Lexicalization and multiword expressions in the Basque WordNet

Eneko Agirre, Izaskun Aldezabal and Eli Pociello*
IXA NLP Group
University of the Basque Country
649 pk. 20.080 - Donostia. Basque Country.
eneko@si.ehu.es

Abstract

In this paper we propose a solution for the representation of a wide range of multiword expressions[1] (lexicalized or not) in the Basque WordNet. We first argue in favor of including non-lexicalized multiword expressions, and propose very simple criteria based on existing dictionaries to mark those that are lexicalized from those that are not. We then motivate and propose a representation based in EuroWordNet relations to represent the inner structure of them. This rich representation will allow for further populating the MEANING Multilingual Central Repository with additional semantic relations.

1 Introduction

Multiword expressions (MWEs) represent a challenge for NLP, both in syntactic and semantic grounds (Sag et al. 2002; Bentivogli & Pianta 2004; Villavicencio et al., 2005). Typically, MWEs cover a range of phenomena like idiomatic expressions, compound nominals, terminology, proper nouns, verb-particle constructions, light verbs, institutionalized phrases etc. The criteria for deciding whether to include a MWE in the lexicon or not depend on a number of factors, but mainly come from the intended use of the lexicon and the MWE. In lexicography and standard dictionaries, the MWE entries are taken to be lexicalized (Contreras & Sueñer, 2004; Cowie, 1990), in contrast to other MWEs, which are not included in the dictionary.

The context of the present paper is the construction of the Basque WordNet (Agirre et al. 2002; Agirre et al. 2006). At present, we don't consider proper nouns, and we will focus on the problem of deciding the criteria to include a MWE (lexicalized or not) in the wordnet, and how we can represent properly all kinds of MWEs (lexicalized or not). Our representation proposal involves several levels of detail ranging from "word with spaces", to full specification of the internal semantic structure of the MWE, including senses and semantic relations. Note that here we are concerned with the semantic level of the representation. The morphosyntactic representation and processing of Basque MWEs is dealt with in (Alegria et al. 2004).

The original WordNet (Fellbaum 1998), as a computational lexicon, only includes lexicalized entries and concepts. There are a few exceptions, usually linked to general concepts that are introduced to better organize the hierarchy, as for instance the concept 'fictional character' which is not lexicalized. The task of deciding which MWEs are lexicalized or not is one of the main tasks of a wordnet builder, but unfortunately, the boundaries for lexicalization are very difficult to draw (Contreras & Sueñer, 2004; Cowie, 1990).

Besides, when building wordnets, the need arises for including non-lexicalized or close-to-lexicalized entries, especially, for treating lexical gaps (concepts that lexicalize in one language, but not in another, such as *to cook* that in Basque needs to be expressed by a non-lexicalized MWE: *janaria prestatu*, lit. 'prepare food'). Those 'less-lexicalized' entries are very useful for translation as well as for word sense disambiguation (Bentivogli & Pianta 2004). The wordnet builders therefore

* Authors listed in alphabetical order.

[1] Note that we use *multiword expression* as a general term to denominate those constructions, either lexicalized or not, containing more than one word (*word* defined as "any string of characters between two blanks" (Fontenelle et al., 94).

need to decide what to do (only include lexicalized entries or also include boundary or non-lexicalized entries) and how to represent all kinds of MWEs (ranging from word-as-spaces, which might be enough for obscure idioms, to the representation of the internal structure).

In this paper we start exploring the problems posed by lexicalization in relation to the Basque WordNet (Section 2). We then mention the motivation for including non-lexicalized MWEs in the Basque WordNet (Section 3), followed by our criteria for lexicalization (Section 4) and for the representation of MWEs in the Basque WordNet (Section 5). Finally Section 6 presents the conclusions and further work.

2 Lexicalization problems

The term *lexicalization* refers to the transformation of an element (or a sequence of elements) into a unique lexical or conceptual element (Lewandowski, 1992)[2]. Therefore, the result of lexicalization can be carried out as (i) a lexical element (a word) or (ii) a sequence of elements or phrase (a MWE). Since the lexicalization problem is much more complex with MWEs than with words, in this paper we will focus on MWEs.

