The Spanish Version of WordNet 3.0

Ana Fernández-Montraveta, Gloria Vázquez and Christiane Fellbaum

Abstract. In this paper we present the Spanish version of WordNet 3.0. The English resource includes the glosses (definitions and examples) and the labelling of senses with WordNet identifiers. We have translated the synsets and the glosses to Spanish and alignment has been carried out at word level, whenever possible. The project has produced two interesting results: we have obtained a bilingual (Spanish and English) lexical resource for WordNet which will be available at no cost, as well as a parallel Spanish-English corpus annotated at word level with not only morphosyntactic information but also semantic information.

1 Introduction

In this paper[1] we present a lexical resource, the Spanish version of the English WordNet 3.0 (cf. Fellbaum 1998; Miller and Miller 1990). This resource is composed of the translations of the English synsets into Spanish and a parallel annotated corpus with the definitions and the examples of each synset.

This corpus will be specially of interest since it will not only be a parallel corpus but will also be partially annotated in both languages with morphosyntactic and semantic information at word level. There are other English-Spanish corpora. In some of them, alignment is established at paragraph level (cf. Gelbukh and AngelăVera-Félix 2006) whereas in others it is at word level (CRATER Corpus (cf. McEnery and Nieto-Serano 1997), GRIAL Trilingual Corpus (cf. Castellón 2005), ACREL Corpus (cf. Ramon 2004)). In all of these examples, as far as we know, annotation is limited to the morphosyntactic level if indeed there is any annotation at all.

We are certain that the information provided in this resource will be very useful for different automatic tasks within the domain of PLN, such as semantic annotation and disambiguation for Spanish or within the scope of applied linguistics to carry out contrastive studies in Spanish and English.

Corpus annotation is an arduous task and, in keeping with the precedent of other projects such as MultiSemCor (cf. Bentivogli and Ranieri 2005; Ranieri and Bentivogli 2004), we based our strategy on reusing the work already carried out for the annotation of the English corpus. Thus, we have worked with the annotated glosses

[1] The present work was funded by project HUM2007-65267.

provided by the University of Princeton.[2] The English glosses were already annotated at both morphological and semantic levels. From this annotated corpus, we translated the variants and the glosses into Spanish and changed the annotation when it was necessary because the morphosyntactic category did not correspond in both languages. Alignment has been carried out at word level, whenever possible, in order to keep the original annotation structure.

Below, we reproduce an example that shows the kind of information annotated in the glosses:

> deed: a notable achievement
>
> a notable [lemma = notable%1; pos = JJ ; SK= notable%3:00:00”] achievement [lemma = achievement%1; pos = NN; SK = achievement%1:04:00::)

As can be observed, the information annotated in the glosses includes the morphological tag (POS), the lemma and the WordNet sensekey of some of the words. From this structure, the resulting Spanish gloss is:

> hazaña: un logro notable
>
> un logro [lemma = logro%1; pos = NN; SK = logro%1:04:00::)] notable [lemma = notable%1; pos = JJ ; SK= notable%3:00:00”]

In the first stages of the project, we attempted to automate the translation process by using several different tools for this purpose. However, the quality of the translation was so poor and required so much manual editing that we discarded this possibility. Additionally, in order to keep the annotation structure and the alignment at word level we decided to translate as literally as possible whenever this kind of translation made sense.

A team of 3 translators who are native Spanish speakers with a high command of English have been working part time on the translations. There has also been a coordinator, who is bilingual, in charge of resolving issues that have been considered problematic by the translators. The coordinator also has had to validate the translations done and make sure the structure and the annotation has been kept. All of this work has been done using an online interface.

Next we present further detail of the process of creation of the resource. First, we will show the interface in order to present a clearer picture of how alignment is established. Afterwards, we will see several examples illustrating the diverse cases found in the process of translation and parallelization.

[2] The annotation of English glosses was carried out by Fellbaum’s team and funded by DTO/IARPA.

