilingual as well as in electronic dictionaries, since the latter, at least those compiled by mainstream publishers, have inherited the problems already present in printed versions. Nevertheless, electronic dictionaries allow us to retrieve hidden information more easily. Almost two decades ago, Fontenelle (1997) built a bilingual collocational database from a bilingual dictionary, although limited by the information contained in a machine readable dictionary. Nowadays, when online dictionaries rely on structured information in XML markup, the idea of "turning" a dictionary into a database is even more compelling.

This paper addresses the problem of how to exploit the collocational information included in the OSED, trying to take the first steps to fill the gap left by the absence of a Spanish–English dictionary of collocations[3]. As a result of the reorganization of the collocational information, it is possible to obtain lexical data for the germ of a Spanish–English collocation dictionary. These data can be used to compile either a dictionary in the strict sense of the term, or an online lexical tool to be exploited by platforms involved in machine translation or other applications. In the next section, I present how collocations are offered in the OSED in the part Spanish–English, and the different problems of accessibility that this display poses. Section 3 elaborates on a possible strategy to obtain a Spanish–English collocation dictionary by establishing different links between the XML tags. Section 4 focuses on the difficulties that this task presents in relation to the selection of the potential bases and to the selection of collocates in English. Finally, Section 5 draws some conclusions and presents an estimation of the viability of the final output.

2 Treatment of collocational information in the OSED

Putting combinatorial information under the collocate entry (instead of under the base)

[2] Within the Explanatory and Combinatorial Lexicology, a different conception of bilingual dictionary of collocations is claimed: a bilingual part aimed at selecting the translation equivalent of the base of the collocation, and a monolingual part where the collocation of the target language is described. See Alonso Ramos (2001), Meyer (1990) and Iordanskaja & Mel'čuk (1997).

[3] According to Ferrando (2012), the appearance of bilingual dictionaries of collocations is recent. This author mentions 1958 as the date of the publication of an English–Japanese dictionary. Nowadays, it is possible to find some bilingual dictionaries of collocations for other pairs of languages; for example, English–Russian (Benson & Benson, 1993), German–French (Ilgenfritz et al., 1989), German–Italian (Konecny & Autelli, 2014).

makes these entries very long and user-unfriendly to look at. The user has to scroll down long stretches of text in order to find the translation of a collocation, such as *poner atención* 'to pay attention', for example. This problem can be solved if the combinatorial information is placed under the entry for the base, in this case, the noun *atención.*

In what follows, I will present the different displays of collocational information in the OSED. There are three main strategies to present collocational information under the collocate entries:

- As an example, sometimes without a translation equivalent. For instance, under the entry for the adverb *encarecidamente*, we can find the collocation *pedir encarecidamente*. Note that no translation equivalent for the adverb is provided. See:

  1) a. *le* ***pido encarecidamente*** *que haga lo posible por ayudarlo*
     b. I **urge** o [formal] ***beg*** *you to do whatever you can to help him*

- As an equivalent construction, in a lemmatized form. For instance, under the entry for the adverb *perdidamente*, a translation equivalent, *hopelessly,* is provided and after that, the equivalent constructions are presented. See:

  2) a. *estar* ***perdidamente enamorado*** *de alguien*
     b. *to be* ***hopelessly in love*** *with somebody*

- By providing the Spanish base in brackets. There are two main distinctions: when the Spanish collocation has the syntactic pattern "N *de* N", the base is introduced with the preposition *de*. For instance, under the entry for the noun *grano*, different translation equivalents are supplied according to the different bases included in brackets. See:

  3) (*de trigo, arroz*) grain; (*de café*) bean; (*de mostaza*) seed

With all other syntactic patterns, the base is included in brackets[4]: a noun in brackets in the entries for adjectives or verbs, on the one hand, and a verb in the entry for adverbs, on the other hand. For instance, under the entry for the adjective *acérrimo*, two translations are provided depending on the noun included in brackets. See:

[4] Atkins and Rundell (2008: 217) refer to these sense indicators as *collocators*. In the jargon used in OUP, these words in brackets are called *collocates*, following the Sinclairian approach to collocations, whereby both elements of a collocation can be considered collocates, since no directionality in the relation is postulated. I would rather avoid this confusing terminology and will limit the term *collocate* to the lexical unit selected by the base. Corpas Pastor (in press) uses the term *collocational sets* for the series of potential collocates of a given base and/or the series of potential bases for a given collocate. However, in this dictionary only series of bases for a given collocate are displayed in this way.

