Lexicalization and multiword expressions in the Basque WordNet

Eneko Agirre, Izaskun Aldezabal and Eli Pociello*

IXA NLP Group

University of the Basque Country (UPV/EHU)

Abstract

In this paper we propose a solution for the representation of a wide range of multiword expressions (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 on EuroWordNet relations to represent the inner structure of them. This rich representation will allow for populating the MEANING Multilingual Central Repository with additional semantic relations.

0. Introduction

The context of the present paper is the construction of the Basque WordNet (Agirre et al. 2002). The Basque WordNet is a Basque lexical-semantic knowledge base (LSKB) developed by the IXA research group.[1] This Basque LSKB is based on WordNet (Fellbaum 1998), as well as on its multilingual counterparts EuroWordNet (Vossen 1998) and the Multilingual Central Repository (Atserias et al. 2004).

The goal of building the Basque WordNet raised several problems and challenges. But in this paper we will deal with the question of lexicalization and multiword expressions (MWEs). MWEs represent a challenge for Natural Language Processing, 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, and etcetera. The criteria for deciding whether to include a MWE in the lexicon or not depend on a number of

* Authors listed in alphabetic order.

[1] http://ixa.si.ehu.es

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), that is, those MWEs that are not considered as lexicalized are not included in the dictionary.

At present, we have focused on the problem of deciding the criteria to include a MWE in the wordnet, and how we can represent properly all kinds of MWEs (lexicalized or not).[2] 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. This approach is based on Bentivogli & Pianta's (2004) representation proposal which has been applied in the Italian WordNet.

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).

We start this paper briefly presenting the lexical-semantic knowledge bases used by the IXA research group to develop a Basque LSKB: WordNet and its multilingual counterparts EuroWordNet and the Multilingual Central Repository. Section 2 describes how those LSKBs deal with lexicalization and MWEs. Afterwards, we explore the theoretic problems posed by lexicalization (Section 3). We then mention the motivation to include non-lexicalized MWEs in the Basque WordNet (Section 4), followed by our proposal for the inclusion of MWEs in the Basque WordNet (Section 5) as well as their representation (Section 6). Finally, Section 7 presents some conclusions and further work.

1. WordNet, EuroWordNet and the Multilingual Central Repository

Natural Language Processing techniques need lexical-semantic knowledge bases (LSKB) in order to perform semantic interpretation. The IXA group decided to develop a Basque LSKB for this reason: the Basque WordNet. Basque WordNet is based on WordNet and its multilingual counterparts EuroWordNet and the Multilingual Central Repository (MCR). The steps followed in the construction of the Basque LSKB are explained in Pociello (2004), Agirre et al. (2005)

[2] However, for this study, we have left aside proper names.

and Agirre et al. (submitted). Below, we briefly present the most relevant characteristics of these LSKBs.

WordNet (Fellbaum 1998) is a semantic lexicon for the English language. It groups English words into sets of synonyms called *synsets*, and records various semantic relations between these synonym sets forming a hierarchy. Moreover, each of these *synsets* corresponds to a lexical concept. For instance, and as we can see in example (1), according to WordNet the noun *T-shirt* has one *synset*, or in other words, it can refer to one lexical concept:

(1) T-shirt, jersey -- (a close-fitting pullover shirt)

The *synset* in (1) is composed by two words: *T-shirt* and *jersey*. Therefore, *T-shirt* and *jersey* are synonyms.

Furthermore, these synonym sets are related by various semantic relations, such as hyperonymy and hyponymy.[3] In (2) and (3) we have the hyperonyms and hyponyms respectively of (1):

(2) => jersey, T-shirt -- (a close-fitting pullover shirt)
=> shirt -- (a garment worn on the upper part of the body)
=> clothing, clothes -- (covering designed to be worn on a person's body)
=> covering -- (an artifact that protects or shelters or conceals)
=> artifact, artefact -- (a man-made object)
=> object, physical object -- (a physical entity)
=> entity, something -- (anything having existence)

(3) => jersey, T-shirt -- (a close-fitting pullover shirt)
=> turtleneck, polo-neck -- (a sweater or jersey with a high close-fitting collar)

Hyperonyms are general or superordinate terms. In (2) we can see that a *T-shirt* is a *shirt*; that a *shirt* is a *clothing*; that a *clothing* is an *covering*; that a *covering* is an *artifact*, and so on. Therefore, when we talk about a *T-shirt* we are talking about something that is also a *shirt*, a *clothing*, a *covering*, an *artifact*, etc.

