Enriching the Spanish EuroWordNet by Collocations

Leo Wanner[1], Margarita Alonso Ramos[2] and Antonia Martí[3]

[1]IULA, Universitat Pompeu Fabra, Barcelona, Spain
[2]Universidade da Coruña, A Coruña, Spain
[3]Universitat de Barcelona, Barcelona, Spain
leo.wanner@upf.edu, lxalonso@udc.es, amarti@ub.edu

Abstract

Collocations constitute an important type of syntagmatic information whose introduction into WordNets has not yet been addressed. The goal of our work is the integration of the collocational material for the field of emotion nouns encoded in the *DIccionario de colocaciones del español* (DICE) in terms of Lexical Functions into the Spanish part of the EuroWordNet (SpEWN). Two features of collocations are decisive in connection with their representation in SpEWN: (i) they are variant-specific rather than synset-specific and (ii) depending on the degree of their idiosyncrasy, they may be generalized to a certain degree. These features are accounted for by introducing new structures into the SpEWN. These new structures are compatible with the general design principles of the EWN. Given that SpEWN and DICE reveal a diverging distinction of the senses of lexical items, prior to the introduction of collocations from DICE into SpEWN, the senses of the elements of collocations in SpEWN and DICE are aligned.

1. Introduction

WordNet, WN (Fellbaum, 1998) has originally been designed as a paradigmatic lexical data base of English with relations such as *synonymy*, *hyperonymy*, and *meronymy* defined between sets of quasi-synonymous lexical items (= *synsets*). To provide a more comprehensive lexical resource – also in order to meet the needs of NLP-applications – WN and its descendants are extended by other types of linguistic information. The information being considered for inclusion concerns, for instance, the domain (or *field*) of a given synset (Magnini & Cavaglià, 2000), its metaphorical (Alonge & Castelli, 2002; Lönneker, 2003) and phraseological elements (Bentivogli & Pianta, 2004), subcategorization and selection restrictions (Fellbaum, 1998; Agirre & Martínez, 2002), etc. We focus on yet another type of information whose representation has not been tackled so far in the WN-model: *collocations*. The availability of collocations is essential for NLP-applications such as *Information Retrieval*, *Word Sense Disambiguation* and *Information Extraction*. The goal of our work is to extend the Spanish part of one of the multilingual descendants of WN, the EuroWordNet, EWN (Vossen, 1998), by collocations.

A collocation is a binary (to a varying degree idiosyncratic) combination of lexical items such that one of the items possesses its full semantics and the other item reveals a semantics that depends on the meaning expressed by the combination as a whole.[1] Cf., e.g., *bitter*/*deep*/*intense*/ *violent*/… *HATRED*, where *HATRED* keeps its semantics and *bitter*, *deep*, etc. express the meaning 'intense'. The former is called the *base* (or *keyword*) of the collocation, and each of the latter – a *collocate*. Collocates and thus collocations as a whole can be semantically classified. For instance, all collocates in the above set have the meaning 'intense'; all verbal collocates in *give* [*a*] *lecture*, *take* [*a*] *walk*, *make* [*a*] *statement*, *deliver* [*a*] *speech*, etc. mean 'perform', and so on. The most detailed semantic typology of collocations available to date is the typology of *Lexical Functions* (LFs) (Mel'cuk, 1996). We use LFs to encode collocations in the Spanish EuroWordNet (henceforth SpEWN). We start from the collocation data on Spanish emotion nouns encoded in the *DIccionario de Colocaciones del Español* (DICE) (Alonso Ramos, 2003). DICE contains about 3,400 LF-encoded collocational relations of emotion nouns. Further LF-classified collocations are acquired automatically (Wanner, 2004).

Two features of collocations are decisive in connection with their representation in EWN: First, collocations are *variant*-specific rather than *synset*-specific (in contrast to paradigmatic relations). Thus, *cólera* 'anger', *ira* 'rage' and *enojo* 'anger' constitute a synset. However, the collocate *ciego* 'blind' is selected only by the two first variants, not by the third one: *ciego de cólera/ira/*enojo*. Second, depending on the degree of their idiosyncrasy, collocations may be generalizable to a certain extent. Three different cases must be distinguished: (i) all keywords that are described by the same *base concept* (in the sense of EWN; cf. Vossen, 1998) co-occur with the same collocate(s); e.g., all 'feeling' nouns co-occur with *sentir* '[to] feel'; (ii) a subset of keywords that are described by the same base concept co-occurs with the same collocate(s); e.g., *esperanza* 'hope', *menosprecio* 'scorn', *odio* 'hatred', *remordimiento* 'remorse', *rencor* 'grudge', and *sospecha* 'suspicion' all co-occur with *abrigar* '[to] harbor'; (iii) only one or a few keywords co-occur with the same collocate(s); e.g., only *odio* 'hatred' co-occurs with *mortal* 'deadly'.