The current resource is composed of 20,000 variants and 10,000 glosses (around 100,000 words). In the future, we expect to continue with the creation of this resource and finish with the translation of the 30,000 annotated glosses available in English. Our intention is to make the work totally available on the Internet. It represents an added value for the scientific community, since the only bilingual (Spanish and English) lexical resource for WordNet, the Spanish EuroWordNet, is only partially free.

2 The translation interface

We have created an interface to work on the translation of the variants, and, specially, of the glosses, the appearance of which is presented in the figure below:

[redacted]

Figure 1. The translation and alignment interface.

The interface is divided into 4 sections. In the first section, identification information is displayed, such as the entry ID and the English synset variant (in the example, 'flying_colors%1:04:00::'). It also shows the field for the translation into Spanish of this variant ('traducción'), the name of the person who translates the gloss ('traductor') and the state of the translation ('validado'), which can be 'to be done', 'done', 'problematic' and 'validated'. First a translation is proposed ('done'); then it is revised and 'validated'. If translators are unsure how to translate a word or a part of the gloss, they label it as 'problematic', and if necessary a comment can be inserted.[3]

In sections 2 and 3, definition ('complete success') and examples ('they passed inspection with flying colors'), if any, are shown vertically. In both cases, the translation is carried out word by word in the column marked as Español in order to keep the annotations (morphosyntactic and semantic) of the English words whenever possible. Only the words in this field are parallel to the English version.

In the first column on the left ('ID'), the word (and punctuation mark) identifiers are shown. If a word is semantically annotated, this annotation is shown in the column beside (SK) by means of the WN sensekey assigned. In this example, all of the words of the English definition are labeled with this information (e.g. 'complete' carries the label 'complete%3:00:00' and 'success' the sensekey 'success%1:04:00').

[3] The field to insert comments is not visualized in the figure.

The morphological annotation can be seen under the POS label. In the example seen, ‘complete’ is labeled as an adjective (JJ) and ‘success’ as a singular noun (NN). Sometimes we need to change the category of a word when translating. To this aim we have created the box (Dif. Cat., different category). In this example this option has not been required but we will see an example later on.

One of the most common problems encountered when translating English to Spanish is that of word order. As can be seen in this case the order of the words that made up the definition is different in the two languages. The adjective always precedes the noun in English whereas in Spanish this is not always the case; in fact, it is usually the other way round. In order to account for order problems we have the column ‘Orden’. The value of this field is numeric and it expresses the order in which the translated words are to be shown in Spanish.[4]

The other three columns in this section correspond to ‘Info extra’ and ‘Multipalabra’ (extra information and multi-word respectively). As for the ‘multi-word’ field, it is used when it is not possible to make the correspondence between two concepts of the two languages word to word. If we take a look at the figure again, we will see that ‘flying colors’ has been established as a ‘multi-word’. In this case, the reason is that it is an expression in English, since its meaning is not compositional. Formally, this type of structure is created by linking the IDs of the words in English to just one field in the Spanish equivalent.

The field ‘extra information’ is used when more words are required in Spanish to express a meaning and will, therefore, not have a straight correspondence in the English annotated gloss. This could be true when, for example, we need a determiner in Spanish, as is the case of the image. As can be seen the articles ‘la’ and ‘un’ have been added to the Spanish structure and do not hold a link to any of the words in the English sentence.

Finally, in the fourth section of the interface we have four types of information: the first ID (further to the left) corresponds to the internal localization of the English WN database, the second ID is used to identify all the variants of a synsetăin the EWN; third, the two variants that are linked to the gloss and that are part of the same synset are visualized (in this gloss, in fact, there are two possible spellings of the same word), and, fourth, the translation to Spanish, that in this case is the same in both variants.

[4] If the number is 0, as in the example (section 3), it means that there are not any changes in the order of words in relation to English.

3 Different problems with parallelization

In this section we will briefly review some of the most common problems encountered when translating the text and aligning the Spanish and the English corpus. Obviously, we refer to mismatches in the translation that have a reflection on the alignment structure.