On the reverse, hyponyms are words that refer to more specific words or concepts. As we can see in (3), the only hyponym of *T-shirt* is *turtleneck*, a T-shirt or jersey with a high close-fitting collar.

[3] Although synonymy, hyperonymy and hyponymy are the most important semantic relations in WordNet, there are much more semantic relations: meronimy, antonymy, holonimy, etc. For more details refer to Fellbaum (1998).

The purpose of WordNet is twofold: to produce a combination of dictionary and thesaurus that is more intuitively usable, and to support automatic text analysis and artificial intelligence applications. This database can be browsed on line at http://wordnet.princeton.edu/perl/webwn.

Considering this English WordNet as a reference, new wordnets have been built in some other languages (Spanish, Italian, German French, Danish, etc.), especially in the framework of the EuroWordNet project.[4] EuroWordNet (Vossen 1998) is a coordinated system of semantic networks for European languages, where each language develops its own wordnet. What EuroWordNet does is, basically, to add multilingual equivalence links across wordnets.

The Multilingual Central Repository[5] (MCR) (Atserias et al. 2004) is a LSKB developed in the MEANING project ("MEANING: Developing Multilingual Web-Scale Language Technologies"). Currently the MCR integrates into the EuroWordNet framework five local wordnets (English, Catalan, Italian, Basque and Spanish) and other semantic information that EuroWordNet did not provide (hundred of thousand of new semantic relations and properties automatically acquired from corpora).

At present, the Basque WordNet is being built basing on MCR framework. Fig. 1 presents the Basque WordNet interface,[6] which can be consulted on line at http://ixa2.si.ehu.es/mcr/wei.html.

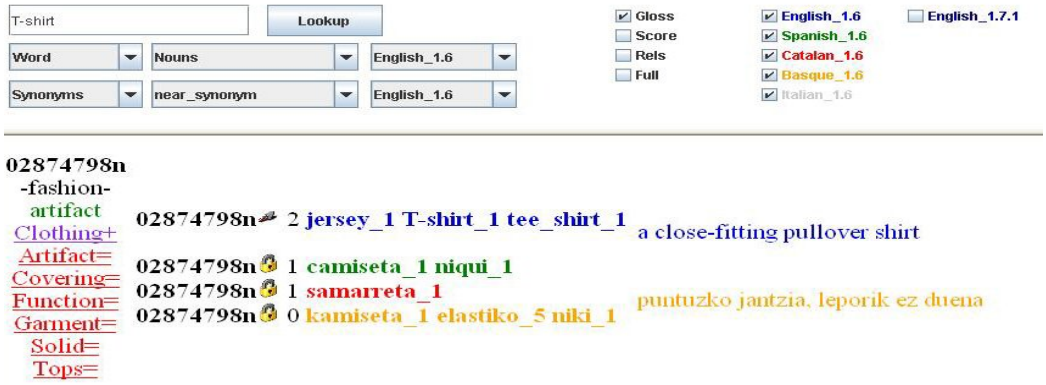


Figure 1: Actual interface for the Basque WordNet.

4 http://www.hum.uva.nl/~ewn

5 http://nipadio.lsi.upc.edu/cgi-bin/wei4/public/wei.consult.perl

6 Further information in order to use the interface is available in Agirre et al. (2005) and Agirre et al. (submitted).

2. Lexicalization and multiword expressions in WordNet frameworks

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' or 'body of water' which are 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).