However, that despite this potential for generalization, collocations must be considered from the angle of individual lexemes. That is, our task is not just a matter of labour (i.e., adding another piece of information in an existing theoretical framework), but also of the extension of the theoretical framework of the SpEWN. In the next section, we introduce DICE. In Section 3, the alignment of senses in SpEWN and in DICE is discussed, and in Section 4 the structures that we introduce into SpEWN to accommodate for collocations. Finally, Section 5 contains the summary, conclusions and directions of future work.

2. The Source of Collocations: DICE

DICE is an ongoing lexicographic project on the compilation of an on-line dictionary of collocations for Spanish encoded in terms of LFs. The DICE is organized in terms of semantic fields. The field that is most widely

[1] A different definition of the notion of collocation that is not compatible with ours is based on frequency: lexical items that co-occur sufficiently often together form a collocation.

covered is the field of emotion nouns, which is particularly rich in collocations. The organization of DICE is strictly base-oriented, i.e., only a base of a collocation can appear as lemma. Thus, in the case of the collocation *despertar odio* '[to] incite hatred', only *odio* will appear in the lemma list, *despertar* will not. However, the implementation of DICE in a relational data base also allows for an access of the information from the collocate side – for instance by a search like "which nominal items co-occur with *despertar* in its causative sense". The field of emotion nouns illustrates well the characteristic features of collocations that must be taken into account for the integration of DICE into SpEWN. Thus, on the one hand, *entrar* lit. '[to] enter' co-occurs with *alegría* (*entrar alegría a X*) but not with its quasi-synonym *contento* (**entrar contento a X*). On the other hand, *entrar* is very productive as an inchoative collocate in combination with emotion nouns: *entrar pena* 'pity'/*ganas* 'desire'/*vergüenza* 'shame'/*alegría* 'joy'/*miedo* 'fear'/…
For a detailed description of the organization of DICE and the types of information provided for each lemma in DICE, see (Alonso Ramos, 2003). For illustration, consider a fragment of the entry for the noun *alegría***1a** 'joy':[2]

N+ADJ Collocations
Magn ('intense'):*grande, intensa, loca, desbordante, indescriptible, extraordinaria, indecible, a raudales, inmensa, enorme, inefable*
Ver ('real'): *verdadera, sincera, franca*
AntiVer ('not real'): *forzada, fingida*
ADJ *de* N Collocations
Magn+A$_1$ ('X is very happy'): *resplandeciente de, rebosante de, lleno de, exultante de*
V+N Collocations
Oper$_1$ ('to feel'): *sentir, tener, llevarse*
CausFunc$_1$ ('to cause'): *causar, dar, despertar, producir, provocar*
CausDegrad ('to spoil'): *nublar, perturbar, turbar, enturbar*
N+V Collocations
Func$_1$ ('to exist'): *reinar en*
Func$_2$('to come from'): *nacer de, emanar de*
IncepFunc$_1$ ('to begin'): *entrar a, embargar a*
Prep+N Collocations
Adv$_2$: *para* ~ de X
N *de* N Collocations
Gener ('generic'): *sentimiento de* ~

Figure 1: Fragment of a lexical entry from DICE

In order to make DICE compatible with SpEWN, the bases and collocates in DICE are assigned the corresponding synset number in the SpEWN. Thus, *cólera*, *ira* and *enojo*, which constitute a synset in SpEWN, receive in DICE the same pointer (<**04807941**>) to SpEWN. Each of the collocates *aliviar*, *aplacar* and *suavizar* is assigned the synset number <**01005913**> because the three variants belong to the same synset. And so on.

2 The names in bold (**Magn**, **Ver**, etc.) are names of LFs; in parentheses, LF-glosses are given. Due to the lack of space, the Spanish collocates are not translated.

3. Examination of SpEWN

A major problem for the integration of any two lexical resources is their diverging distinction of the senses of lexical items. As is well-known, dictionaries often display extremely different senses (Fillmore *et al*., 1994). The different WNs have often been cited as resources with fine-grained sense distinction. A number of works focuses on the clustering of WN-senses (Peters *et al.* 1998). In DICE, the distinction of senses is equally given special attention. Therefore, a thorough study of the compatibility of the distinctions made in SpEWN and in DICE prior to the integration seems appropriate.

