Methodology and Construction of the Basque WordNet

Elisabete Pociello

Elhuyar R&D

http://www.elhuyar.com
Zelai Haundi kalea, 3. Osinalde Industrialdea, 20170 Usurbil. Basque Country

E-mail: eli@elhuyar.org

Fax: (+34) 943.36.31.44

Eneko Agirre, Izaskun Aldezabal

IXA NLP Research Group

http://ixa.si.ehu.es
649 pk. 20.080 – Donostia. Basque Country

E-mail: {e.agirre, izaskun.aldezabal}@ehu.es

Fax: (+34) 943.01.55.90

Abstract

Semantic interpretation of language requires extensive and rich lexical knowledge bases (LKB). The Basque WordNet is a LKB based on WordNet and its multilingual counterparts EuroWordNet and the Multilingual Central Repository. This paper reviews the theoretical and practical aspects of the Basque WordNet lexical knowledge base, as well as the steps and methodology followed in its construction. Our methodology is based on the joint development of wordnets and annotated corpora. The Basque WordNet contains 32,456 synsets and 26,565 lemmas, and is complemented by a hand-tagged corpus comprising 59,968 annotations.

Keywords: *lexical semantics, lexical knowledge bases, WordNet.*

1 Introduction

This paper presents work on a Basque lexical knowledge base, the Basque WordNet, and describes its construction from the quest for an appropriate model to its development

Natural Language Processing (NLP) techniques for semantic interpretation require lexical knowledge bases (LKB). LKBs are structured lexical resources that organize the information in the lexical entry in order to prevent redundancy. Nowadays, LKBs dominate the lexical-semantic field of NLP, as they offer a number of advantages for knowledge representation: information in the lexical entries can be structured, redundancy can be resolved, data can be controlled, consistence can be achieved and information capture can be made easier. Besides, information can be maintained and updated, including the management of versions.

In order to deal with computational semantics, our research group set the following requirements for the Basque LKB:

- **The LKB should cover a wide range of language phenomena**, including senses, semantic classes and syntactic-semantic information such as thematic roles, subcategorization and selectional preferences.
- **It should have a large coverage of the vocabulary**, so it can be used in free text.
- **It should not be linked to a single theory,** in other words, it should have the capacity to take advantage of other models or formalisms.
- **It should be computational**, one that can be used in NLP.
- **It should be multilingual**, so in addition to lexical entries in Basque, it would make equivalents in other languages available.

There are many and very different proposals for designing an LKB. We examined and evaluated them according to the above criteria, including theory oriented models —Jackendoff (1990), Levin (1993), Pustejovsky (1995)— and computational models –*FrameNet* (Fillmore and Baker, 2001), *WordNet* (Miller, 1985; Fellbaum, 1998), *EuroWordNet* (Vossen, 1997), *Multilingual Central Repository* (MCR) (Atserias *et al.*, 2004), *Volem* (Fernández *et al.*, 2002), *PropBank* (Palmer and Kingsbury, 2003). From our analysis we concluded that the large coverage of WordNet was a very important feature. The proven multilingual extensions of WordNet were also taken into account. Although WordNet misses information in the syntactic-semantics interface, these were left for later development[1]. Another feature was that there were already several wordnets under development. At present more than 50 national languages are registered within the Global WordNet Association[2]. The Global WordNet Association is a free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world.

Our team started to build the Basque WordNet following the EuroWordNet design in 2000; and in 2003, in the context of the MEANING Project (Rigau *et al.*, 2003)— the Basque WordNet was moved to the MCR, an advanced version of EuroWordNet.

The paper is organized as follows. We first briefly describe WordNet, EuroWordNet and the MCR in Section 2. Section 3 presents the methodology for developing our LKB. Section 4 explains the treatment of linguistic phenomena, giving special attention to the criteria defined for representing them. Finally, Section 5 outlines some conclusions and summarizes future work.

2 WordNet, EuroWordNet and the MCR

WordNet (Miller, 1985; Fellbaum, 1998) is an LKB for English based on psycho-linguistic theories developed at Princeton University. Nouns, verbs, adjectives and adverbs are grouped together into synonym sets or **synsets**, each one corresponding to a single lexical concept. For example, the English noun *tree* has two senses in WordNet, which are represented as two different synsets:

(1) Sense 1: tree (a tall perennial woody plant having a main trunk and branches. . .)
Sense 2: tree, tree diagram (a figure that branches from a single root; “genealogical tree”)

[1] In order to see the specific analysis and the conclusions drawn from it, refer to (Pociello, 2008).
[2] At http://www.globalwordnet.org

The first sense corresponds to the ‘plant’ meaning, and the second to the ‘diagram’ meaning. The first synset is made up of a single lexical unit (*tree*), in other words, the noun *tree* in that synset has no other synonym. The second synset contains an additional lexical unit (*tree diagram*), so these two lexical units (*tree* and *tree diagram*) are synonyms. The lexical unit in each synset is known as a **literal**. **Synonymy** is an important relation in WordNet, and the structure of the LKB is based on the meanings of the lexical units; when the same meaning is shared by more than one lexical unit, the lexical units are grouped together into a synset.

In addition to synonymy, WordNet represents several relations. For instance, the **hypernymy** relation links general synsets to more specific ones[3]. **Hyponymy** is the inverse relation. The hypernymy chain and a subset of the hyponyms of the synset corresponding to the first sense of ‘plant’ in Example 1 can be seen in Examples 2 and 3, respectively[4].

(2) Sense 1
tree (a tall perennial woody plant having a main trunk and branches…)
=> woody plant, ligneous plant – (a plant having hard lignified tissues…)
=> vascular plant, tracheophyte – (green plant having a vascular system…)
=> plant, flora, plant life – (a living organism lacking the power of locomotion)
=> life form, organism, being, living thing – (any living entity)
=> entity, something – (anything having existence (living or nonliving))

(3) Sense 1
tree (a tall perennial woody plant having a main trunk and branches…)
=> yellowwood, yellowwood tree (any of various trees having yellowish wood…)
=> lancewood, lancewood tree (source of most of the lancewood of commerce)
=> Guinea pepper, negro pepper, Xylopia aethiopica (tropical west African evergreen tree…)
=> anise tree (any of several evergreen shrubs and small trees of the genus Illicium)
=> winter’s bark, winter’s bark tree, Drimys winteri (South American evergreen tree…)
=> zebrawood, zebrawood tree (any of various trees or shrubs having or striped wood)
=> granadilla tree, Brya ebenus (West Indian tree yielding a fine grade of green ebony)
=> acacia (any of various spiny trees or shrubs of the genus Acacia)
=> …

Example 2 gives an idea of the WordNet hierarchy or taxonomy, indicating that a tree is a woody plant, which is a vascular plant, which is a plant, which is a life form, which is an entity. In Example 3 we show a partial list of kinds of trees. The hypernymy hierarchy can be used to define **semantic classes**, that is, a synset can be seen as the semantic class that groups all its hyponyms. For example, all the different kinds of trees are direct or indirect hyponyms of the synset representing the ‘plant’ meaning of *tree*. We can thus take this synset as the semantic class together all tree species. In the case of verbs, **troponymy** is used to encode the hierarchy of verbs, where verb *Y* is a troponym of the verb *X* if the activity *Y* is doing *X* in some manner.