The aforementioned "transformation" is an obscure process. Many authors point out (Calzolari et al., 2002) that lexicalization should be understood as a continuum from full-fledged compositional and productive constructions to fixed and frozen expressions. This is due to the fact that lexicalization is the result of the combination of a number of factors, which can occur either totally or partially. Although, there is no agreement in the number of factors that make lexicalization, we can mention the most important ones: *co-occurrence frequency* or *collocation* > *fixation* > *semantic specialization* > *idiomatization.* In those cases that the combination of factors occurs totally –in other words, when the construction goes through all those factors– then, we will have a frozen expression. On the other hand, when the combination of factors is partial –when the construction does not go through all those factors– the construction may be at any point in that continuum.

Therefore, depending on the point of the continuum constructions are, they have different characteristics, and consequently, they will be named with different terms, which has brought authors to create a classification and terminology to distinguish among them. Unfortunately, there is no uniformity either in the classification or in the terminology related to MWEs.

According to Sag et al. (2002) there are two main kinds of MWE: **lexicalized phrases** and **institutionalized phrases**. They describe lexicalized phrases as "having at least partially idiosyncratic syntax or semantics, or containing 'words' which do not occur in isolation*"*. They can be further broken down into idioms proper (*kick the bucket*), decomposable idioms (*spill the beans*), compound nominals including terminological MWEs (*car park*, *central processing unit*), proper names (*Los Angeles*), verb-particle constructions (*set up*) and light verb constructions (*make a mistake*). For this study we leave out proper names and verb-particle constructions, which don't occur in Basque.

Idioms are relatively frozen expressions whose meaning cannot be built compositionally from the meanings of their component words. Moreover, the component words cannot be substituted with synonyms (as in *adarra jo* in Basque, literally meaning 'to play the horn' and translated to English as another idiom such as *to pull sb's leg*). Decomposable idioms are sequences of words which habitually co-occur and whose meaning can be derived compositionally. However, they show a kind of semantic cohesion which limits the substitution of their component words –as in *to spill the beans* in English, where *spill* and *beans* can be taken to have the appropriate senses that produces the compositional reading, or in Basque *burua jan* ('to brainwash', lit. 'to eat the head'). A similar phenomenon occurs with light verbs –*fall asleep* in English, or *lan egin* in Basque ('to work', lit. 'to do work')– and compound nominals –*buruhauste* in Basque ('problem', lit. 'broken head').

Apart from these kinds of lexicalized phrases, the other kind of MWEs is institutionalized phrases. These are not usually taken as elementary lexical units, that is, they are not taken as lexicalized forms, and do not belong to the lexicon**.** Institutionalized phrases

[2] Other approaches to lexicalization are Talmy (1985) and Traugott (1996), which are not explained here due to space limitations.

are combinations following only the general rule of syntax: the word meanings combine compositionally but can not always be substituted by synonyms. They are often conventionalized, and they take only one of the possible readings available (for instance *traffic light* means *stop light*, and not *turning light* which would be also a possible meaning). Moreover, they are characterized by having much higher frequency than alternative verbalizations (*traffic director* or *intersection regulator* to mean *traffic light*). Thus, institutionalized phrases are semantically and syntactically compositional, but statistically idiosyncratic.

Alternatively, other authors (Bentivogli & Pianta, 2004) distinguish between **lexicalized MWEs** such as idioms and restricted collocations (which would include all the above except institutionalized phrases), and **free combinations** (which would include institutionalized phrases). Both idioms and restricted collocations are considered to be a sequence of elements that act as a single unit at some level of linguistic analysis and that are lexicalized (i.e. they belong to the lexicon). However, idioms are frozen expressions whose meaning cannot be built compositionally (as the examples mentioned before, *adarra jo - to pull sb's leg*), whereas restricted collocations consist of words which habitually co-occur and whose meaning can be derived compositionally but with some degree of semantic cohesion (as the aforementioned *to spill the beans* in English, or *burua jan* in Basque). On the contrary, free combinations follow the general rule of syntax, are compositional and allow for synonym substitution. For instance, the English verb *to bike* is translated into Basque as *bizikletan ibili* (lit. 'to walk on a bicycle'). However, we can use a synonym to express exactly the same: *bizikletan joan* (lit. 'to go on a bicycle'). This is the reason why they are considered as non-lexicalized forms, and therefore, they do not belong to the lexicon**.**