4) (*partidario/defensor*) staunch; (*enemigo*) bitter

In a similar way, in the entry for the verb *cometer*, we find different translations associated with different nouns. See:

5) (*crimen/delito*) to commit; (*error/falta*) to make; (*pecado*) to commit

In this case, the noun acts as the grammatical object of the verb, but it also can be its grammatical subject. See for instance the entry for the verb *estallar*:

6) (*guerra/revuelta*) to break out; (*tormenta*) to break

The same procedure is used with collocations following the pattern "V+Adv", but not systematically. Thus, in the entry for the adverb *bulliciosamente*, we find two translations associated with different verbs. See:

7) (*festejar/protestar*) noisily; (*jugar*) boisterously

However, an adverbial collocate is not always treated in the same way. Sometimes the translation is given without any information about the base; for instance, under the entry for *radicalmente*, only the translation *radically* is found irrespective of the base. The possible explanation is that in Spanish as well as in English this adverbial collocate is selected by the verb *cambiar* or its equivalent in English *to change*. In other cases, a translation equivalent is provided, but different translations appear in the examples. This is the case of the adverb *definitivamente*. See:

8) (*resolver/rechazar*) once and for all

a. *el texto quedó terminado* ***definitivamente*** *en la sesión de ayer*
the text was finalized at yesterday's meeting (no translation)
the ***final*** *o* ***definitive*** *version of the text was drawn up at yesterday's meeting*
b. *mientras se resuelve* ***definitivamente*** *el problema*
while waiting for a ***final*** *o* ***definitive*** *solution to the problem*

None of these strategies have been devised to introduce collocational information, but rather to try to provide semantic cues in order to choose the best translation equivalent in the context of a given base.

Although it is not very frequent, it is also possible to find collocational information under the entry for the bases, especially for collocations following the syntactic pattern "V+N" or "N+V". This is done by means of examples. For instance, in the entry for the noun *guerra* ('war'), we find different verbal collocations in the examples. See:

9) a. *nos declararon la guerra*
b. *they declared war on us*

10) a. *están en guerra*

b. *they are at war*

11) a. *cuando estalló la guerra*
b. *when war broke out*

A further source of collocational information is what this dictionary calls *compounds*[5]. For instance, under the *café* entry, we find *café americano* ('large black coffee'), *café con leche* ('white coffee'), *café cortado* ('coffee with a dash of milk'), etc.

In sum, the procedures for including collocational information do not favour the use of the dictionary in terms of production. As stated in the introduction, an L1 Spanish user who wants to know how to say *coger una enfermedad* 'to catch an illness' in English will not find the answer in the entry for the noun, but in the entry for the verb, after scrolling through the rather long entry of *coger* in search of the translation *to catch an illness*. This procedure yields long entries highly difficult to look up. For instance, the entry for the verb *coger* offers 68 translations including senses and examples. With the removal of the translations linked to collocations, the entry would contain only 22 translations and would be, therefore, considerably more accessible. Some headwords functioning only as collocates could remain with the single role of providing part of speech or any other morphological information, but they would not need a whole entry. In the case of the adjective *mortal*, out of the four senses included in this entry, only the first one should remain, since the other three are collocates that should be given in the entry for the nouns in brackets. See:

12. (*ser*) mortal; (*herida*) fatal mortal; (*dosis*) fatal, lethal; (*odio/enemigo*) mortal; (*aburrimiento*) *fue un aburrimiento mortal – it was lethally (inglés norteamericano) o (inglés británico) deadly boring*

The inclusion of the collocational information under the collocate entry does not favour the use of the dictionary for comprehension either, due to the length of the entry and the lack of organisation in the microstructure. If an L1 English user wants to know what *coger* means with *enfermedad*, it is possible to devise an option consisting of launching a query which goes through the whole dictionary. In this way, entries for collocates will only be the result of a query[6].