As an illustration of the richness of relations in WordNet we will briefly mention three. A relation which holds between nominal synsets is **meronymy**, which is used to represent the *part of* relation, e.g. a *finger* is part of a *hand* and a *hand* is a part of an *arm*. Verbal synsets can be related by **entailment**, e.g. *snoring entails sleeping*. Adjectival synsets can be linked to nominal synsets with the **related-to** relation, e.g. *nice* and *niceness*.

[3] The hypernymy/hyponymy relation is also referred as the subset/superset relation.
[4] All the expressions have been taken from WordNet 3.0 (http://wordnetweb.princeton.edu/perl/webwn), with some editing in synsets, literals and glosses due to space limitations.

WordNet is one of the most cited lexical resources in the NLP literature, with more than 38,000 hits in Google Scholar[5] and many applications in wide range of tasks. WordNet is complemented with SemCor, a corpus hand-tagged with WordNet senses (Miller et al., 1994; Fellbaum et al., 2001). WordNet is freely available[6].

2.1.1 EuroWordNet

The EuroWordNet project (Vossen, 1998) is a European project that was started in 1996 and went on until 1999, and produced wordnets[7] for eight European languages (English, Danish, Italian, Spanish, German, French, Czech and Estonian). EuroWordNet follows the Princeton WordNet model, but incorporates cross-lingual links. Each language in EuroWordNet has an "independent" wordnet with its own relations, but the synsets in one language can be linked to the so called Inter-Lingual-Index (ILI), which is largely based on the Princeton WordNet. EuroWordNet is available from ELRA[8].

In addition to the ILI, EuroWordNet includes several new features. EuroWordNet has more kinds of language-internal relations, and some of the semantic relations of WordNet are refined and/or enriched. Domain ontologies and a Top Ontology were added. The first one organizes synsets according to domains like *free time*, *restaurant*, or *traffic*. The second one enables relevant synsets of the different wordnets to be classified according to basic semantic features[9] based on linguistic features (e.g. [+/– living], ([+/– agent]).

Finally, EuroWordNet introduced the notion of Base Concepts[10]: the concepts that play the most important role in the various wordnets of different languages, as measured by their high position in the semantic hierarchy and their having many relations to other concepts. The motivation was to reach maximum overlap and compatibility across wordnets in different languages, while at the same time, allow for the distributive development of wordnets in the world.

2.1.2 The Multilingual Central Repository (MCR)

The MCR was devised in the context of MEANING (Rigau *et al.*, 2003), an European project which run from 2002 to 2005. The MCR follows the EuroWordNet model, including five languages: Basque, Catalan, English, Italian and Spanish. The wordnets were enriched with new kinds of information, like domain tags for synsets, the Suggested Upper Merged Ontology (Niles and Pease, 2001), or selectional preferences (Agirre and Martínez, 2002).

3 Methodology for building the Basque WordNet

In this Section we will present the phases and methodological issues regarding the construction of the Basque WordNet. We will first introduce general issues, followed by the methodology for nouns, and the joint development of a hand-

[5] A WordNet bibliography with more than 400 is maintained at http://lit.csci.unt.edu/~wordnet/.
[6] http://wordnet.princeton.edu/
[7] We use *WordNet* (upper case) for the original Princeton WordNet, while we use *wordnet* (lower case) for the rest.
[8] http://catalog.elra.info/
[9] Although top ontologies classify a limited number of synsets, the synsets below them can also inherit the classification.
[10] http://www.globalwordnet.org/gwa/gwa_base_concepts.htm

tagged corpus. Finally, we will describe the methodology for verbs. Note that we have not addressed adjectives and adverbs yet.

3.1 Design and methodology

There are two main options to create a new wordnet: we could create the Basque WordNet afresh based on Basque corpora and dictionaries, or we could take the Princeton WordNet and translate its synsets into Basque. Vossen (1999) referred to these two approaches as *merge approach* and *expand approach*, respectively.

In the first approach the senses and hierarchies in the Basque WordNet would be independent of the senses and hierarchies in the Princeton WordNet. This involves heavy lexicographic work in order to build the sense inventory and the hypernymy hierarchy. In addition, the multilinguality will require to manually add cross-lingual links to the ILI (cf. Section 2.1.1). In the second approach, the work is basically reduced to linking Basque words to the English concepts via the ILI, i.e. we can reuse the synsets and relations in the English wordnet, and translate the literals in the synsets into Basque. We would thus avoid most of the lexicographic work and the need to link Basque synsets to the ILI. On the weak side, there is the risk to misrepresent cultural differences in the sense inventories and hierarchies.

After analyzing the pros and cons of each approach, the decision was taken to use the expand approach, taking the English WordNet as the starting point for building the Basque WordNet. Special care will be placed in detecting cultural differences. For instance, some new concepts will be needed for words like *trikitixa* –Basque accordion and related songs– or *ikastola* –schools where Basque is the main language. In parallel, we also decided to study automatic construction of LKBs from dictionaries, in order to explore the potential of the merge approach and possible combinations (Agirre and Lersundi, 2001; Lersundi, 2005).

3.2 Methodology for nouns

The methodology to build the Basque WordNet changed during the different stages in its evolution. In a first stage, the goal was to build a first fast version of the Basque WordNet, with an emphasis on wide coverage, i.e. the number of lemmas. In this stage, the 1,024 Base Concepts of EuroWordNet (cf. Section 2.1.2) were manually translated into Basque, and then Basque-English bilingual dictionaries were used to automatically create Basque equivalents for the rest of English synsets (Agirre et al. 2002).

In the next stage, the main goal was to ensure quality. We initially devised two complementary steps. Firstly, a team of linguists manually inspected the automatically generated synsets for Basque, **concept by concept**. In this process the linguists checked to see whether the Basque equivalent for the synset was appropriate or not; and a check was also made to see whether any other equivalents of Basque were needed in the synset. The focus of this process was to ensure that the literals in the Basque synsets were correct. After this inspection was completed, the team embarked on the second step, inspecting the words and the respective synsets **word by word**, trying to ensure that the main senses of the words as occurring in a dictionary (Elhuyar, 1998) where properly represented. These two steps involved looking at the same data from two complementary perspectives, ensuring proper quality in the synsets of the Basque WordNet.