Alegria et al. (2004) use the term **multiword expressions** to refer to any word combinations ranging from idioms, over proper names, compounds, lexical and grammatical collocations, lexicalized phrases etc. to institutionalized phrases: *lan egin* ('to work', lit. 'to do work'), *noizik behin* ('once in a while'), *ahopeka abestu, ahopeka kantatu, ahopean abestu, ahopean kantatu* ('to hum', lit. 'to sing in whispers'). On the other side, they use the term **multiword lexical units** to refer to lexicalized MWEs (those MWEs that are semantically non-compositional or syntactically idiosyncratic): *adarra jo* ('to pull sb's leg', lit. 'to play the horn').

3 The need for non-lexicalized multiword expressions

In order to provide the basis for the semantic interpretation of Basque, it is obvious that the Basque WordNet needs to provide the meaning for lexicalized MWEs. There are four reasons or situations why we need to also include non-lexicalized MWEs: difficulty of defining lexicalization, lexical gaps, translation tasks, facilitate semantic interpretation and a richer Lexical Knowledge Base.

The first reason is that we do not want to have lengthy debates about the lexicalization status of a MWE. In case of doubt, we want to incorporate as many MWEs as possible, without making claims of their lexicalization status, and thus, allow for non-lexicalized MWEs.

In the process of building the Basque WordNet, we have followed the expand approach, which means that we based our work on the English WordNet synsets, and substituted the English variants by Basque variants (Vossen et al. 1998). Additionally, we also incorporate new synsets that exist for Basque but not for English. In many cases, the English synsets have a dubious lexicalization in Basque, that is, they can be translated by a MWE which is not found in a Basque dictionary. If we were to follow a rigid approach for including only lexicalized variants, those synsets would be gaps in the Basque WordNet, for instance, *bizikletan ibili* ('to bike', lit. 'to walk on bicycle') or, *ahopeka kantatu* ('to hum', lit. 'to sing in whispers'). We nevertheless want to include such translations, as they are very useful information for translation tasks.

Regarding semantic interpretation in general, and word sense disambiguation in particular, the more MWEs are included in WordNet, the easier is the task for a word sense disambiguation program. For non-compositional MWEs this is obvious, but considers also the decrease of ambiguity for institutionalized phrases or free phrases.

Linked to this, a rich Lexical Knowledge Base, where the internal semantic structure of MWEs is represented, would aid in the semantic interpretation process. For instance, *fall_asleep* is a variant for a synset in WordNet 2.0, and capturing the relation between *asleep* and *fall_asleep* (very similar to *lo_hartu* and *lo* in Basque) would allow to better understand the consequences of falling asleep.

4 Introducing multiword expressions in the Basque WordNet

As previously seen, MWEs are usually analyzed from different perspectives and criteria. In general terms, MWEs can be defined by some or all of the following criteria (Calzolari et al., 2002):

1. reduced syntactic and semantic transparency;
2. reduced or lack of compositionality;
3. more or less frozen or fixed status;
4. possible violation of some otherwise general syntactic patterns or rules;
5. a high degree of lexicalization (depending on pragmatic factors).
6. a high degree of conventionality.

When facing concrete examples these criteria are not easy to apply. Even for lexicographers, sometimes it is very difficult to distinguish among those constructions, especially, between those that are not frozen. This is why some constructions do have a dictionary entry in some dictionaries, but not in others. For instance, we have looked up *buruz ikasi* ('to memorize', lit. 'to learn by heart') in three Basque monolingual dictionaries[3]; in two of them *buruz ikasi* is a dictionary entry, so it has been considered as a lexicalized construction. Still, when looking up to a similar construction (*buruz esan* – 'to recite', lit. 'to say by heart'), it does not appear in any of the dictionaries. It seems to have been treated as a non-lexicalized construction, although, perhaps, it has been overlooked.