The difficulty to differ between what is a lexicalized MWE and what is not, is emphasized by the development of manually sense-annotated corpora based on the Basque WordNet *synsets* (called the Basque Semcor), together with the Basque WordNet itself (Agirre et al. 2006). All occurrences in the corpus that are part of a MWE are marked in order to signal that the word is part of a lexicalized MWE. For instance, if an occurrence of the noun *urte* ('year') is followed by the word *berri* ('new'), it will be marked as part of a MWE: *urte berri* ('new year'). The aim of detecting MWEs in the corpora was to update the Basque WordNet with the most current MWEs in Basque. However, this marking has been problematic, because, quite often, taggers did not agree on what a lexicalized MWE is.

Therefore, when building wordnets, the need for including non-lexicalized or close-to-lexicalized entries arises, especially, when 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, 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). As a consequence, and in order to make wordnet builders' job easier, we have defined some criteria to represent all kinds of MWEs (Section 5). Still, before going deeply into the explanation of our criteria, in Section 3, we will focus on the lexicalization phenomenon itself, and on the different points of view it has been treated from.

3. 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).[7] Therefore, the result of lexicalization can be carried out as (i) a lexical element (a word)[8] or (ii) a sequence of elements (more than one word), that is, 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 (Calzolari et al. 2002) point out 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 types of MWEs: **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* as in (4), *decomposable idioms* as in (5), *compound nominals* including terminological MWEs as in (6), *proper names* as in (7), *verb-particle constructions* as in (8) and *light verb constructions* as in (9).

---

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

[8] Defined as "any string of characters between two blanks" (Fontenelle et al., 1994).

(4) a. English: to kick the bucket
b. English: to pull somebody's leg
c. Basque: adarra jo
'to pull somebody's leg'; lit: 'to play the horn'
d. Basque: larru bizirik
'stark naked'; lit: 'raw-skinned'

(5) a. English: to spill the beans
b. English: to play truant
c. Basque: burua jan
'to brainwash'; lit: 'to eat the head'
d. Basque: sudurra sartu
'to poke somebody's nose'

(6) a. English: car park
b. English: central processing unit
c. Basque: buruhauste
'problem'; lit: 'broken head'
b. Basque: sudur-zapi
'handkerchief'; 'lit: nose-cloth'

(7) a. English: Los Angeles
b. English: Chicago Bulls
c. Basque: Euskal Herri
'Basque Country'
d. Basque: Alderdi Popular
'Popular Party'

(8) a. English: set up
b. English: go after
c. Basque: -tzat (prolative case) + verb
'to take someone for'; lit: 'to take as'

d. Basque: -i (dative case) + eutsi (verb)
'defend'; lit: 'to hold to something'

(9) a. English: make a mistake
b. English: fall asleep
c. Basque: lan egin
'to work'; lit: 'to do work'
d. Basque: min hartu
'to hurt; lit: 'to take hurt'

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; for instance, (4c) literally means 'to play the horn' and it is translated to English as the idiom in (4b): *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 (5a), where *spill* and *beans* can be taken to have the appropriate senses that produces the compositional reading, or in Basque (5c).

A similar phenomenon occurs with *light verbs* –see examples in (9)– and c*ompound nominals* –see examples in (6).

Regarding to *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 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* in (10a) 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.

(10) a. English: traffic light
b. English: telephone box
c. Basque: zirkulazio-argi
'traffic light'
d. Basque: telefono-kabina
'telephone box'

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 in (4)), 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 examples in (5)).

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* –see example (11a). However, as example (11b) shows, 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**.**

(11) a. Basque: bizikletan ibili
'to bike'; lit: 'to walk on a bicycle'
b. Basque: bizikletan joan
'to bike'; lit: 'to go on a bicycle'
c. Basque: ahopeka abestu, ahopeka kantatu, ahopean abestu, ahopean kantatu
'to hum'; lit: 'to sing in whispers'

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.*

On the other side, they use the term **multiword lexical units (MLU)** to refer to lexicalized MWEs, those MWEs that are semantically non-compositional or syntactically idiosyncratic –all the examples mentioned from (4) to (9). As it can be seen, in this paper we have followed Alegria et al.'s (2004) terminology.