Halfway through the word-by-word inspection, we realized that linguists were paying increasing attention to real word examples as occurring in a corpus. In fact, the linguists had to examine existing corpora to check that the main senses of the words were properly represented in the Basque WordNet. Since they were

already analyzing the examples of a target word, we thought that they could actually annotate the examples with the senses of the target word, and produce a Basque semantic concordance (Basque SemCor for short). This methodology was inspired by Fellbaum *et al.* (2001) who propose that dictionaries and corpora should be used together. We thus started the joint development of the Basque WordNet and SemCor.

3.3 Joint development of Basque WordNet and SemCor

First of all, we compiled a corpus of approximately 300,000 words[11], including samples from a balanced corpus and a newspaper corpus. The goal is to coordinate the tagging of the corpus with the word-to-word review of the Basque WordNet. The synsets corresponding to the target word will be edited according to the examples in the corpus, thus ensuring that the Basque WordNet contains the synsets and literals as used in the corpus.

The motivations of this methodology are the following: (i) the manual annotation of the corpus guarantees that the sense-inventory and sense boundaries fit those found in the corpus (in particular, all senses occurring in the corpus will be reflected in the Basque WordNet), (ii) the senses in the Basque WordNet are tuned to real occurrences of the words, and not only to existing monolingual dictionaries (thus ensuring that the synsets reflect the real usage of the words), (iii) the annotated corpus provides a companion resource for enriching Basque WordNet with richer semantic relations acquired from corpora (Atserias et al., 2004), including the relative frequency of the senses for a given word and (iv) the annotated corpus will enable to build word sense disambiguation programs for Basque.

We implemented the joint development with a team of five linguists with the following roles: one supervisor, one editor, two taggers and one referee. The editor is the one who edits the synsets. The taggers tag the occurrences of the word that needs to be tagged. The referee compares the work of the two taggers and resolves any disagreements. The supervisor coordinates the team.

In short, the methodology followed by this team is as follows: (i) The editor selects a handful of words[12], edits the synsets corresponding to those words introducing the necessary changes. (ii) The editor tries to convene the meaning of the target words to taggers and referee, ensuring that they have a common understanding. (iii) The taggers tag the occurrences of the target words. (iv) Basque glosses and examples are added to the synsets (Agirre et al. 2005). (v) When all these tasks have been completed, the taggers inform the editor and the referee and explain the problems they encountered while tagging. (vi) The referee compares the results of the two taggers, resolving inconsistencies. (vii) In addition, should new senses of the words appear in the corpus, the editor will examine the suitability of these new senses that appeared in the corpus prior to deciding whether to incorporate them into the Basque WordNet for posterior tagging. Figure 1 summarizes this cyclical methodology:

[11]Given that Basque is an agglutinative language, it has a higher lemma/word rate than English. Estimates in parallel corpora allow us to think that 300.000 words in Basque are comparable to 500.000 words in English.

[12] Nouns in the corpus were ordered according to frequency, from most to least frequent. The editor follows this order to select words. That way it is possible to ensure that the most frequent nouns are properly edited and tagged.

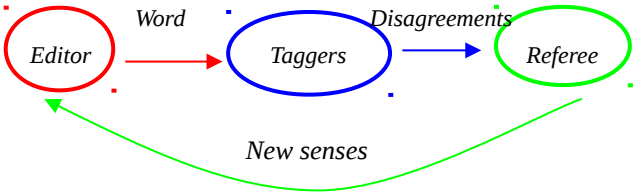

Figure 1: Roles and workflow for the joint development of the Basque WordNet and SemCor.

At present, we have applied this methodology to nouns. We organized the tagging starting with the most polysemous nouns. We also reviewed all monosemous nouns in the most frequent list, leaving aside those which we think needed a new sense in the Basque WordNet. These were edited and tagged in the next stage. The words not in the Basque WordNet are mainly proper nouns, but the list needed to be revised, in order to find common nouns that need to be included in the Basque WordNet, and tagged accordingly. The conclusion Section shows the current figures for the Basque WordNet.

We next present the agreement figures among the taggers. As already mentioned, each occurrence in the corpus was tagged by two different taggers. The referee had to resolve all disagreements between the taggers. In order to facilitate his work a number of data was presented to him, including confusion matrixes, and agreement figures. Inter-tagger agreement (ITA) was computed as the percentage of occurrences where the two taggers agreed over the total of the occurrences. In case of any of the taggers assigning more than one tag to an occurrence, a tag in common between the two taggers is sufficient to be considered an agreement. Inter-tagger agreement can be misleading for words with different numbers of senses or senses with different distributions, i.e. an agreement of 80% for a word with two senses where one sense accounts for 90% of all occurrences is very low, while it would be a very satisfactory figure for a word with 10 evenly distributed senses.

The Kappa coefficient (Carletta, 1996) overcomes the shortcomings of the ITA measure by substracting from ITA the chance agreement (given the number and distribution of the senses) and normalizing from 0 to 1. Our referee was satisfied with the use of the Kappa figure, but she also found the ITA measure useful as a more intuitive measure of agreement.

On average, the taggers attained 84% ITA and a Kappa coefficient of 0.68. Tables 2 and 3 show the 5 words with lowest and highest scores respectively.

	Kappa	ITA	senses	occ.
familia	-0.46	0.18	6	81
indarkeria	-0.44	0.08	5	114
aste	-0.19	0.36	5	173
histori	-0.18	0.18	7	54
urrats	-0.05	0.41	7	63

Table 2: 5 words with worst Kappa (respectively *family*, *violence*, *week*, *history*, *step*). ITA, senses and number of occurrences are also given.

	Kappa	ITA	senses	occ.
ipar	1.00	1.00	5	102
kontratu	1.00	1.00	3	52
hiri	1.00	1.00	4	87
partidu	1.00	1.00	5	465
anaia	1.00	1.00	3	44

Table 3: 5 words with best Kappa (respectively *north*, *contract*, *city, match, brother*). ITA, senses and number of occurrences are also given.

We want to mention that Kappas over 0.7 are deemed reasonable for well-defined tasks. While most of our words are over this threshold, some words attain very low scores. We have found that most of the disagreements are systematic for each word, i.e. each of the taggers understands differently the sense boundaries and applies his conceptualization systematically, leaving certain kind of occurrences under different senses each. The meetings between the taggers and the referee highlighted that most of these differences were due to an insufficient characterization of the senses, where the glosses were not clear. These meeting served to review the glosses and sense differentiations in the Basque WordNet, and complement WordNet with a number of examples which have been coherently tagged with its senses. In fact, we think that if the taggers were given a representative number of tagged examples to supplement the WordNet glosses, the agreement rates would be much higher.