Consequently, we needed to define some criteria which can be easily applied when classifying MWEs in the Basque WordNet. Regarding the criteria above, and being the goal of the Basque WordNet to provide the building blocks for the semantic interpretation for Basque, our main concern with MWEs is that of reduced or lack of compositionality. The syntactic behavior of MWEs (reduced syntactic transparency, frozen or fixed status, violation of syntactic patterns) are covered in a separate work in our group (Alegria et al. 2004). In the following section, we will describe the criteria adopted to classify MWEs as lexicalized or as non-lexicalized.

Before introducing our criteria, we will see how the actual WordNet deals with MWEs. WordNet does include lexicalized synsets which may contain either single words or MWEs, or sometimes, both together:

English WN {*girlfriend, girl, lady_friend*}

The Basque WordNet builders must take this into account, and they need to decide whether a synset in the English WordNet –expressed as a single word or as a MWE– can be translated into Basque, using a single word or a MWE, or using both.

English WN {*girlfriend, girl, lady_friend*}
Basque WN {*neska-lagun, adiskide, lagun, neska*}

Once the builder of the Basque WordNet has detected which synsets are MWEs in Basque, his next objective will be to decide whether these MWEs are lexicalized in Basque or not, in order to add them in the Basque WordNet.

A simple way to do this is to distinguish, for each synset, between fully lexicalized variants, non-lexicalized MWEs and lexical gaps. In the first case, we would have a normal synset with its variants (simple and/or MWE). In the second case we can include a broader range of MWEs, marked with a special flag. In the last case, the synset would not have a counterpart in Basque.

The Basque WordNet team decided to define a simple criterium for distinguishing fully lexicalized variants from the rest: the MWE needs to have an entry in a monolingual dictionary (Elhuyar, 2000; Sarasola, 1996 and Euskaltzaindia, 2000) or terminological glossary (UZEI, 1987). If a MWE is a dictionary entry, then, the builder of the Basque WordNet will add this MWE in the Basque WordNet. For instance, *to memorize* is translated into Basque as *buruz ikasi* (lit. 'to learn by heart'). Being *buruz ikasi* a dictionary

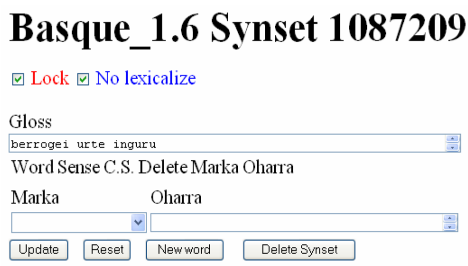
[3] *Euskal Hiztegi Modernoa* (Elhuyar, 2000), a terminological data bank for Basque (*Euskalterm*) and *Euskal Hiztegia* (Sarasola, I., 1996).

entry, the builder of the Basque WordNet will add this multiword in the synset:

English WN {*memorize, memorise, con, learn*}
Basque WN {*memorizatu, buruz_ikasi*}

The reason for such a simple criterium is mainly the lack of human resources to do an appropriate in-depth study of lexicalization for all potential Basque WordNet entries.

But, what happens when a synset is expressed by a MWE that is not a dictionary entry? It depends on the type of the MWE. Sometimes, the only way to express a synset in Basque is using a definition:



Basque_1.6 Synset 10872096

☑ Lock ☑ No lexicalize

Gloss

berrogei urte inguru

English WN { forties, mid-forties -- (the time of life between 40 and 50) }
Basque WN {GAP -- (berrogei urte inguru)}

Figure 1: Representation of a GAP in the actual interface for the Basque WordNet (Atserias et al., 2004)

In Basque, the only way to translate *forties* (see Fig. 1) is using a kind of definition (e.g. a long and complex syntactic construction): *berrogei urte inguru* (lit. ‘around forty years old’). As we can see in Fig. 1, those constructions are considered as non-lexicalized (or lexical gaps) and we do not introduce them in the synset, but in the gloss.