4. 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 LSKB.

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, examples (11a), (11b) and (11c). We nevertheless want to include such translations, as they are very useful information for translation tasks. Bentivogli & Pianta (2004) also emphasize the need to avoid lexical gaps as much as possible.

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 LSKB, where the

internal semantic structure of MWEs is represented, would aid in the semantic interpretation process. For instance, *fall_asleep* in (12) is a variant for a synset in WordNet 2.0, and capturing the relation between *asleep* and *fall_asleep* (very similar to *lo* and *lo_hartu* in Basque) would allow to better understand the consequences of *falling asleep.*

(12) fall asleep, dope off, flake out, […] -- (change from a waking to a sleeping state)

5. 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 head') in three Basque monolingual dictionaries;[9] 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 head'), 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.

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

Obviously, the first task is to detect the possible MWEs in Basque. They are detected in different stages during the development of Basque WordNet (Agirre et al. 2006). This way, the basis of the Basque MWEs are mainly the Basque counterparts of the English variants in the dictionaries. WordNet does include lexicalized synsets which may contain either single words or MWEs, or sometimes, both together:

(13) English WN {*girlfriend, girl, lady_friend*}

Then, the Basque WordNet builder must 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.

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

We have adopted the next criteria: if the MWE is an entry in a monolingual dictionary (Elhuyar 2000; Sarasola 1996; Euskaltzaindia 2000) or terminological glossary (UZEI 1987), then, the builder of the Basque WordNet will add this MWE in the synset, and it will be considered as a fully lexicalized MWE. For instance, *to memorize* is translated into Basque as both *buruz ikasi* (lit. 'to learn by head') and *memorizatu* (a loanword). Being *memorizatu* and *buruz ikasi* dictionary entries, the builder of the Basque WordNet will add both the loanword and the MWE in the synset:

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

In addition, it often happens that a MWE is the most usual way –and sometimes the only way– to express a concept, in spite of not being a dictionary entry. For instance, the English verb *to recite* is expressed in Basque either by the loanword *errezitatu* or either by the MWE *buruz esan*. Although this construction (*buruz esan*) is very similar to *buruz ikasi* ('to memorize' or 'to learn by head') and it is the most frequent and natural way to express this concept, according to our criteria, *buruz esan* will not be included in the synset. And as a consequence, it will not be

considered lexicalized MWE because it is not a dictionary entry. Therefore, 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.

In order to avoid this risk, we have decided to consider this type of MWEs *syntagmatic concepts* (Artola 1993), and to include them in the Basque WordNet. These refer to those concepts that are expressed by a phrase and that have become widespread in most of the cases. 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. Below, we present some more examples of Basque syntagmatic concepts:

(16) a. English WN {*recite, recite*}
Basque WN {*buruz_esan, errezitatu*}
b. English WN {*retranslate*}
Basque WN {*berriro_itzuli*} (lit. ‘translate again’)
c. English WN {*hum*}
Basque WN {*ahopeka_kantatu*} (lit. ‘sing in whispers’)
d. English WN {*bike*}
Basque WN {*bizikletan_ibili*} (lit. ‘move on a bike’)
e. English WN {*two-dimensional_figure* }
Basque WN {*irudi_bidimentsional*}

Therefore, instead of representing a lexical gap by adding an empty synset aligned with a non-empty synset of the other language (see Fig. 2), 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 the ones that are dictionary entries, they are marked with the syntagmatic concept label in the database, IXALEX (see Fig. 3).