Another reason for the low agreement is that the team would need more time to prepare each of the words. In contrast to other hand-tagging tasks like PoS tagging or treebanking, sense-tagging has the peculiarity that each word is in fact a different task. Knowing and interiorizing the sense boundaries can be very time-consuming, and needs to be repeated for each word. After the tagging-refereeing-editing cycle we are quite sure that the tagged examples and the sense definitions are a coherent set produced by a well-interiorized model of the word.

3.4 Methodology for verbs

The methodology for verbs was slightly different. As a first step, we attached Basque verbal literals to Base Concepts, as we did for nouns, and then applied automatic methods followed by a synset-to-synset review. Given the richer syntactic-semantic information encoded in the verbs, we wanted to make sure that the word-to-word review as done for nouns was convenient for verbs. The next Subsection presents verb sense distinction as defined in WordNet, and we will then see the verb polysemy is defined differently from nouns, as we will see in the next Section. We will now review the representation of verbs in WordNet, and then present out methodology for Basque verbs.

3.4.1 Verbs in WordNet

WordNet uses syntactic-semantic information to form verbal synsets. The synset components have to have the same selectional restriction and subcategorization. Failure to abide by this will result in the verbs being distributed among different synsets, as in the following examples.

(4) Mary ate an apple.
(5) Mary ate.

The verb *eat* can be used as a transitive (4) or intransitive verb (5), and each use is distinguished in two senses (*eat*_1 and *eat*_2) corresponding to two senses as shown in Example 6.

(6) {eat 1} (take solid food; "She was eating a banana")
{eat 2} (eat a meal; "We did not eat until 10 P.M.")

The syntactic-semantic information encoded in the synset also influences the hierarchy and semantic classes. For instance, each of the synsets in 6 defines a different semantic class. *eat*_1 has transitive troponyms like *gobble*, *gulp* or *devour*, and *eat*_2 has intransitive troponyms like *dine*, *snack*, *picnic* and *breakfast*. In the former, the troponyms indicate 'ways of eating', while the troponyms of the later incorporate that which is eaten.

Unfortunately, bilingual dictionaries don't always include such syntactic-semantic nuances, and the wordnet editor needs to study the syntactic-semantic behavior of the Basque equivalents. For instance, the Basque equivalent of *eat* (*jan*) also has an intransitive form (*Hagina kendu diote eta ezin du jan* ["He's had a tooth out and can't eat"]) and a transitive form (*Bazkaltzeko haragia jan dut* ["I've had meat for lunch"]) and thus the two synsets in Example 6 also apply for Basque.

3.4.2 Analysis for incorporating verbs into the MCR

Given the importance of syntactic-semantic features when deciding sense differences and troponyms for verbs, we considered whether a **hierarchy oriented** edition of the Basque WordNet would be preferred over the **word-by-word** method we had been using for nouns**.** Thus, we did two pilot studies following each of the possible methods.

In the word-by-word pilot we chose to study five highly polysemous verbs: *esan* ["to say"], *banandu* ["to separate"], *banatu* ["to distribute"], *abestu* ["to sing"] and *ekarri* ["to bring"]). Given the limited syntactic information available in the dictionaries used (*Elhuyar Hiztegia* (Elhuyar, 1996) and the *Elhuyar Hiztegi Modernoa* (Elhuyar, 2000)) we had to take into account the classification and sub-categorization information included in Aldezabal (2004). In our experience, this pilot showed that the word-by-word edition ensures that all the senses of the verb are properly edited, but in could lead to errors and imbalances in the hierarchy. For example, some of the literals in a troponym could be more general than the literals of their hypernym, because the editor focused on the word and its senses, but not on the hierarchy. Furthermore, in order to understand the syntactic-semantic information inherent in some synsets and choose the appropriate Basque literals, the editor had to check the hierarchy, as in the troponyms of *eat* such as *devour* or *picnic*.

For the hierarchy oriented pilot, we chose a hierarchy with an average number of synsets {*express*_2, *give_tongue*_1, *utter*_1}, and proceeded top-down starting from the top synset. Using this method the editors were satisfied in that they ensured that the hierarchy was balanced and that the Basque literals had a coherent syntactic-semantic behavior, but they observed that some meanings of the verbs would be easily missed.

Given our experience in the two pilots, we saw that neither method was completely satisfactory. One solution would be to first follow the hierarchies, and later do the word-by-word check, but unfortunately this could be too costly. Another alternative would be to work word-by-word and do limited checks in the immediate hypernym and troponyms of the involved synsets. The advantage of the latter alternative is that it can be coupled with the manual tagging of the verbs in the Basque SemCor. Given the added value of a coupled WordNet-SemCor development, we concluded that this was the preferred solution, also for verbs.

4 From WordNet to Basque WordNet: distinguishing features and enhancements

In this Section some distinguishing linguistic features that emerged during the edition of the Basque WordNet will be presented, and how we coded them in the underlying MCR database. Section 4.1 presents some features related to lexicalization. Section 4.2 reviews the hierarchical organization. Finally, Section 4.3 presents a proposal for a richer internal representation of multiword expressions (MWE).

4.1 Lexicalization

The term *lexicalization* refers to the transformation of an element (or a sequence of elements) into a unique lexical or conceptual element (Lewandowski, 1992). Therefore, the result of lexicalization can be carried out as (i) a lexical element (a word) or (ii) a sequence of elements or multiword expressions (MWEs). The aforementioned "transformation" is an obscure process, especially, with MWEs.

Lexicalization is a key issue when building a wordnet, as the editors of the wordnet will need to decide whether a word or sequence of words should be an entry in the wordnet or not, but unfortunately, in practice, the boundaries for lexicalization are very difficult to draw (Contreras & Sueñer, 2004; Cowie, 1990; Calzolari et al., 2002; Sag et al. 2002), and this is the reason why the job of deciding whether the word or sequence of words is lexicalized is usually very difficult. This difficulty becomes apparent when comparing two languages, or, as in this case, when taking one LKB built for one language (WordNet) as the starting point for the LKB of another language (Basque WordNet).

In WordNet only lexicalized concepts are included, whether they are lexicalized by single words (*pet*, *lyrics*, *sleep*, etc.) or MWEs (*mid-forties*, *tree diagram*, *military man*, etc.). However, in the process of constructing the network of words and concepts, WordNet developers found that in many cases it was necessary to postulate general concepts that happen not to be lexicalized in English (Fellbaum, 1998). These general concepts have been added with the aim of organizing the hierarchy (cf. Section 4.2.1).

When English literals are to be translated into appropriate Basque literals, the editor often comes up against lexicalization problems, because there are **conceptual level imbalances** and **expression level imbalances.**

Among conceptual imbalances there are *cultural concepts*, concepts that appear linked to a particular culture and which do not exist in other languages, e.g. a simnel cake in English is "a cake eaten in England around Easter time", and a *trikitixa* in Basque is a "Basque accordion". These concepts are expressed in other languages by means of explanations or definitions and translated just as they are, using the same word as in the source language. Such synsets in EuroWordNet

used to be left empty (without literals) and are referred to as *cultural gaps* (Vossen, 1999). We will explicitly code that they are not lexicalized.