On the contrary, there are some recurrent MWEs that are used to express a lexicalized concept in English (*to recite*) but not in Basque (*buruz esan*). Although, with regard to construction, *buruz esan* is very similar to *buruz ikasi* (‘to memorize’ or ‘to learn by heart’), according to our criteria, *buruz esan* will not be considered as lexicalized because it is not a dictionary entry. Yet, its synonym *errezitatu*, which is a borrowed term for the translation of *to recite*, is considered to be lexicalized because it has a dictionary entry in Basque dictionaries. As a consequence, the editor of the Basque WordNet would introduce *errezitatu* in the Basque WordNet, but not *buruz esan*. Nevertheless, *buruz esan* is the most frequent and natural way to translate the English verb *to recite* into Basque. So, this approach seems to be quite risky, since applying these criteria leads to the consequence that a considerable number of frequently used expressions can be excluded from the Basque WordNet as they are considered to be not lexicalized.

To avoid this risk, we have decided to consider this type of MWEs as *syntagmatic concepts* (Artola, 1993), and to include them in the Basque WordNet. These are those concepts that need to be expressed by a phrase and that have become widespread. This approach has already been used by Bentivogli & Pianta (2004). These authors introduce those frequent MWEs as phrasets and they also add them in the Italian WordNet. Some examples of syntagmatic concepts follow:

English WN {*recite, recite*}
Basque WN {*buruz_esan, errezitatu*}

English WN {*retranslate*}
Basque WN {*berriro_itzuli*} (lit. ‘translate again’)

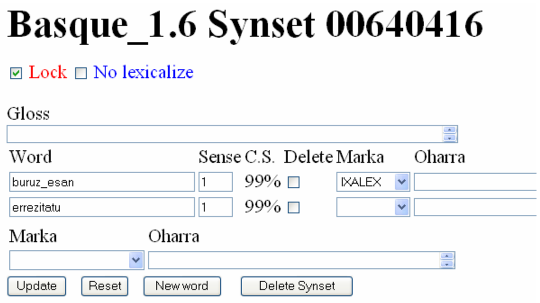
English WN {*hum*}
Basque WN {*ahopeka_kantatu*} (lit. ‘sing in whispers’)

English WN {*bike*}
Basque WN {*bizikletan_ibili*} (lit. ‘move on a bike’)

English WN {*frank, frankfurter, hotdog, hot dog*}
Basque WN {*Frankfurt_saltxitxa*}

English WN {*two-dimensional_figure* }
Basque WN {*irudi_bidimentsional*}

As Bentivogli & Pianta (2004) support, the main reasons to add syntagmatic concepts in wordnets is to manage lexical gaps, that is, cases in which a language expresses a concept with a lexical unit whereas the other language does not. Therefore, instead of representing a lexical gap by adding an empty synset aligned with a non-empty synset of the other language (see Fig. 1), we propose to represent it following Bentivogli & Pianta's (2004) approach: adding the syntagmatic concept in the synset. However, in order to differ these MWEs from lexicalized MWEs, all syntagmatic concepts are marked with the syntagmatic concept label in the database (see Fig. 2).



Basque_1.6 Synset 00640416

☑ Lock □ No lexicalize

Gloss

Word	Sense	C.S.	Delete	Marka	Oharra
buruz_esan	1	99%	□	IXALEX	
errezitatu	1	99%	□		

Marka

Oharra

Figure 2: The actual interface, showing a syntagmatic concept (IXALEX is our shorthand for syntagmatic concept).

5 Full representation of phrasal concepts in the Basque WordNet

The above representation (Section 4) is limited to listing the MWEs together with their lexicalization status, and fails to reflect the inner structure and semantic relations in the MWE. This more detailed representation is especially desirable for decomposable idioms, compound nouns (incl. terminology), light verbs and institutionalized expressions, where we would like to keep semantic links between components. It is also necessary for a proper coupling of the syntactic analysis of the MWE and its semantic interpretation (Sag et al. 2002). For instance, in Basque the auxiliary verb agrees with both the ergative case (the subject) and the absolutive case (the object). In the case of some light verb constructions like *lo egin* (which is considered a lexicalized MWE; ‘to sleep’, lit. ‘to do sleep’) its nominal component *lo* (nominal ‘sleep’) is sintactically the object of the sentence in ‘*umeak lo egin zuen*’ (‘the baby slept’), and the semantic interpreter needs to make sense of the role of this object, which is really part of a MWE. From another perspective, as mentioned in Section 3, the internal relations would allow the semantic interpreter to infer that in the previous sentence a *sleep* state is involved.