After this overview of the treatment of collocational information in the OSED, the main conclusion is that it contains abundant information, but this is not appropriately

[5] The distinction between compounds and collocations is not trivial. As an illustration, in the Spanish part, the collocation *diente de ajo* ('clove of garlic') is treated as a compound, but in the English part, it is treated as other collocations: under the entry for the collocate *clove*, we find: "(of garlic) *diente*". For an overview of the distinction between compounds and collocations in Spanish, see Alonso Ramos (2009).

[6] Queries of this kind are already available, although some refinements would be necessary, since now they return not only collocations. See the query for COGER: http://www.oxforddictionaries.com/search/spanish-english/?q=COGER&multi=1.

organized nor displayed. In the next section, I put forward a proposal to build a bilingual collocational database with this information.

3 Taking advantage of implicit collocational information

The fact that the OSED relies on structured information with XML markup makes possible the retrieval of collocational information. Two tags are used to indicate special co-occurrences. These tags are <cs> and <co>. The first one is used to mark the noun acting as the typical subject of a given verb. For example, a typical subject of the verb *contagiarse* 'to spread' is the noun *enfermedad*. This information appears in the entry for the verb:

13. CONTAGIARSE$_{V}$

[<**cs** enfermedad> ('illness')] to spread, be transmitted

The tag <co> is more frequently used because it covers different relations: verb and object, noun and modifying adjective and finally, verb and adverb.

14. COGER$_{V}$

[<**co** enfermedad> ('illness')] to catch; [<**co** insolación> ('sunstroke')] to get

15. GRAVE$_{ADJ}$

[<**co** enfermedad> ('illness')] serious; [<**co** voz> ('voice')] deep

16. AUTOMATICAMENTE$_{ADV}$

[<$_{co}$ abrirse/cerrarse ('to open/ to close')] automatically

For the collocations following the pattern "N *de* N" as *grano de café* ('coffee bean'), a further tag is used: <ind>. This tag is also employed to introduce quasi-synonyms of the headword and, therefore, its automatic exploitation in retrieving collocational information is more complicated. Retrieving the collocations contained in the examples is not trivial either. All examples are tagged with the tag <ex> irrespective of whether or not they include collocations. For instance, under the entry for *pegar*, we find an example including a collocation and another including an idiom:

17) a. <ex no te acerques, que te **pego la gripe**_don't come near me, I'll give you my flu>
b. <ex la verdad es que **la pegamos** con su regalo_we really were dead on o spot on with her gift>

Therefore, this study will be limited to the information which can be more easily exploited automatically, the collocational pairs tagged as <co> and <cs>. After

extracting all the words tagged with <co/cs> and the headwords in the Spanish–English dictionary, I obtained a file with 21,358 pairs consisting of a noun linked with an adjective, a verb, and much less frequently, a verb with an adverb by means of the tags <co/cs>. The nouns appear in singular and in plural, and in some occasions with the article (see the entry for *romper* where we find <un amigo> or <un novio>). After the lemmatisation, there are 3024 words with the tag <co>, 140 of which are verbs and 2880 nouns; and 889 words with the tag <cs>, all of which are nouns, since this tag covers the relation between a noun as grammatical subject and the verb. The intersection between <cs> and <co> is 729 words. The total number of words disregarding the distinction between <co> and <cs> is 3184. This means that the bilingual collocational dictionary could have about 3184 bases for the Spanish part. By way of example, the verb *vivir* ('to live'), which appears tagged as <co> in the entry for the adverb *despreocupadamente* ('in a carefree way'), or the noun *zapato* 'shoe', which appears tagged as <co> in the entry for the adjective *plano* ('flat') and for the verb *acordonar* ('to lace') or as <cs> in the entry for the verb *apretar* ('to be too tight') are presented in an Excel file in the following way:

vivir	co	**despreocupadamente**
zapato	co	**plano**
zapatos	co	**acordonar**
zapatos	cs	**apretar**