Expression level imbalances occur when a concept is known in the two languages, but when different expressions are used in each one. For example, some synsets in English are translated into Basque through multiword expressions (*pet*: *konpainia-animalia; cook*: *janaria egin*), or through an inflectional suffix (*cold*: *hotz**ez**, hotz**ik***) or through a number mark (*furnishing*: *altzari**ak***). It is not easy to rule on the lexicalization of these pragmatic gaps, especially if dictionaries are taken as the basis: *lo egin* ["to sleep"; lit. "to do sleep"] is a dictionary entry, whereas *janaria egin* ["to cook"; lit. "to do food"] is not; *hotzik* is a dictionary entry whereas *hotzez* is not.

Insofar as language is creative, it goes on creating new word combinations, and even though we understand them it is difficult to say whether they are lexicalized or not. This, of course, leads to problems when deciding whether or not to include such a word in the Basque WordNet. Such things in EuroWordNet used to be left blank as in conceptual imbalances, but Vossen (1999) refers to these cases as *pragmatic gaps*. But in the Basque WordNet, aware of the difficulty in ruling on lexicalization, a decision was taken to include these expressions of doubtful lexicalization in the LKB, as we will see next.

4.1.1 Need for expressions of doubtful lexicalization

As a general rule, the criteria used to decide whether or not to incorporate certain equivalents in the LKB are specified according to external factors and the use that one wants to make of the LKB. In our case, we want a Basque WordNet that is good basis for the semantic interpretation of Basque, so that it can be helpful in certain NLP tasks. Our aim is to enrich Basque WordNet with as large a number of equivalents as possible, since they are very useful for conducting semantic interpretation and sense disambiguation: the more equivalents there are in the Basque WordNet, the easier it will be for a program to disambiguate the senses. On the other hand, conducting deep reflection on lexicalization does not figure among the aims of our work; moreover, if too much time is spent on deciding about the lexicalization of each equivalent, the development of Basque WordNet would be slowed down tremendously.

In order to work coherently on lexicalized, non-lexicalized Basque WordNet literals and ones of doubtful lexicalization, the criteria and tags presented in the next Section have been specified.

4.1.2 Criteria for incorporating Basque literal into the Basque WordNet and marking them

We developed three criteria for adding Basque literals into a Basque synset, as follows.

- **First criterion**: if the Basque expression is a **dictionary entry** in the following dictionaries *Elhuyar Hiztegia*, *Hiztegi Modernoa*, *Euskal Hiztegia*, *Euskalterm* or *Hiztegi Batua*[13], then the editor will regard this expression as lexicalized and will incorporate it into the synset with the LEX mark:

  (7) **Synset number**: 00009805

[13] The reasons for choosing these dictionaries should be pointed out: firstly, we were given the chance to use them electronically, because of the close contacts the IXA Group has with the dictionary makers; and secondly, because the dictionaries are widely used for specialised (Euskalterm) and general purposes.

=> **Lexicalization situation of the synset**: LEX
=> **Gloss**: Lo-egoeran egon [“to be in a sleep situation”]
=> **Synonyms**: *lo egin* [“to sleep”]

- **Second criterion**: If the Basque expression is an MWE, and if it is not a dictionary entry in *Elhuyar Hiztegia, Hiztegi Modernoa, Euskal Hiztegia, Euskalterm* or *Hiztegi Batua:*

  (a) if the concept can be translated **without using a definition** in Basque, then the editor will incorporate the expression as a literal, and will mark it as a **syntagmatic expression** (SYNTAG-LEX) to indicate that it is an MWE that it is not a dictionary entry (see Example 8).

  (b) If **a definition** has to be used to express the concept, then the editor will incorporate the MWE not as a literal but as a gloss. These would be *cultural gaps* (Vossen, 1999) and have been marked as **non-lexicalized** (NOLEX) (see Example 9, which corresponds to *simnel*).

(8) **Synset number**: 01143604
=> **Lexicalization situation of the synset**: SYNTAG-LEX
=> **Gloss**: elikagaiak jateko prestatu [“to prepare food for eating”]
=> **Synonyms**: *janaria prestatu* [“to prepare food”], *janaria egin* [“to cook”]

(9) **Synset number**: 05678078
=> **Lexicalization situation of the synset**: NOLEX
=> **Gloss**: Ingalaterran Pazko inguruan jaten den gozokia [“a sweet eaten in England at Easter”]
=> **Synonyms**:

- **Third criterion**: If a form having a **plural** or **inflectional suffix** has to be used to express a concept, then the editor will incorporate the literal without the plural or inflectional suffix, and will mark it with PLU (see Example 10, which corresponds to *altzariak*) or INFL (see Example 11, which corresponds to *hotzek* and *hotzik*), to show that the concept takes the plural quality or the inflectional suffix, respectively.

(10) **Synset number**: 02729592
=> **Lexicalization situation of the synset**: PLU
=> **Gloss**: Hainbat zereginetarako erabiltzen diren objektu higigarriak [“movable objects used for many purposes”]
=> **Synonyms**: *altzari* [“piece of furniture”]

(11) **Synset number**: 01199751
=> **Lexicalization situation of the synset**: INFL
=> **Gloss**: Bero-gabeziak gorputzean eragiten duen sentsazioa [“sensation felt by the body caused by lack of warmth”]
=> **Synonyms**: *hotz* [“cold”]

4.2 Hierarchical distinctions

Since we are using the merge approach, the Basque WordNet follows the same hierarchical classification as WordNet. Unfortunately Basque literals cannot

simply be inserted into a synset just because it they are a translation of a literal in the English synset, as they need to share the meaning expressed by the sysnset, and because coherence has to be maintained in the hierarchy.

In this respect, we recognized two major issues: an equivalent which is not lexicalized has to be invented for the purposes of organizing the hierarchy (what will be referred to as *conceptual organizers*), and when the English hypernymous-hyponymous literals are lexicalized with the same equivalents in Basque, known as *autohyponimy* (Cruse, 2000). These issues are also linked to lexicalization, but they refer to lexicalization problems from the hierarchical organization of WordNet.

4.2.1 Conceptual organizers

The term *conceptual organizer* refers to general concepts devised to organize the hierarchy. They tend to appear at the top of the hierarchy and are necessary for classifying semantic classes. For example, the English synset which groups together the types of characteristics distinguished by sight (color, darkness, texture, etc.) is called *visual property*. This concept is not lexicalized in Basque, but it can be used for giving a name to the semantic class that brings together all the synsets that express types of visual property (150 hyponyms in all).