A proposal for the representation of the inner structure was made by (Bentivogli & Pianta, 2004). They propose the use of a *composed-of* link between the MWE variant and its components, including their word sense specification whenever possible (see Fig. 3c). This proposal does not make explicit the semantic relation between the MWE and its components. EuroWordNet defined a richer set of semantic relations than the original WordNet, including the involved relation, defined as follows:

> “The INVOLVED relation is used to encode data on arguments or adjuncts lexicalized within the meaning of a 2nd order entity”
>
> (Alonge et al 1998, p. 29).

We think that these relations are very well suited for encoding the inner relations. Scheme d) in Fig. 3 shows a possible representation for *lo_egin* where *lo* is the *involved_theme*[4] of the MWE verb. An additional advantage of this representation is that those semantic relations carry over to other languages, and apply also in English (the *sleep* is the *involved_theme* for a sleeping event). In addition to these possibilities, Fig. 3 also shows the other two possibilities for completeness: a) for non including MWEs, and b) for including them as words with spaces and no inner structure. At the current development stage, all MWEs have been marked following the b) scheme.

The same scheme as in Fig. 3d can be applied to complex MWEs like *gabon kantak abestu* (‘to carol, lit. ‘to sing Chirstmas songs’) or *arinki lo egin* (‘to snooze’, lit. ‘to sleep lightly’). In fact, we will show that it can be applied to all kinds of MWEs.

6 Summary and further work

In this paper we have proposed a solution for the representation of the wide range of MWEs (lexicalized or not) in the Basque WordNet. We first argue in favor of including non-lexicalized MWEs, and propose a very simple criterion based on existing dictionaries to mark those that are lexicalized from those that are not. We then propose a representation based in EuroWordNet relations to represent the inner structure of them. Currently, **noun and verb** MWEs in the Basque WordNet have been marked according to their lexicalization status, i.e. either non-lexicalized or syntagmatic concepts. This corresponds to scheme b) in Fig. 3. Table 1 shows the current figures for the Basque WordNet (Agirre et al., 2006) and it also reviews the amount of synsets marked as non-lexicalized or as syntagmatic concepts.

[4] *involved_theme* is a specialization of the *involved* relation, where the semantic role is *theme*. We also allow for 2nd order entities as fillers for these relation. Note that in English, sleep$_V$ and sleep$_N$ are also related by a *xpos_near_synonym* relation.