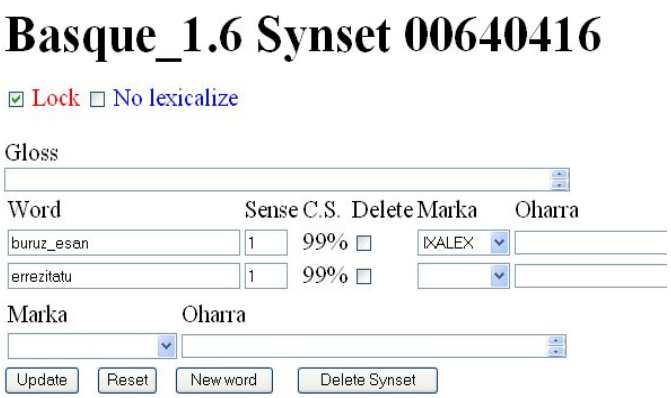
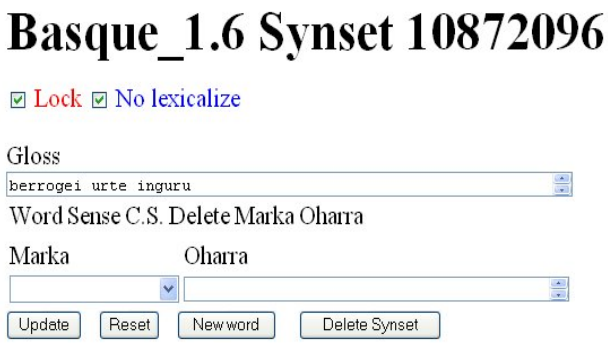


Figure 2: The actual interface, showing a syntagmatic concept (IXALEX is our shorthand for syntagmatic concept).

Finally, there are synsets that can be only expressed by a kind of definition. That is, they are expressed in a very different way than in English, using different syntactic categories as well as different phrase constructions.



English WN { forties, mid-forties -- (the time of life between 40 and 50) }

Basque WN {GAP -- (berrogei urte inguru)}

Figure 3: Representation of a GAP in the actual interface for the Basque WordNet.

For instance, in Basque, the only way to translate *forties* (see Fig. 2) is to use a kind of definition: *berrogei urte inguru* (lit. 'around forty years old'). We have decided not to include this kind of expressions in the synset but in the gloss. Therefore, these concepts will be lexical gaps in Basque.

6. Full representation of phrasal concepts in the Basque WordNet

The above representation (Section 5) 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 or the subject, depending on the transitivity). 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 syntactically 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 4, 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. 4c). But 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)