Table 2: Sample of potential Spanish collocations

From this point, the procedure to be followed in order to build a collocational tool can be synthesized in the following steps:

1) Obtaining the English translation related to the tag <cs/co> from the entry for the Spanish headword. For example, in the XML markup entry for ATACAR and in the entry for CONTAGIAR :

   18) ATACAR

   <trg ><cs >virus/enfermedad</cs><**tr**>to attack</tr></trg>

   19) CONTAGIAR

   <trg ><cs >enfermedad</cs><**tr**>to spread</tr> <**tr**> to be transmitted</tr></trg>

2) Aligning the Spanish headword with the English translation in order to have the translation of collocates. For example:

   20) ATACAR –ATTACK

21) CONTAGIAR –TO SPREAD, TO BE TRANSMITTED

3) Aligning the Spanish and English collocates with the word tagged as <co/cs>. For example:

BASE	**SyntRel**	**COLL-ES**	**COLL-EN**
enfermedad	co	ARRASTRAR	DRAG ON
enfermedad	cs	ATACAR	ATTACK
enfermedad	co	ATAJAR	KEEP IN CHECK
enfermedad	co	BENIGNO	BENIGN

Table 3: Sample of bilingual collocational database

This file can be seen as a germ of a collocational dictionary since we have turned a file of headwords and the values of the tags into what can be a starting point of a bilingual collocational database consisting of a potential base, a syntactic relation and the collocate in both languages[7]. Not all words tagged as <co/cs> are equally productive: out of the total, only 214 are used 20 or more times; among them, there is the noun *persona* ('person'), which appears as <co/cs> in 1261 entries, and the noun *resultado* ('result'), which appears in 24 entries. After an exploration of the data, we can see different cases: highly productive values, as *persona* (1261), *ropa* 'cloth' (129), *animal* 'animal' (109), or *situación* 'situation' (103), and much less productive ones, such as *acceso* 'access' (3), *abanico* 'fan' (2) or *abeja* 'bee' (1). The four former nouns are the most frequently used, but note the difference between the first and the second noun: from 1261 to 129 entries. About 1600 words are used only in one entry. However, between the highly productive words and the very unproductive ones, there is a significant number of words that can become the bases of a collocational entry with 30 or 40 collocates in average. For instance, the noun *enfermedad* ('illness') will contain 42 bilingual collocations; the entry for *acuerdo* ('agreement') will contain about 25 collocates, etc. The entries for these nouns in some collocational dictionaries in the respective languages are much longer (for instance, the entry for *agreement* in OCD contains 56 collocates and the entry for *acuerdo* in DCP 179). However, since a bilingual Spanish–English collocation dictionary does not yet exist, poor entries are

[7] Note that in this way we do not obtain the translation of the base. This translation should follow another strategy based on semantic grounds to be described in the bilingual dictionary, rather than in the collocational bilingual dictionary. For instance, translating *enfermedad* as *sickness*, *disease* or *illness* does not depend on which are its collocates, but on semantic differences existing between these three English equivalents. See the help note that appears under the CSED dictionary: http://www.collinsdictionary.com/dictionary/spanish-english/enfermedad?showCookiePolicy=trueootnote 1.

better than no entries. This file is merely a starting point because it also needs to be filtered. Some distinctions should be established among the words tagged as <co/cs>, which will result in that many pairs will not be part of the collocational tool. In what follows, I will focus on the difficulties or challenges regarding the selection of bases and the selection of the translation of collocates.

4 Filtering Spanish bases and English collocates

In order to arrive at the situation depicted in Table 3, it is necessary, first, to be sure that the relation between the word tagged as <co/cs> and the headword is a collocational relation. Secondly, it is necessary to identify with precision which is the translation equivalent, since, in many cases, the OSDE does not propose any and gives only an example.