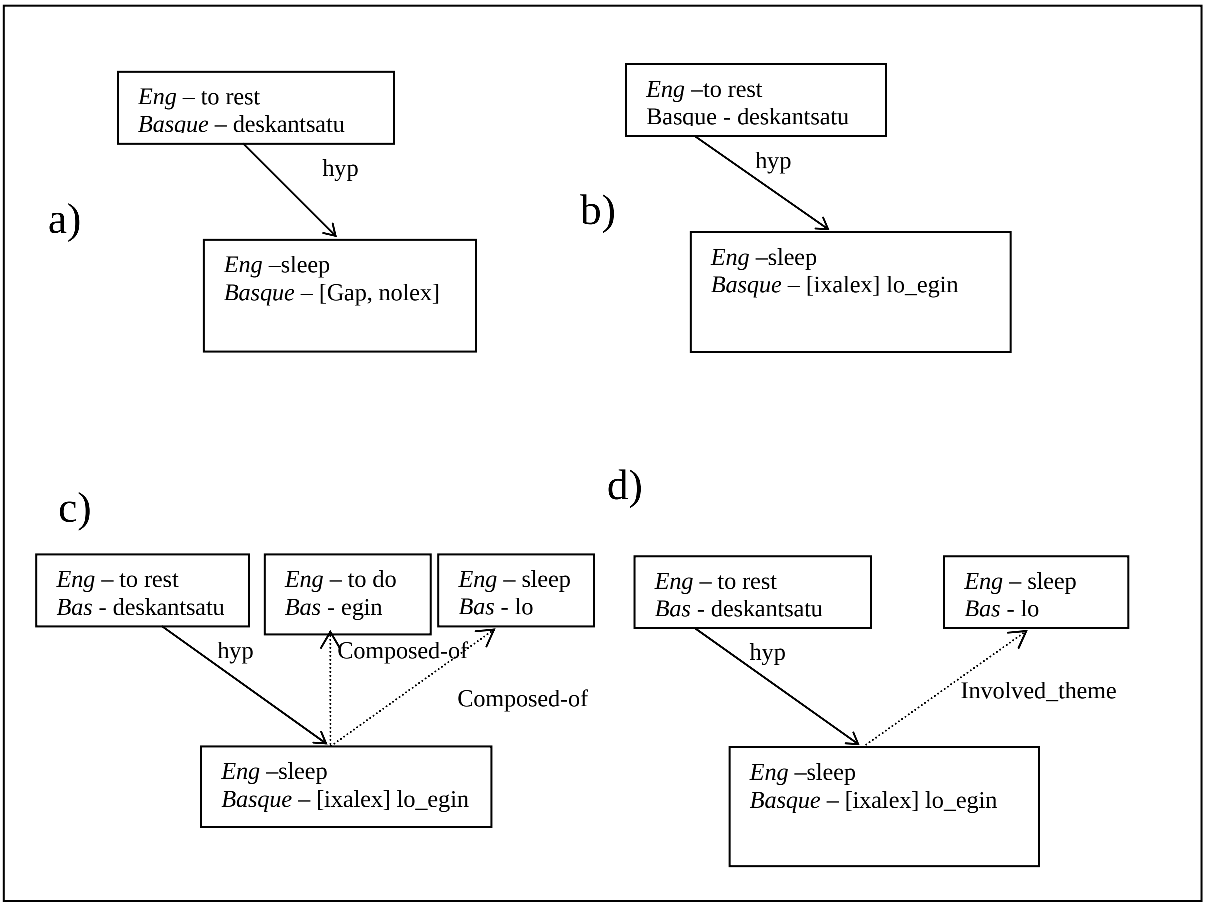


Figure 3: Different representation choices in the Basque WordNet: **a)** representing only lexicalized multiwords, **b)** including syntagmatic concepts (ixalex being the internal tag for those), **c)** describing the inner structure using phrasets, and **d)** describing the inner structure using EuroWordNet relations.

	TOT	N	V	ADJ	ADV
Word Senses	51423	41833	9450	140	0
Lemmas	25755	22492	3368	50	0
Synsets	31585	27880	3592	113	0
Basque gaps (no lex)	1439	1223	208	8	0
Proper Nouns		680			
MWE (no lex)	5730	2935	2439	0	0
Syntagmatic concepts	356	79	273	4	0

Table 1: Current figures for the Basque WordNet and for non-lexicalized and syntagmatic concepts.

In the future, we are planning to further enrich the MWE with the representation of their inner structure, following the proposal in Section 5 (corresponding to scheme d) in Fig. 3). We plan to apply semi-automatic methods to disambiguate both the semantic relation and the synsets involved in the inner structure, using a method which has been already applied to derivation relations (Agirre & Lersundi, 2001). These relations will help populate the relations in all wordnets designed in the EuroWordNet style (linked to a common interlingual index) and further enrich the MEANING Multilingual Central Repository (Atserias et al., 2004).

We would also like to join the morphosyntactic and semantic representation of MWEs. This is a subtask in the process of merging the morphosyntactic lexicon for Basque (EDBL, Alegria et al. 2004) and the semantic lexicon (Basque WordNet).

Acknowledgments

The work has been partially funded by the European Commission (MEANING project IST-2001-34460), by the Basque Government (Saiotek, GO765) and by the Education Department of the Spanish Government (HUM2004-21127-E). Eli Pociello has a PhD grant from the Basque Government.

References

Agirre E. and Lersundi M. (2001) Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. In P*roceedings of SEPLN 2001*. Jaén (Spain).

Agirre E., Ansa O., Arregi X., Arriola J., Díaz de Ilarraza A., Pociello E., Uria L. (2002) Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis. In *Proceedings of First International WordNet Conference.* Mysore (India).