3.1 On the Distinction of Senses in SpEWN

SpEWN has been derived semi-automatically from the English WN (Vossen, 1998). This led to a strong bias of the sense distinction in SpEWN towards English. As a consequence, SpEWN multiplies the senses for Spanish words that correspond to several words in English – even when in Spanish no sense distinction can be detected. Thus, the noun *alegría*, which possesses only one reading of feeling, receives in SpEWN two 'feeling' senses, one of them being linked to the English *gladfulness* and the other to *joy*. On the other hand, SpEWN does not introduce necessary sense distinctions of a Spanish word if the English equivalent lacks them. For instance, the meaning of *llenar* lit. 'to fill' as it appears in collocation with *alegría* (*llenar de alegría*) is not covered by any sense in SpEWN because in English no equivalent collocation is available (the equivalent of *llenar de alegria* is the single verbal lexeme *to overjoy*; cf. *La noticia nos llenó de alegría* vs. *We were overjoyed by the news*). That is, in order to make DICE and SpEWN compatible with respect to the distinction of senses, we have to: (i) reduce the polysemy in SpEWN in the case of unjustified sense inflation (which will be discussed with respect to bases); (ii) introduce new senses in the case they are not available in SpEWN (which will be discussed with respect to collocates).

3.2 Aligning the Senses of Bases

To illustrate the procedure of base sense aligning, we use the noun *ansia*. Consider Figure 2.

04789334n	lock 5 **craving_1**
feeling	lock 5 **ansia_1 regosto_1**
04790562n	lock 0 **hankering_1 yen_1**
feeling	lock 0 **ansia_2**
	lock 1 **avidity_1 keenness_2 eagerness_1**
04829857n	**avidness_1**
feeling	lock 1 **afán_2 anhelo_2 avidez_2 ardor_5 ansia_3**
08684458n	lock 6 **nausea_1 sickness_2**
state	lock 6 **náusea_2 mareo_3 ansia_4 asco_3 arcada_2 basca_3**
	lock 9 **anxiety_2**
04812078n	lock 9 **preocupación_5 intranquilidad_7**
feeling	**inquietud_5 ansiedad_2 ansia_5 zozobra_3 desasosiego_5 angustia_6**
08580489n	lock 0 **passion_2 rage_2**
state	lock 0 **pasión_6 ansia_6**

Figure 2: The six synsets of *ansia* in SpEWN

Ansia with the meaning 'desire' appears in the first three synsets shown in Figure 2. However, *ansia*_1, *ansia*_2, and *ansia*_3 are merely translations of different English words, they do not stand for different senses in Spanish. In contrast, the introduction of distinct senses for *ansia*_4 and *ansia*_5 is well justified. This distinction is also buttressed by their different co-occurrence. Cf., e.g., the co-occurrence of *ansia* 'anxiety' (i.e. *ansia*_5) when contrasted to that of *ansia* 'desire' (i.e. *ansia*_1). Only *ansia* in the sense of *ansia*_1 can be modified by intensifiers such as *insaciable*, *incontenible*, *irrefrenable* and co-occur with verbs such as *exacerbar*, *alimentar*, *avivar*, etc. All of these verbs also co-occur with the noun *deseo* 'desire'. *Ansia* in the sense of *ansia*_5, shares collocates with *ansiedad* 'anxiety' and with *angustia* 'anguish'. The sense distinction *ansia*_6 glossed as 'lo que se desea' (what is desired) is, again, not well justified. In Spanish, in contrast to *deseo* 'desire', *ansia* lacks the meaning 'what is desired': *Hacerte feliz es mi mayor deseo* 'To make you happy is my great desire' vs. **Hacerte feliz es mi mayor ansia*. That is, the six senses of *ansia* in SpEWN are reduced in DICE to three senses; cf. Figure 3.

ANSIA**1**: *Un hombre que había pagado con su vida su ansia de libertad*
EWN: <**04789334**>+<**04790562**>+<**04829857**>
ANSIA**2**: *No pases ansia, seguro que están todos bien*
EWN: <**04812078**>
ANSIA**3**: *Después le vinieron las ansias y Marialuisa y Juanita se la llevaron hacia el gallinero para que vomitase*.
EWN: <**08684458**>

Figure 3: Reduction of the senses of *ansia*

As a result, the variants *ansia*_1, *ansia*_2 and *ansia*_3, in the SpEWN will share the same collocational information, *ansia*_4 and *ansia*_5 will receive collocational information of their own, and *ansia*_6 will not receive any collocation link because this sense is erroneous. That is, by aligning the senses of bases, we cluster several variants, and we mark the variants whose introduction is not justified in Spanish.