4.1 Selection of bases: semantic and lexical tags

As I have pointed out, the purpose of the words in brackets is to help to find the translation of the headwords in combination with these words, not necessarily to give collocational information. For this reason, the words tagged as <co/cs> sometimes represent meanings and sometimes stand for lexical units. In the first case, I will call them *semantic tags*, and in the second case, *lexical tags*. Words are used as semantic tags when their role is to provide a semantic restriction on the nouns that can instantiate the object of a verb[8]. For instance, under the entry of *coger*, we can find:

22) [trabajo ('work')/casa ('house')] to take

The example provided for that sense is:

23) *no puedo coger más clases – I can't take on any more classes*

The nouns <[trabajo/casa]> restrict semantically what could be the object of *coger* when it means 'to accept', but it is possible to use the verb *coger* without these nouns as well, as illustrated with the example: *no puedo coger más clases* ('I can't take on any more classes'). Here we do not have the word *trabajo* ('work'), but the meaning 'trabajo', which can be associated to the meaning of (dar) *clases* 'to teach'.

In contrast, most occurrences of <cs/co> are lexical tags. By lexical tag, I mean the specific word or lexical unit that is combined with the headword. For instance, again in the entry of *coger*, we find:

24) [tren ('train')/autobus ('bus')/taxi] to catch, take

The three nouns in brackets are given to provide the translation of the collocations

[8] Regarding the role of selectional restrictions and collocations as markers of senses in the dictionaries, see Atkins and Rundell (2008: 302).

resulting from combining *coger* with any of these nouns, as *coger un tren ('to catch a train*').

The problem is that it is not always clear for the user when the tag is used as a semantic restriction, i.e. as a semantic cue to help find the correct meaning of the headword, and when it is used as a lexical tag, i.e. when it specifies the base of a specific collocation which serves to give the translation of this collocation. This ambiguity will make the automatic treatment difficult. For instance, under the entry for the verb *acometer* 'to undertake', we can find:

25) [empresa ('undertaking')/proyecto ('project')] to undertake, tackle

With this information, it is not possible to know with certainty when the word tagged as <co/cs> is representative of a semantic group and when it is only a specific combination. For instance, the noun *tarea* ('task') inherits the collocate *acometer,* because *tarea* can be considered a hyponym of *empresa* or *proyecto*, but it is not explicitly indicated. In the case of semantic tags with this hyperonymic role, it would be useful to study the possibility of automatically deriving collocations by means of some formalism establishing paradigmatic relationships such as Eurowordnet (Vossen, 1998). For instance, if under the entry of the verb *abandonar* ('to abandon'), the noun *actividad* ('activity') is treated as a <co>, all nouns which are considered activities could inherit the collocate *abandonar*: *estudios* ('studies'), *lucha* ('fight'), *curso* ('course'), etc. Therefore, by using some formalism which serves to infer relationships, the initial collocational database could be enriched with new information. However, the formalism should also have the possibility of blocking the inheritance for tags as *persona* 'person' which most of the time represents a semantic restriction and can be eliminated as a potential base to be included in a collocational tool. Thus, in the entry for *abandonar*, it is possible to find *persona* as <co>, as in the following examples of the OSED:

26) a. *abandonó a su familia – he abandoned* o *deserted his family*
b. *abandonó al bebé en la puerta del hospital– she abandoned* o *left the baby at the entrance to the hospital*

Nevertheless, the combinations *abandon his family/the baby* are not collocations. Here, the tag <persona> is used to outline the meaning of *abandonar*, but *abandonar* is not a lexical unit selected by the nouns *familia* or *bebé*. Therefore, pairs such as "persona-abandonar" should be eliminated of the collocational database.

In sum, from the initial file, some potential bases should be eliminated, such as *persona* because it is mostly used as a semantic restriction, but some others could be added by using some formalism handling inheritance relationships.