3.1 Adding functional words

One of the most common problems is the one just described: we need words in Spanish that do not have a direct counterpart in English. These elements would be left unlinked to any words in English but in a position between two elements of the sentence. Mostly, these are problems related to the different use of determiners in the two languages, as in the example above. As we can see in the examples below, this is also quite often the case with the use of the possessive adjectives (su) and with prepositions (de) that convey relations that in English are expressed by means of order.

(1) *Transporting alcoholic liquor for sale illegally.*
Transportando alcohólicas bebidas para venta ilegalmente.
'Transportar bebidas alcohólicas para su venta ilegal'.

(2) *Pocket-sized paperback book.*
Tamaño-de-bolsillo tapa-blanda libro
'Libro de tapa blanda de tamaño de bolsillo'.

3.2 Problems related to a different word order

As we have already pointed out this is an extremely common problem, and it usually affects the sequence 'adjective(s) plus noun', as in the example:

(3) *The experiencing of* emotional states.
La experimentación de emocionales estados.
'La experimentación de *estados emocionales*'.

At other times it is more complicated because what we essentially have are different structures in both languages.

3.3 Multi-word expressions

As we have said, the level at which alignment is established is the word since this is the level at which annotation in the English corpus is established. On occasion, this equivalence is not possible and links have to be established from an English expression to another expression in Spanish, as in the case of ‘flying colors’ and ‘éxito absoluto’, or to just one word, such as occurs in the following example:

(4) *There was* too much *for a single person to do.*
Había demasiado para una sola persona hacer
‘Había *demasiado* que hacer para una persona sola’.

Another common problem are clitic pronouns because they are graphically connected to the verb in Spanish when it is a gerund or an infinitive. In order to align them, we use the same mechanism we use with expressions, but a further morphological annotation process should analyze this form as a complex one formed by a verb and a pronoun.

It is also possible the contrary case, an English word is translated into more than one word in Spanish. This happens when English uses a synthetic process and Spanish an analytic one, for example, in the formation of some compounds (Eng. *trademark*, Spa. *marca registrada*) and also in some comparative and superlative forms:

(5) *The* biggest *overturn since David beat Goliath.*
El más-sorprendente resultado desde-que David ganó a-Goliat
‘El resultado *más sorprendente* desde que David ganó a Goliat’.

3.4 Different grammatical requirements

Given the fact that we are translating dictionary definitions and examples, the complexity of the grammatical structures to be translated is limited. Nevertheless there are some mismatches at this level that are worth remarking upon. A very common type is the use of gerund in English versus the use of infinitive in Spanish:[5]

Sometimes the differences between the two languages are even greater because there is not a complete correspondence at word level. In some instances, for example, a subordinate clause is required instead of an infinitive construction.

(6) *The Prohibition amendment made bootlegging profitable*
La Ley Seca hizo haciendo-contrabando rentable
‘La Ley Seca hizo *que el* contrabando *fuese* rentable’.

We keep the alignment of every word that can be linked even though the grammatical structure is different. If necessary, the changes of category are codified, as in the case of the gerund *bootlegging*, which is expressed by a singular noun (*contrabando*) in

[5] (cf. Izquierdo 2006) for a analysis of the possibilities of translation of the -ing forms to Spanish.

Spanish. We used the field ‘extra-information’ to accommodate any extra words that do not have a counterpart in English. The words *que*, *el* and *fuese* are left unlinked to any English words but linked to the Spanish words that form their context.

Let′s examine another example of grammatical mismatch between English and Spanish:

(7) *A miscalculation that recoils in its maker.*
Un error-de-cálculo que afecta a su realizador
‘Un error de cálculo que afecta *al que lo* realiza’.

In this case, in Spanish a verb (*realize*) is used instead of a noun (*maker*) and, as a consequence, the resultant syntactic structure is quite different in both languages, since what in English is expressed by the possessive that determines the noun, in Spanish it is expressed by means of the subject (*al que*) and the object (*lo*) of the verb used.