Agirre E., Aldezabal I., Etxeberria J., Izagirre I., Mendizabal K., Pociello E., Quintian M. (2006) Improving the Basque WordNet by corpus annotation. In *Proceedings of Third International WordNet Conference*. Jeju Island (Korea).

Alegria I., Ansa O., Artola X., Ezeiza N., Gojenola K., Urizar R. (2004) Representation and Treatment of Multiword Expressions in Basque. In *Proceedings of the ACL workshop on Multiword Expressions*. Barcelona (Catalunya).

Alonge A., Calzolari N., Vossen P., Bloksma L., Castellon I., Marti T., Peters W. (1998) The Linguistic Design of the EuroWordNet Database. In: N. Ide, D. Greenstein and P. Vossen (eds.), Special Issue on EuroWordNet. *Computers and the Humanities*, Volume 32, Nos. 2-3, (91-115).

Atserias J., Villarejo L., Rigau G., Agirre E., Carroll J., Magnini B., Vossen P. (2004) The MEANING Multilingual Central Repository. In *Proc. of the 2nd Global WordNet Conference.* Brno (Czech Republic).

Artola, X. (1993) *HIZTSUA: Hiztegi-sistema urgazle adimendunaren sorkuntza eta eraikuntza. Hiztegi-ezagumenduaren errepresentazioa eta arrazonamenduaren ezarpena.* PhD Thesis. University of the Basque Country.

Bentivogli L. and Pianta E. (2004) Extending wordnet with syntagmatic information. In *Proceedings of Second Global WordNet Conference* (47–53).

Calzolari N., Fillmore C., Grishman R., Ide N., Lenci A., MacLeod C., Zampolli A. (2002) Towards Best Practice for Multiword Expressions in Computational Lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, (1934-1940).

Contreras JM. and Sueñer A. (2004) Los procesos de lexicalización. In E. Perez Gaztelu, I. Zabala and L. Gràcia (eds.), *Las fronteras de la composición en lenguas románicas y en vasco* (47-109). Universidad de Deusto.

Cowie, A.P., Mackin R., McCaig I.R. (1990) *Oxford Dictionary of Current Idiomatic English.* v2.

Fellbaum C. (1998) *WordNet: An electronic Lexical Database*. The MIT Press, Cambridge, Massachusetts. London (England).

Fontenelle, T., Adriaens, G., De Braekeleer, G. (1994) The Lexical Unit in the Metal® MT System. In *MT*, Volume 9. 1-19. The Netherlands.

Lewandowski, T. (1992) *Diccionario de Lingüística*. Cátedra.

Talmy, L. (1985) Lexicalization patterns: semantic structure in lexical forms. In T. Sophen (ed.), *Language Tipology and Syntactic Description*. Cambridge University Press.

Traugott, E.C. (1996) Lexicalization and Lexicalization. In K. Brown and J. Miller (eds.), *Concise Encyclopedia os Syntactic Theories*: (181-187). Cambridge University Press.

Sag I. A., Baldwin T., Bond, F., Copestake A., and Flickinger, D. (2002) Multiword expressions: A pain in the neck for NLP. In *Proceedings of the Third International Conference on Intelligent Text Processing and ComputationalLinguistics (CICLING 2002)*, pages 1–15. Mexico City (Mexico).

Villavicencio A., Bond F., Korhonen A., McCarthy D. (2005) Introduction to the special issue on multiword expressions: Having a crack at a hard nut. In *Computer Speech & Language*, Volume 19, 4.

Vossen, P. ; L. Bloksma; S. Climent; M. Anonia Marti; G. Oreggioni; G. Escudero; G. Rigau; H. Rodriguez; A. Roventini; F. Bertagna; A. Alonge; C. Peters; W. Peters. 1998. The Reestructured Core wordnets in EuroWordnet: Subset1. EuroWordNet(LE-4003) Deliverable D014/D015, University of Amsterdam.

Dictionaries

Elhuyar, 2000. *Euskal Hiztegi Modernoa*.

Euskaltzaindia, 2000. *Hiztegi Batua*.

Sarasola, I. 1996. *Euskal Hiztegia*.

UZEI, 1987. *Euskalterm.* http://www.uzei.com/en/euskalter.htm.