3.3 Aligning the Senses of Collocates

Collocate lexemes have a vague status in most of the lexical repositories. Therefore, it is not surprising that the amount of data on collocate lexemes is rather unbalanced in SpEWN. A study of the collocates of emotion nouns in SpEWN reveals three cases: (a) a collocate meaning is not given; (b) the collocate meaning is given, but it is too specific; and (c) the collocate meaning is given, with the semantics of the corresponding base being indicated as well as base concept, top concept, or in the gloss of the collocate.
The lack of collocates can be due to incomplete data, which is frequent in any lexical resource, but, again, also due to the bias towards English already identified above. Thus, *mortal* in sense 'intense' is not available in SpEWN for the collocation *odio mortal* 'mortal hatred', but it is available in co-occurrence with *enemigo* 'enemy' (*enemigo mortal*). This is because in English, *mortal* and *deadly* do not co-occur with feeling nouns (cf. *bitter hatred*), but rather with the noun enemy (*mortal enemy*).

The data on collocates in SpEWN are too specific in a number of cases. For instance, the verb *coger* in the sense 'to begin to have' and *contraer* 'to contract' appear in the same synset. Therefore, this synset can only be associated with nouns denoting illnesses, and, e.g., the collocation *coger cariño* 'to begin to have affection' is not present in SpEWN.
The third case is the most interesting for our goal. Certain synsets encode collocate meanings characterized by the base concept 'emotion'; cf. Figure 4.

01005913v emotion	lock 7 **appease_1 placate_1 pacify_1 mollify_1 lenify_1 gruntle_1 gentle_1 assuage_1** lock 0 **aliviar_6 suavizar_7 sosegar_2 pacificar_1 aplacar_4 apaciguar_4**

Figure 4: A collocate synset for emotion nouns

If available, the gloss of the synset can serve as a hint that the synset in question is a collocate. In the same way as in traditional dictionaries the definition of a collocate includes the corresponding bases in parentheses, in the following example of SpEWN, the gloss indicates the possible bases of the collocate variants; cf. Figure 5.

01033880v emotion	lock 5 **allay_1 still_3** of anxieties and fears **relieve_4 ease_4** 1 **aliviar_7 disipar_2 calmar_6**

Figure 5: Example of a collocate gloss in SpEWN

4. Defining Collocation Structures

The extension we propose follows the general design principles of the EWN. In analogy to the notion of synset, two new notions are introduced: *key*(*word*)*set* and *coll*(*ocate*)*set*. A *keyset* is a set composed of lexemes that share one or more collocates. A *collset* is a set composed of lexemes that are collocates of the same keyword(s).
In order to capture the different degrees of generalization of the collocations, we need three different structures (called *indexes*): (1) a *field index* (FI), (2) a *keyword index* (KI), and (3) a *collocate index* (CI).
An FI is an extension of the notion of the *base concept* in EWN. It specifies, for an individual base concept, the list of collocations (given in terms of LFs) shared by all synsets characterized by this base concept. Furthermore, it contains a list of pointers to collsets or to individual collocates of each collocation. Thus, for the base concept 'feeling', the following FI-record would be generated (the LF-gloss may be omitted since it is redundant; we cite it here for higher transparency):

[…]
1 base_concept “feeling”
1 collocations
2 LF Oper1
3 LF_GLOSS “perform, experience, or be in a state”
3 WORDNET_POINTER <01008772v>
2 LF Magn
3 LF_GLOSS “intense”

3 WORDNET_POINTER <01149202a>
2 LF …

Figure 6: A Fragment of a Field Index for 'feeling'

That is, in this case, *alegria* 'joy', *miedo* 'fear', *odio* `hatred´, and all other 'feeling'-nouns would inherit the information that they co-occur with the verbal lexeme referred to by the pointer <01008772v> (namely *sentir* '[to] feel'), with the adjectival lexeme referred to by <01149202a>, etc. A KI specifies the collocations for a keyset or an individual keyword; cf. two examples:

[...]
1 base_concept “feeling”
1 keyset {esperanza:2 menosprecio:5 odio:1 remordimiento:2 rencor:2 sospecha:1}
1 collocations
2 LF Oper1
3 LF_GLOSS “perform, experience, or be in a state”
3 WORDNET_POINTER <01009946v>