4.2 Selection of the translation of collocates

The policy of the OSED is not very systematic with respect to the way of providing

translation equivalents for collocates. In the ideal situation, we would have a translation equivalent of the collocate with an example in both languages. Thus, under *alcanzar*, we find:

27) (acuerdo) to reach
los acuerdos alcanzados en materia de desarme
the agreements reached in the field of disarmament

This information could be easily turned into a bilingual collocation entry:

BASE	**SyntRel**	**COLL-ES**	**COLL-EN**
acuerdo	co	ALCANZAR	TO REACH

Table 4: Bilingual collocation entry

However, the OSED does not always provide a translation equivalent and sometimes gives only an example. In these cases, several possibilities exist:

1. The translation equivalent is recoverable from the example. We have two parallel collocations in the two languages. See the entry for *levantar*:

28) (ojos)
me contestó sin ***levantar los ojos*** *del libro*
she answered me without looking up o without ***lifting her eyes*** *from her book*

From the example, the following equivalence could be established, through an automatic syntactic parsing:

BASE	**SyntRel**	**COLL-ES**	**COLL-EN**
ojos	co	LEVANTAR	TO LIFT

Table 5: Bilingual collocation entry

2. The translation equivalent represents a different construction in English. This kind of mismatch is very frequent when comparing collocations in different languages (see Mel'čuk & Wanner, 2001). For instance, under the entry of *arder* ('to burn'), we find:

29) (estómago)
me arde el estómago
I've got heartburn

In Spanish, the noun *estómago* ('stomach') is the subject of the verb *arder* 'to burn', but the English noun *heartburn* is not the translation of *estómago*: this noun expresses

the meaning expressed by the verb *arder* in Spanish. In this case, the correspondence between both collocations is more difficult to be derived automatically, because the following mapping is wrong:

30) estómago arder to have got

When the meaning of the collocation is distributed between the base and the collocate in different ways in both languages, it is necessary to give the translation of the base (see footnote 6).

Another example, similar to the previous one, could be the mismatch between a light verb construction in Spanish and a single verb in English. In Spanish, it is possible to express the meaning *golpe* ('blow') by the suffix -azo, as in *codazo* 'blow given with the elbow'. Any noun created in this way selects a light verb such as *dar* 'give' or in Mexico *arrimar*. In contrast, English uses a single verb *to elbow*. In the entry for *arrimar*, we find:

31) (golpe)
me arrimó un codazo – he elbowed me

In this case, the correspondence is between a collocation and a single verb.

3. In some occasions, lexical gaps prevent a translation. Consequently, the OSED provides a paraphrase of the Spanish collocation. This is the case of *habitación interior*:

32) (habitación/piso) *(with windows facing onto a central staircase or patio)*

5 Conclusion

This paper has described the process of construction of a bilingual collocation database from information already included in an online bilingual dictionary. The approach of reusing existing resources was frequently used in the beginning of the 1990s, but even though nowadays NLP applications tend to rely on big corpora by extracting linguistic knowledge from statistical regularities, I believe that the lexicon is still necessary; especially a lexicon which has been informed by lexicographers. The construction of lexicons from scratch continues to be time-consuming and costly, as in the time when Fontenelle (1997) proposed his collocational database. For this reason I consider that it is worth the effort to reuse the collocational information included in the OSED. This approach of reusing previous lexicographic work can be complemented with current techniques of extracting collocational information from a parallel Spanish–English corpus, especially to provide frequency information. In this way the bilingual collocation dictionary would be corpus-based, not corpus-driven, because the collocations have been established previously in the OSED, not induced from a corpus. Nonetheless, if the final goal is to build a comprehensive bilingual

collocation dictionary, the information extracted from the OSED should be complemented by corpus-induced combinatorial information.

The work presented here only concerned the Spanish–English part of the OSED, but it can be assumed that a similar XML encoding is used in all other bilingual dictionaries from this publisher. Therefore, the potential of bilingual collocational databases is big. As pointed out earlier, the bases and the translations in the database need to be filtered by lexicographers, but according to my estimates this task is not especially time-demanding. In order to obtain a definitive collocational database, technological and lexicographical skills are needed. First, it is necessary to implement a program which automatically establishes the new links between the words involved in collocations. Second, collocational relationships need to be verified by expert lexicographers.