3.5 Non-existence of a lexical counterpart

Sometimes, the synset we are translating belongs to a cultural reality (most of the time American) that does not have a straight counterpart in Spanish, at least lexically speaking, and thus the literal translation of the definition is impossible.

(8) *He came all the way around on William’s* hit.
él llegó todo el camino alrededor en william’de golpe
‘Llegó a la meta gracias al *golpe* de William.’

This example belongs to the domain of baseball. Baseball is barely known in Spain and thus the rules of the game are unknown to most people; so in the translation we decided to rephrase it to make it more understandable. ‘Llegar a la meta’ is more general than ‘come all the way around’ but the concept is the same; to reach a point that is the goal. So we have explained the meaning as much as possible keeping the pointers to the English semantic annotation (‘hit, *golpe*’; SK: hit%1:04:03::).

Other examples of this type of mismatch are the well known verbs and deverbal nouns expressing manner in English. Manner in English can be expressed more generally than it is in Spanish where it usually requires a specification by means of an adjunct, as it can be seen in the case ‘smack, *beso sonoro*’.

4 Conclusions

We have presented the results of the lexical and textual resource we have built aligning the WN glosses in English-Spanish. The lexical resource contains the translation of the English variants. Sometimes the equivalence is not a one-to-one since a language can have more synonyms for a concept than the other; thus, synsets have not necessarily the same number of variants in both languages. As for the glosses, they are annotated with POS and semantic information. They parallel WordNet 3.0 entries by keeping the annotation from this source whenever possible. We have tried to make translations as literal as possible, since, even though the morphosyntactic annotation is easily done, the semantic annotation is an arduous task and it is worthwhile to take advantage of work already completed.

Both resources will be very useful for PLN researchers working with Spanish since currently there is not any completely public resource for this language linked to any version of WordNet and, on the other hand, there are very few corpus for Spanish with annotation at semantic level. Also, it presents the added value of it being aligned to the English corpus and therefore it can contribute information in both languages from a contrastive perspective.

References

Bentivogli, Emanuele Pianta, Luisa and Marcello Ranieri (2005). Multisemcor: an english-italian aligned corpus with a shared inventory of senses. In *Proceedings of the Meaning Workshop*, 90, Trento, Italy.

Castellón, Ana Fernández Gloria Vázquez, Irene (2005). Creación de un recurso textual para el aprendizaje del inglés. In *NOVATICA. Revista de la Asociación de técnicos de informática*, 51–54.

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gelbukh, GrigoriăSidorov, Alexander and José AngelăVera-Félix (2006). Paragraph-level alignment of an english-spanish parallel corpus of fiction texts using bilingual dictionaries. In *Text, Speech and Dialogue*, 61–67, Berlin: Springer.

Izquierdo, Marlen (2006). Contrastive analysis and translation english-spanish: functions of the english -ing form and its equivalents in spanish. In *Multilingua*, http://multilingua.uib.no/marlen.page.

McEnery, Andrew Wilson Fernando Sánchez-León, Tony and Amalio Nieto-Serano (1997). Multilingual resources for european languages: Contributions of the crater project. In *Literary and Linguistic Computing*, 219–226.

Miller, Richard Beckwith Christiane Fellbaum-David Gross, George and Katherine Miller (1990). Introduction to wordnet: An on-line lexical database. In *International Journal of Lexicography*, 235–244.

Ramon, Noelia (2004). Building an english-spanish parallel corpus for teaching and research: The actres project. In *Proceedings of the Sixth Teaching and Language Corpora Conference*.

Ranieri, Emanuele Pianta, Marcello and Luisa Bentivogli (2004). Browsing multilingual information with the multisemcor web interface. In *Proceedings of the LREC 2004 Workshop The Amazing Utility of Parallel and Comparable Corpora*, 38–41, Lisbon, Portugal.