In WordNet they are listed as exceptions, because in this LKB they are the only non-lexicalized synsets (Fellbaum, 1998), and an MWE is needed to express their meanings.

In the Basque WordNet we will include a description as the literal, and a mark to signal that the synset is not lexicalized and that it has been added for the purpose of organizing the hierarchy. The mark is NOLEX-GENERAL, general in English.

(12) **Synset number**: 03871460
=> **Lexicalization situation of the synset**: NOLEX-GENERAL
=> **Gloss**: ikusmenak duen ezaugarria [“the property of vision”]
=> **Synonyms**: *ikusmenezko ezaugarri* [“visual property”]

4.2.2 Hierarchies and lexical specificity

For some Basque words, we found that it was not easy to find the right level in the hierarchy. Before going into details we will review *autohyponymy*. The senses of a polysemous lexical unit can be hypernyms/hyponyms of each other. Basque WordNet, for example, gives the following example:

(13) {pertsona_1, gizabanako_1, lagun _15} (a human being)
=> {adiskide_7, lagun_10} (a person you know well and regard with affection and trust)

Lagun can thus mean a human being, but also can refer to a friend, where one synet is hyponym of the other. Cruse (2000) calls this kind of polysemy *autohyponymy*:

> “Autohyponymy occurs when a word has a default general sense, and a contextually restricted sense which is more specific in that it denotes a subvariety of the general sense." (Cruse, 2000, p. 110)

In the process of building the Basque WordNet it is possible to generate what we call *false autohyponym*. Synsets are translated while conducting the editing, and sometimes the same word in Basque was used both for the hypernym and

hyponym, without considering whether these senses in Basque were really distinguished. When we started the word by word manual editing (Section 3.2), more attention was paid to the hierarchy, and it was at that point that it became clear that in the Basque hierarchy the number of autohyponym synsets was much higher than in the English hierarchy: there were more than four thousand autohyponyms in Basque and only 26 in English. Example 14 gives a partial list of the hyponyms of merrymaking[14], and Example 15 the corresponding Basque literals.

(14) {celebration, festivity} (any festival or other celebration)
=> {merrymaking} (boisterous celebration)
=> {revel, revelry} (noisy partying)
=> {bout, spree} (a drunken revel)
=> {bender, bust} (an occasion for heavy drinking)
=> {carouse} (a merry drinking party)
=> {orgy} (a wild gathering involving drinking and promiscuity)
=> {whoopee} (noisy and boisterous revelry)
=> {…}

(15) {festa, jai} (event or party organised to celebrate something)
=> {parranda} (boisterous celebration)
=> {parranda} (noisy partying)
=> {parranda} (a drunken revel)
=> {parranda} (an occasion for heavy drinking)
=> {parranda} (a merry drinking party)
=> {orgia} (a wild gathering involving drinking and promiscuity)
=> {parranda} (noisy and boisterous revelry)
=> {…}

If these hierarchies are compared, we can see that English uses different words to refer to each of the synsets, while many of those synsets can be lexicalized by the Basque word *parranda*. When doing the word-by-word review and consulting the dictionaries, it was clear that the Basque word *parranda* did not differentiate all those meanings, and was thus a case of *false autohyponymy*, in contrast to Example 13, which is a genuine autohyponym.

In order to deal with *false autohyponymy*, it was decided that the lowest hyponyms (insofar as they are translated by a literal in the hypernym) would be left without literals, and a different mark would be used to distinguish them from other non-lexicalized synsets, namely NOLEX-AUTOHYPO. For instance, Example 16 shows how we finally coded one of the hyponyms of *merrymaking* in Basque (the synset corresponding to *revelry*).

(16) **Synset number**: 00328944
=> **Lexicalization situation of the synset**: NOLEX-AUTOHYPO
=> **Gloss**: jai zaratatsua (noisy party)
=> **Synonyms**:

In order to decide whether we are facing a true autohyponymy or not, we resort to dictionaries. Basque is a language currently undergoing a standardization process, and equivalents of these concepts could exist outside dictionaries, as some words from the dialects and specific domains have yet made it to our dictionaries.

[14] The whole semantic class of the example has 22 hyponyms, but in the example only the direct hyponyms of the hyponym *merrymaking* have been given. The number of literals of the synsets has also been reduced.

Autohyponymy has been also treated in other wordnet projects such as BalkaNet (Stamou *et al*., 2002). In order to detect those synsets or cases that could indicate lexicographers mistakes, in BalkaNet project, a set of checks were developed, and autohyponymy review was included in one of those checks (Tufis *et al.*, 2004). In other approaches (Gonzalo *et al.*, 2000, and Peters *et al.*, 2000), autohyponymy has been also used to cluster senses.

To conclude, note that the process to enrich the Basque WordNet has been done on the basis of the English synsets. Although we have the impression that English has more lexicalized concepts due to a more specific and precise vocabulary, we would need to perform complementary experiments, that is, take a native Basque hierarchical organization, and translate it into English.

4.3 Semantic internal representation of MWEs in the Basque WordNet

MWEs are common place in wordnets, but their internal representation has not been included in WordNet, EuroWordNet or the MCR. Bentivogli and Pianta (2002) proposed a model for internal representation based on the MWEs of the Italian wordnet. These authors used a *composed-of* lexical relation between a MWE literal and its component words. In section A) of figure 2, the MWE *lo egin* [“to sleep”; lit. “to do sleep”] has been given as an example. This synset, like any other synset, will be semantically linked to its hypernym (*deskantsatu* [“to rest”]) and its troponyms (*siesta egin* [“to have a nap”], *kuluxka bat egin* [“to doze”], *hibernatu* [“to hibernate”], etc.). But in addition, each component (*lo* [“sleep”] and *egin* [“do”]) that forms the MWE will have a *composed-of* link with its corresponding word form, indicating that the MWE in that synset is made up of two word forms belonging to two other synsets.

We will be using the lexical relation *composed-of* in the Basque WordNet, because the components of the MWEs that are formed compositionally seemed suitable to us for representation purposes. Nevertheless, in addition to that relation, the internal representation of the components making up the MWE can be specified further. For example, this *composed of* lexical relation does not express the syntactic-semantic relation between the MWE’s components. Let us take the sentence *umeak lo egin zuen* [“the child slept”]) as an example in which we have a light verb structure: *lo egin*. Semantically, the *composed-of* relation in this sentence does not indicate that as part of the act *lo egin* [“to sleep”] is the situation of being *lo* [“asleep”]. Syntactically, neither does it indicate that the nominal component of this MWE (*lo*) is the syntactic object of the multiword verb expression (*lo egin*) and that the latter will assume a thematic role.