[...]
1 base_concept “feeling”
1 keyset {alegria:1 gozo:2 orgullo:3 felicidad:1}
1 collocations
2 LF Magn+A1
3 LF_GLOSS “X experiencing an intense feeling”
3 WORDNET_POINTER <TARGET_CI>

Figure 7: Fragments of two keyword indexes

In the first example, the keyset shares one collocate, in the second example a collset. A CI is inverse to the KI. Its record contains a collset or an individual collocate, the LF-names of collocations they occur in, and a pointer to an individual base or a keyword index of a keyset. For instance, the TARGET_CI in the last KI-example looks as follows:

[...]
1 collset {rebosante:2 radiante:1 desbordante:1}
1 collocations
2 LF Magn+A1
3 LF_GLOSS “X experiencing an intense feeling”
3 WORDNET_POINTER <TARGET_KI>

Figure 8: Fragment of a collocate index

The proposed structures are very flexible. Thus, they can be used to define collocation relations between both individual lexemes (variants) and sets of lexemes, and they can be used in both directions: from the base of a collocation to the collocate, and vice versa. That is, they allow for both an efficient introduction of collocational information and its retrieval.

5. Summary and Conclusions

Our proposal to add new structures for the representation of collocational information seeks to be in agreement with the philosophy of EWN. As Miller (1998) points out, the notion of synonymy in WordNet does not entail interchangeability in all contexts. As became clear from the discussion above, one of the reasons why two synonyms cannot be interchanged are their collocational relations. Let us give here a final example. Thus, the two variants of the synset <*coger*_1 *contraer*_1> cannot be interchanged in all contexts because even if both co-occur with all nouns of diseases, only *coger*_1 co-occurs with many 'feeling'- nouns. In our proposal, the variants *coger*_1, *contraer*_1 constitute together with *atrapar*_2, *pillar*_2 a collocation set which is selected by the nouns of diseases. *Coger*_10, *cobrar*_2, *concebir*_2, and *tomar*_11 constitute another collocation set, which is selected by a specific subset of nouns of feeling. In sum, the addition of collocational information also shows how synsets can be reorganized from the syntagmatic point of view.

So far, our work focused on the linguistic aspects of the integration of collocations from DICE in SpEWN. Future work includes the implementation of the structures introduced above, the implementation of a user interface to ensure access to collocational information by human users, and research on the multilingual representation of collocational information in EWN.

References

Agirre. E. & Martínez, D. (2002) Integrating selectional preferences in WordNet. In Proceedings of the First Global WN-Conference. Mysore, India.

Alonge. A. & Castelli, M. (2002) Metaphoric expressions: an analysis of data from a corpus and the ItalWordNet database. In Proceedings of the First Global WN-Conference (pp. 3432-350). Mysore, India.

Alonso Ramos, M.M. (2003). Hacia un Diccionario de Colocaciones del español y su codificación. In M.A. Martí et al. (eds.). Lexicografía computacional y semántica (pp. 11-34). Barcelona: Edicions de l’Universitat de Barcelona.

Bentivogli, L & Pianta, E. (2004). Extending with Syntagmatic Information. In Proceedings of the Second Global WN-Conference. Brno, Czech Republic.

Fellbaum, C. (ed.) (1998). WordNet. Cambridge: MIT Press.

Fillmore, Ch. J. & Atkins, B. T. S. (1994): Starting where the dictionaries stop: The challenge for computational lexicography, In Atkins, B. T. S. & A. Zampolli (eds.) Computational Approaches to the Lexicon (pp. 349-393). Oxford: Oxford University Press.

Lönneker, B. (2003). Is there a way to represent metaphors in WordNets? In Proceedings of the ACL Workshop on the Lexicon and Figurative Language, ACL 2003 (pp 18-26). Sapporo, Japan: ACL.

Magnini, B. & Cavaglià, G. (2000). Integrating subject field codes into WordNet. In Proceedings of LREC 2000. Athens, Greece.

Mel’cuk, I. (1996). Lexical Functions. In L. Wanner (ed.) Lexical Functions in Lexicography and NLP (pp. 37-102). Amsterdam: Benjamins.

Miller, G. (1998). Nouns in WordNet. In C. Fellbaum. (ed.) WordNet (pp. 23-46). Cambridge: MIT Press.

Peters, W., Peters, I. & P. Vossen (1998). Proceedings of LREC 1998.

Vossen, P. (ed.) (1998). EuroWordNet. Dordrecht: Kluwer.

Wanner, L. (2004). Towards Automatic Fine-Grained Semantic Classification of Verb-Noun Collocations. In Natural Language Engineering Journal, 10(2).