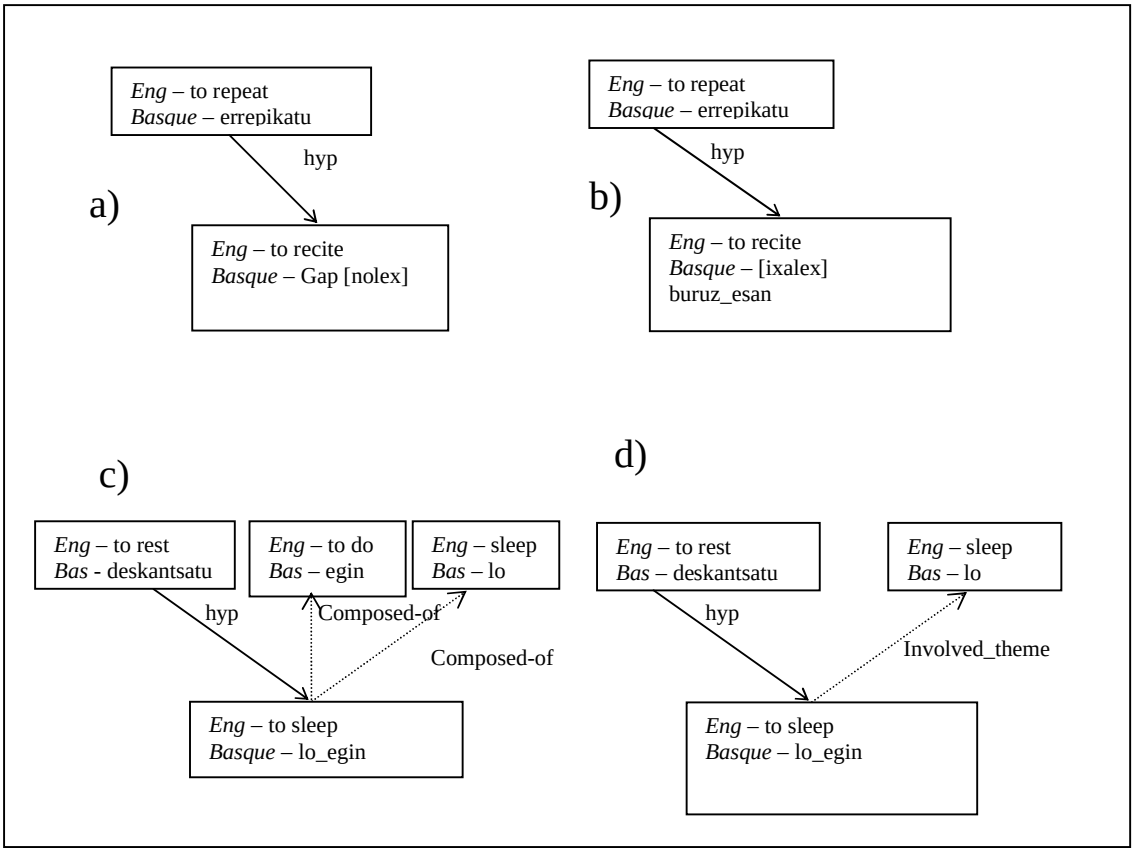


*Figure 4: Different representation choices in the Basque WordNet: **a)** representing only lexicalized multiwords, **b)** including syntagmatic concepts (ixalex being the internal tag for those), **c)** describing the inner structure using the "composed-of" relation, and **d)** describing the inner structure using EuroWordNet relations.*

We think that these relations are very well suited for encoding the inner relations. Scheme d) in Fig. 4 shows a possible representation for *lo_egin* where *lo* is the *involved_theme*[10] 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. 4 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. 4d can be applied to complex MWEs like (17a) or (17b). In fact, we will show that it can be applied to all kinds of MWEs.

(17) a. gabon kantak abestu
'to carol'; lit: 'to sing Chirstmas songs'

b. arinki lo egin
'to snooze'; lit: 'to sleep lightly'

7. 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. 4. 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.

---

[10] *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 this relation. Note that in English, $sleep_V$ and $sleep_N$ are also related by a *xpos_near_synonym* relation.

	TOTAL	N	V	ADJ	ADV
Word Senses	51423	41833	9450	140	0
Lemmas	25755	22492	3368	50	0
Synsets	31585	27880	3592	113	0
Basque lexical gaps	1439	1223	208	8	0
Proper Nouns		680			
MWE (lex)	5730	2935	2439	0	0
Syntagmatic concepts (ixalex)	356	79	273	4	0

Table 1: Current figures for the Basque WordNet and for gaps, lexicalized MWEs 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 6 (corresponding to scheme d) in Fig. 4). 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, Eneko & Mikel Lersundi. 2001. “Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición”. In *Proceedings of SEPLN 2001*. Jaén (Spain). 157-165.

Agirre, Eneko, Olatz Ansa, Xabier Arregi, Jose Mari Arriola, Arantza Díaz de Ilarraza, Eli Pociello & Larraitz Uria. 2002. “Methodological issues in the building of the Basque WordNet: quantitative and qualitative analysis”. In *Proceedings of First International WordNet Conference.* Mysore (India). 32-40.

Agirre, Eneko, Izaskun Aldezabal, Jone Etxeberria, Elixabete Izagirre, Karmele Mendizabal, Eli Pociello, & Mikel Quintian. 2005. “EuskalWordNet: euskararako ezagutza-base lexiko-semantikoa”. In *Euskalingua 7* (http://www.mendebalde.com). 212-219.