As a possible future line of research, the bilingual collocation database could also be enriched with the *lexical functions* (Mel'čuk, 1996). The apparatus of lexical functions is used in the dictionaries issued from the Explanatory and Combinatorial Lexicology to describe semantically and syntactically collocations:

IncepOper_1(enfermedad) = coger, pillar

IncepOper_1(illness) = to catch

The role of *interlingua* played by the lexical functions could be exploited for search engines involved in machine translation or in information retrieval since they can be used for sense disambiguation. Finally, collocations tagged with lexical functions are unquestionably useful in the field of second language learning.

6 Acknowledgements

The work presented this paper has been partially supported by the Spanish Ministry of Economy and Competitiveness (MINECO) and the FEDER Funds of the European Commission under the contract number FFI2011-30219-C02-01. I would like to express my gratitude for the help given by the team working in the Global Division of Oxford University Press during my stay in 2014. I would also like to thank Marcos García Salido and Orsolya Vincze for their careful reading and fruitful comments.

7 References

Alonso Ramos, M. (2001). Construction d'une base de données des collocations bilingue français-espagnol. *Langages*, *143*, pp. 5-27.

Alonso Ramos, M. (2009). Delimitando la intersección entre composición y fraseología. *Lingüística española actual*, 31(2), pp. 5-37.

Atkins, S.B.T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.

Benson, M. & Benson, E. (1993). *Russian English Dictionary of Verbal Collocations*, Amsterdam/Philadelphia: John Benjamins.

Corpas Pastor, G. (in press). Collocations in e-bilingual dictionaries: from underlying theoretical assumptions to practical lexicography and translation issues". In S. Torner & E. Bernal (eds.). *Collocations and other lexical combinations in Spanish. Theoretical and Applied approaches*. Chicago, IL: Ohio State University Press.

CSED: Collins Spanish-English Dictionary (On-line version). Accessed at:www.collinsdictionary.com/dictionary/spanish-english/ (23 May 2015)

DCP: Bosque, I. (dir.) (2006). *Diccionario combinatorio práctico del español contemporáneo*. Madrid: SM.

Ferrando, V. (2012). *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE*. PhD Dissertation. Tarragona: Universitat Rovira i Virgili.

Fontenelle, T. (1997). *Turning a Bilingual Dictionary into a Lexical Semantic Database*. Tübingen: Max Niemeyer Verlag.

Ilgenfritz, P., Stephan-Gabinel, N. & Schneider, G. (1989). *Langenscheidts Kontextwörterbuch Französisch-Deutch*, Berlin/München: Langenscheidt.

Iordanskaja L. & Mel'čuk, I. (1997). Le corps humain en russe et en français. Vers un Dictionnaire explicatif et combinatoire bilingue. *Cahiers de Lexicologie*, 70(1), pp. 103-135.

Konecny, C. & Autelli, E. (2014). *Kollokationen Italienisch-Deutsch*. Hamburg: Helmut Buske.

Makkai, A. (1972) *Idioms Structure in English*. The Hague: Mouton.

Mel'čuk, I. (1996). Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon". In L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: John Benjamins, pp. 37-102

Mel'čuk, I. (2012). Phraseology in the Language, in the Dictionary, and in the Computer. *Yearbook of Phraseology*, 3 (1), pp. 31–56.

Mel'čuk, I. & Wanner, L. (2001). Towards a Lexicographic Approach to Lexical Transfer in Machine Translation (Illustrated by the German-Russian Language Pair). *Machine Translation*, 16(1), pp. 21-87.

Meyer I. (1990). Interlingual Meaning-Text Lexicography: Towards a New Type of Dictionary for Translation. In J. Steele (éd.), *Meaning-Text Theory: Linguistics, Lexicography, and Applications*, Ottawa: University of Ottawa, Ottawa, pp. 175-270.

OCDSE: *Oxford Collocations Dictionary for Students of English*. (2009). 2nd edition. Oxford: Oxford University Press.

OSD: Oxford Spanish-English Dictionary (On-line version). Accessed at www.oxforddictionaries.com/translate/spanish-english/ (23 May 2015).

Vossen. P. (1998). Eurowordnet. A multilingual database with lexical semantic networks for European Languages. Dordrecht: Kluwer Academic Publishers.