To express syntactic-semantic information, the EuroWordNet lexical relation called *involved relation* will be taken as the basis. The *involved* relation starts from a noun (a word form of synset) that expresses a verb or action in order to lexically link it to a concrete or abstract noun (another word form of another synset). For example, the English verb *to hammer* will be linked to the noun *hammer* through an *involved instrument* relation. There are eight types of *involved* relations: *agent, patient, instrument, result, location, direction, source direction and target direction.*

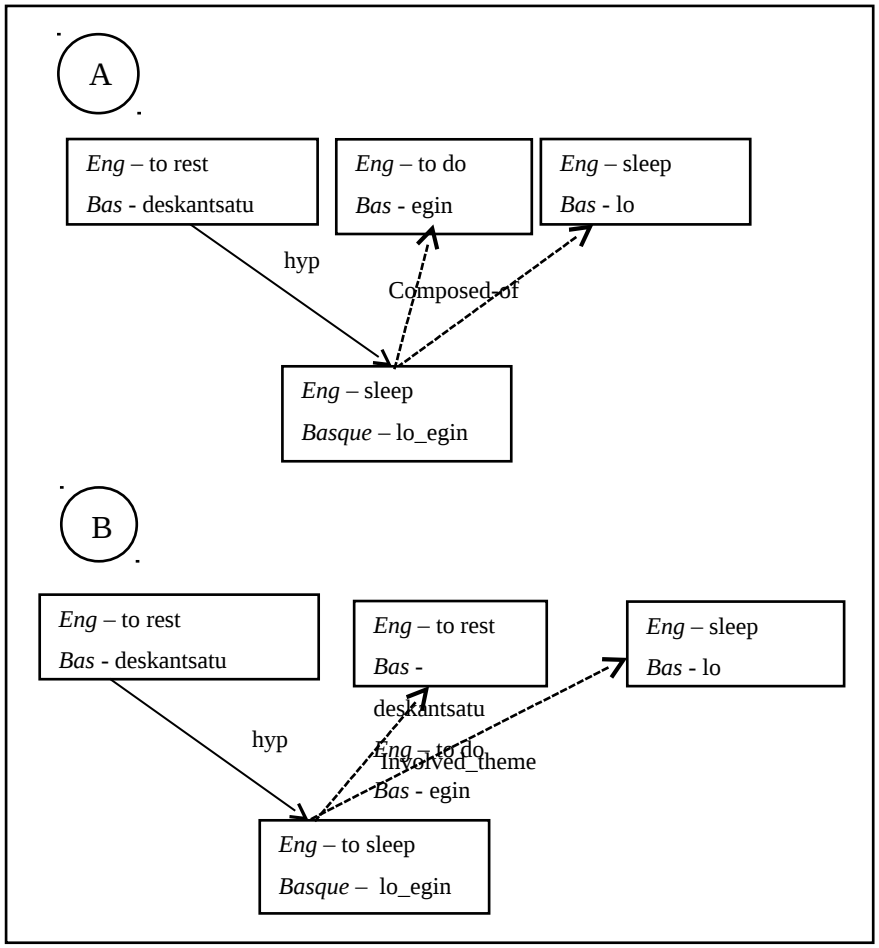

Figure 2: Different MWE internal representations.

In our view the *involved relation* is highly suited to representing internal structures. In section B) of Figure 2, one has the representation of the MWE *lo egin* in which, besides the *composed-of* relation, the *involved relation* is also used: the word form *lo* is the subject of the MWE (*involved patient*), and it enables us to know that *lo egotea* [“being asleep”] is necessary for *lo egiteko* [“to go to sleep”].

	Total	***Nouns***	***Verbs***
Senses	50,670	41,160	9,510
Lemmas	26,565	23,069	3,496
Synsets	32,456	28,705	3,751
Lexical Gaps	2,499	2,198	301
Named Entities	722	722	0

Table 4: Basque WordNet figures, corresponding to the senses, lemmas and synsets. The lexical gaps correspond to synsets which are not lexicalized in Basque. Named entities correspond to lemmas which are proper nouns, and to synsets that are instances, rather than semantic classes.

	Done				*Total*	
	Lemmas		**Occurrences**		**Lemmas**	**Occurrences**
Polysemous	1,015	30,3%	51,427	72,9%	3,354	70,546
Monosemous	307	16,2%	9,179	54,0%	1,897	16,990
Not in Basque WordNet	118	1,1%	1,374	3,6%	10,959	37,877
Total	1,355	8,4%	59,968	47,8%	16,210	125,413

Table 5: Basque SemCor figures for nouns. We list separately polysemous, monosemous and those lemmas not in the Basque WordNet. In the rows we have lemmas and their respective cooccurrences, with two columns for those who have already been tagged (including percentage with respect to the total).

5 Conclusions and future work

The main outcome of this piece of research is the design and development of a multilingual LKB, the Basque WordNet, which is fundamental for the applied semantic analysis of Basque. We first have developed a quick core Basque WordNet using semi-automatic methods that include a concept-to-concept manual review, and later performed an additional word-to-word review based on Basque lexical resources that guarantees the quality of the wordnet produced. Moreover, we have also presented our methodology for the joint development of the Basque WordNet and a complementary corpus for Basque, the Basque SemCor. This methodology consists on editing the the words in the Basque WordNet, double-blind tagging of Basque SemCor with a referee for adjudication, and a farther editing-tagging cycle when required. We have compared this methodology to the hierarchical method, and have concluded that the word-to-word review and joint corpus tagging is the best method to guarantee quality. One shortcoming of the word-to-word method is that we created autohyponyms along the way, but a quick check of the hypernym and hyponyms while doing the review would suffice to prevent this problem in the future.

Table 4 and 5 show the current figures for the Basque WordNet and the nouns in the Basque SemCor, respectively. Note that we have tagged the most frequent lemmas, which correspond to 47,8% of all occurrences. Our word-to-word review has gone through 1,015 nominal lemmas, accounting for 72,9% of the total number of occurrences.

We are satisfied for the results so far. The cost of developing both resources jointly is higher than doing it separately, but the quality justifies the effort, as attested for the improvements of the Basque WordNet after annotating the corpus, and the improved annotation after reviewing WordNet. The joint development guarantees high-quality Basque WordNet and SemCor, which we are confident now that can be used to treat real corpora.

We have also described the linguistic phenomena that emerge when creating a multilingual LKB, defining the required criteria and enriching the MCR model for representing these issues in wordnets. These criteria cover lexicalization, hierarchical distinctions, conceptual organizers and autohyponymy issues. In addition we have enriched the wordnet model with a proposal for the internal representation of the internal structure of MWE, which will be also useful to include more internal relations in the wordnets.

In the future, we plan to finish the tagging of polysemous nouns, and the joint review and tagging of verbs. We are also working on the extension of WordNet to particular domains, including the connection to terminological dictionaries using

semi-automatic methods [Pociello *et al.*, 2008]. We would also like to feed the internal representation of MWEs, following the semi-automatic methods (Agirre and Lersundi, 2001).