Agirre, Eneko, Izaskun Aldezabal, Jone Etxeberria, Elixabete Izagirre, Karmele Mendizabal, Eli Pociello & Mikel Quintian. 2006. “Improving the Basque WordNet by corpus annotation”. In *Proceedings of Third International WordNet Conference*. Jeju Island (Korea).

Agirre, Eneko, Izaskun Aldezabal & Eli Pociello. Submitted. “EuskalWordNet: euskararako ezagutza-base lexiko-semantikoa”. In *GOGOA.*

Alegria, Iñaki, Olatz Ansa, Xabier Artola, Nerea Ezeiza, Koldo Gojenola & Ruben Urizar. 2004. “Representation and Treatment of Multiword Expressions in Basque”. In *Proceedings of the ACL workshop on Multiword Expressions*. Barcelona (Catalunya). 48-55.

Alonge, Antonietta, Nicoletta Calzolari, Piek Vossen, Laura Bloksma, Irene Castellón, Toni Marti & W.Peters, 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, Jordi, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini, & Piek Vossen. 2004. “The MEANING Multilingual Central Repository”. In *Proceedings of the Second Global WordNet Conference.* Brno (Czech Republic). 23-30.

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

Bentivogli, Luisa & Pianta, Emanuele. 2004. “Extending wordnet with syntagmatic information”. In *Proceedings of Second Global WordNet Conference,* 47-53. Brno (Czech Republic).

Calzolari, Nicoletta, Charles J. Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci Catherine MacLeod & Antonio Zampolli. 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, Joan Miquel & Avellina, Sueñer. 2004. “Los procesos de lexicalización”. In E. Perez Gaztelu, I. Zabala & L. Gràcia (eds.), *Las fronteras de la composición en lenguas románicas y en vasco*. Universidad de Deusto. 47-109.

Cowie, Anthony P., Ronald Mackin, Isobel McCaig. 1990. *Oxford Dictionary of Current Idiomatic English.*

Fellbaum, Christiane. 1998. *WordNet: An electronic Lexical Database*. Cambridge, (Mass.); London (England): MIT Press.

Fontenelle, Thierry, Geert Adriaens & Gert de Braekeleer. 1994. “The Lexical Unit in the Metal® MT System”. The Netherlands. In *MT*, Volume 9, 1-19.

Lewandowski, Theodor. 1992. *Diccionario de Lingüística*. Cátedra.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger. 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)*. Mexico City (Mexico), 1-15.

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

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

Villavicencio, Aline, Francis Bond, Anna Korhonen & Diana McCarthy. 2005. “Introduction to the special issue on multiword expressions: Having a crack at a hard nut”. In *Computer Speech & Language*, Volume 19, 4. 365-377.

Vossen, Piek, Laura Bloksma, Salvator Climent, Toni Marti, Gabriel Oreggioni, Gerard Escudero, German Rigau, Horacio Rodriguez, Adriana Roventini, Francesca Bertagna, Antonietta Alonge, Carol Peters & Wim Peters. 1998. *The Reestructured Core wordnets in EuroWordnet: Subset1*. EuroWordNet (LE-4003) Deliverable D014/D015, University of Amsterdam.

Vossen, P. 1998. *EuroWordNet: A multilingual database with lexical semantic networks.* Kluwer Academic Publishers.

Dictionaries

Elhuyar, 2000. *Euskal Hiztegi Modernoa*.

Euskaltzaindia, 2000. *Hiztegi Batua.*

Sarasola, Ibon. 1996. *Euskal Hiztegia.*

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

Eneko Agirre Bengoa, IXA NLP Group.

E. Agirreren posta arrunta: UPV-EHU. Informatika Fakultatea. 649 p.k. - 20.080. DONOSTIA. Euskal Herria

E. Agirreren posta helbidea: e.agirre@ehu.es

Izaskun Aldezabal Roteta, IXA NLP Group, izaskun.aldezabal@ehu.es

Eli Pociello Irigoien, IXA NLP Group, elisabete@si.ehu.es