We would also like to explore the complementarities of the expand and merge approaches. We plan to incorporate the hierarchies and semantic relations extracted from other Basque dictionaries at a large scale (Agirre *et al.*, 2003).

The Basque WordNet is available from ELRA[15], following the WordNet-LMF dialect of the Lexical Markup Framework (Francopoulo et al. 2007). WordNet-LMF is the first application of LMF to wordnet-like applications, and allows for a rich and principled representation of the information contained in wordnets. The release of a free subset is planned in the near future. Both the Basque WordNet and Basque SemCor can be browsed online[16].

References

Agirre, E., & Lersundi, M. (2001). Extraccióon de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. In Proceedings of the Annual SEPLN Meeting, 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., & Martinez, D. (2002). Integrating Selectional Preferences in WordNet. In Proceedings of First International WordNet Conference, Mysore (India).

Agirre, E., Ansa, O., Arregi, X., Artola, X., Zubillaga, X., Díaz de Ilarraza, A., & Lersundi, M. (2003). A conceptual schema for a Basque lexical-semantic framework. In Conference on Computational Lexicography and Text Research, Budapest (Hungary).

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, E., Mendizabal, K., Quintian, M., & Pociello, E. (2005). EuSemCor: euskarako corpusa semantikoki etiketatzeko eskuliburua: editatze- etiketatze- eta epaitze-lanak. Technical report, University of the Basque Country.

Aldezabal, I. (2004). Aditz-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa. Levin-en (1993) lana oinarri hartuta eta metodo informatikoak baliatuz. PhD thesis, University of the Basque Country.

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

Bentivogli, L. & Pianta, E. (2002). Extending WordNet with syntagmatic information. In Proceedings of Second Global WordNet Conference, Brno (Czech Republic).

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), Las Palmas (Spain).

Carletta, J. (1996). Assessing agreement on classication tasks: the kappa statistic. Computational Linguistics, 22(2):249-254.

[15] http://catalog.elra.info/

[16] Basque WordNet: http://ixa2.si.ehu.es/mcr/wei.html.
Basque SemCor: http://sisx04.si.ehu.es:8080/euSemCor

Contreras, JM., & Sueñer, A. (2004). Los procesos de la lexicalización. In E. Perez Gaztelu, I. Zabala, & L. Gràcia (Eds.), Las fronteras de la composición en lenguas románicas y en vasco.(pp. 47-109). University of Deusto.

Cowie, A.P., Mackin, R., & McCaig, I.R. (1990). Oxford Dictionary of Current Idiomatic English: Verbs With Prepositions and Particles, v2. London: Oxford University Press.

Cruse, A. (2000). Meaning in Language: An Introduction to Semantics and Pragmatics. London: Oxford University Press.

Elhuyar (1996). Elhuyar Hiztegia: euskara-gaztelania. Donostia: Elhuyar Kultur Elkartea.

Elhuyar (1998). Elhuyar Hiztegi Txikia. Donostia: Elhuyar Kultur Elkartea.

Elhuyar (2000). Hiztegi Modernoa. Donostia: Elhuyar Kultur Elkartea.

Euskaltzaindia (2000). Hiztegi Batua. Donostia: Elkar.

Fellbaum, C. (1998). WordNet. An Electronic Lexical Database. MIT Press, Cambridge (Massachussetts).

Fellbaum, C., Palmer, M., Dang, H.T., Delfs, L., & Wolf, S. (2001). Manual and automatic semantic annotation with WordNet. In Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources, Pittsburgh.

Fernández, A., Saint-Dizier, P., Vázquez, G., Kamel, M., & Benamara, F. (2002). The Volem Project: a framework for the construction of advanced multilingual lexicons. In Proceedings of Language Engineering Conference (LEC'02), Hyderabad (India).

Fillmore, C.J. & Baker, C.F. (2001). FrameNet: Frame semantics meets the corpus. In Proceedings of WordNet and Other Lexical Resources Workshop, Pittsburgh.

Francopoulo, G., Bel, N., George, M., Calzolari, N., Monachini, M., Pet, M. & Soria, C. (2007). Lexical Markup Framework: ISO standard for semantic information in NLP lexicons. GLDV (Gesellschaft für linguistische Datenverarbeitung), Tubingen.

Jackendoff, R.S. (1990). Semantic Structure. MIT Press, Cambridge (Massachussetts).

Lersundi, M. (2005). Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktikosemantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak . PhD thesis, University of the Basque Country.

Levin, B. (1993). English Verb Classes and Alternations. A Preliminary Investigation. The University of Chicago Press.

Lewandowski, T. Diccionario de Lingüística. Cátedra, 1992.

Miller, G.A. (1985). WordNet: a dictionary browser. In Proceedings of the First International Conference on Information in Data, Waterloo.

Miller, G.A., Chodorow, M., Landes, S., Leacock, C., & Thomas, R.G. (1994). Using a semantic concordance for sense identification. In Proceedings of the ARPA Human Language Technology Workshop, San Francisco.

Niles, I. & Pease, A. (2001). Towards a standard upper ontology. In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems,FOIS 2001, Ogunquit, (Maine)

Kingsbury, P. & Palmer, M. (2002). From TreeBank to PropBank..In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas (Spain).

Pociello, E. (2008). Euskararen ezagutza-base lexikala: Euskal WordNet. PhD thesis, University of the Basque Country.

Pociello, E., Gurrutxaga, A., Agirre, E., Aldezabal, I. & Rigau, G. (2008). WNTERM: Combining the Basque WordNet and a Terminological Dictionary. Proceedings of the 6th International Conference on Language Resources and Evaluations (LREC), Marrakech.

Pustejovsky, J. (1995). The Generative Lexicon. Cambridge: MIT Press.

Rigau, G., Agirre, E., & Atserias, J. (2003). The MEANING project. In Proceedings of the XIX Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), Alcalá de Henares (Madrid).

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

Stamou, S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D. & Grigoriadou, M. (2002). Balkanet: A Multilingual Semantic Network for the Balkan Languages. In Proceedings of First International WordNet Conference, Mysore (India)

Tufis, D., Cristea, D., & Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. Romanian journal of Information science and technology, 7-1-2, pp. 9-44.

UZEI (1987). Euskalterm. http://www1.euskadi.net/euskalterm/indice_c.htm. Accessed 17 March 2010.

Vossen, P. (1997). EuroWordNet: a multilingual database for information retrieval. In Proceedings of the DELOS Workshop on Cross-language Information Retrieval, Zurich.

Vossen, P. (1998). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers.

Vossen, P. (1999). EuroWordNet general document. EuroWordNet (LE2-4003, LE4-8328), Part A, Final Document Deliverable D032D033/2D014.