A methodology for the joint development of the Basque WordNet and Semcor

Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Izagirre, Karmele Mendizabal,
Eli Pociello and Mikel Quintian*
IXA NLP Group
649 pk. 20.080 – Donostia. Basque Country
e.agirre@ehu.es

Abstract

This paper describes the methodology adopted to jointly develop the Basque WordNet and a hand annotated corpora (the Basque Semcor). This joint development allows for better motivated sense distinctions, and a tighter coupling between both resources. The methodology involves edition, tagging and refereeing tasks. We are currently half way through the nominal part of the 300.000 word corpus (roughly equivalent to a 500.000 word corpus for English). We present a detailed description of the task, including the main criteria for difficult cases in the edition of the senses and the tagging of the corpus, with special mention to multiword entries. Finally we give a detailed picture of the current figures, as well as an analysis of the agreement rates.

1. Introduction

This paper presents current work on the Basque WordNet and Semcor. Our team started to build the Basque WordNet (Agirre et al., 2002) following the EuroWordNet design (Vossen et al., 1998) in 2000. The Basque WordNet has been constructed with the expand approach, which means that the English synsets have been enriched with Basque variants. The Basque WordNet is currently aligned with WordNet 1.6, which is the main version of the MEANING Multilingual Central Repository (Atserias et al., 2004).

The initial stage of the construction of the Basque WordNet was focused on coverage; we generated automatically Basque equivalents using bilingual dictionaries (Atserias et al., 1997), and then, we performed a concept-to-concept review where the linguists focused on the correctness of the variants in the synset.

We then focused on quality and started a word-to-word review of word senses. The goal was twofold: to ensure the quality across word senses and to try to cover the main senses for most frequent/relevant words (for more details refer to Agirre et al., 2002).

This review was half way through when we started the Basque Semcor project. At this stage we decided to change our methodology and turned to the coordinated development of the word-to-word review of the Basque WordNet and the manual annotation of a sizeable Basque corpus.

The benefits of this decision 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 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.

This paper is structured as follows. Section 2 explains the methodology for this joint development, including the criteria for editing the Basque WordNet and for tagging the Basque Semcor. Special attention will be paid to multiword expressions, which is a recurrent problem in the design of lexical knowledge-bases. Section 3 presents current figures of the Basque WordNet and Semcor, as well as an analysis of the inter-tagger agreement and kappa figures for the tagging process. In Section 4 we will briefly mention some related work. Lastly, Section 5 summarizes future work and outlines some conclusions.

2. Methodology

Five people, graduate linguistics students, take part in this coordinated development: a supervisor (part-time), an editor (part-time), two taggers (part-time) and a referee (full-time). The editor *edits* the Basque WordNet; he takes care of revising the synsets of the Basque WordNet, ensuring that all variants for the synsets are properly placed, and conversely, that all senses for a word are linked to appropriate synsets. The two taggers independently tag all the examples for the target word. The referee reviews the disagreements between both taggers and takes the final decision.

The detail of the process is the following. The editor looks up a word in the dictionary, and checks that all the senses are correctly represented in the Basque WordNet. In this process, he may add new synsets or delete incorrect ones according to a sample of the target corpus and the available monolingual dictionaries[1]. The editor is the one who decides the preliminary sense inventory of a word. The word to be reviewed by the editor is chosen from a word-list arranged in descending order by their frequency in the corpora.

Once the sense inventory of a word is reviewed, the two taggers independently tag the same examples for that word. The tagging method is based on what Kilgarriff (1998) called *transversal annotation*: instead of tagging the sentences in the corpora token by token, the taggers

* Authors listed in alphabetic order.

[1] Consider that at this stage we are revising an imperfect Basque WordNet, so errors and omissions are possible.

annotate word-type by word-type, that is, all the occurrences of a word first, then all the occurrences of another word, and so on. Through this approach, the semantic characteristics of each word are taken into consideration only once, and the whole corpus achieves greater consistency. In the other alternative, the linear process, the annotator must remember the sense structure of each word and their specific problems each time the word appears in the corpus, making the annotation process much more complex, and increasing the possibilities of low consistency and disagreement between the annotators (Navarro et al., 2003).

The referee, helped by a program that computes the agreement rates (inter-tagger agreement and kappa) and a confusion matrix, reviews the disagreements and decides which the correct tag is.

Finally, if new senses of a word have come out in the corpus, the referee will inform the editor, and the editor, after checking whether those new senses are correct, will add them in the Basque WordNet. The taggers then update the corpus tags accordingly. Figure 1 shows a schematic diagram of this process.

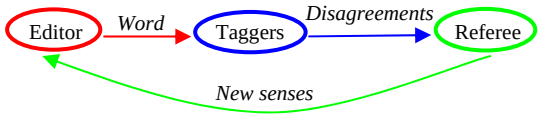

Figure 1: Representation of the cyclic editing-tagging process.

We will se now in more detail the main features of the edition and tagging process, followed by a subsection on the treatment of multiword expressions in both steps.

2.1. Editing the Basque WordNet

As already mentioned, the Basque WordNet is being constructed enriching English synsets with Basque variants. Finding an appropriate translation of an English synsets into Basque can be problematic: the category used to express a concept may not match accross both languages, or there may be different ways of writing a word in Basque, etc. In order to have a consistent treatment of these issues, we defined a detailed set of criteria, and a set of labels for the Basque variants. Due to space constraints, below we only describe, very briefly, some of the most relevant criteria. Refer to Agirre et al. (2005) for more details.

- **Specific terminology**

The English WordNet includes some specific words and terms (e.g. *bar mitzvah, focal ratio, pond scum*, etc). In order to translate this kind of synsets into Basque, the Basque WordNet editor would need to look up these concepts in specific terminological dictionaries. As our first aim is to try to cover the main senses of the most frequent/relevant words in Basque, we have decided to mark these synsets as non-lexicalized for the time being. However, we add a *specific concept* label to them, in order to differentiate them from the other non-lexicalized synsets.

- **Dialectal concepts**

Some words in Basque are mainly dialectal and they are not as frequent as their synonyms. For instance, the Basque word *egunkari* is mainly used to refer to English word *newspaper*. However, dictionaries also indicate that it also means *day labourer*, without mentioning that this only happens in a Basque dialect, and therefore this meaning is rarely used. People would rather use its synomym *jornalari*. We include this kind of words in their respective synsets in the Basque WordNet, but we mark them with the label *rare*.

- **Different ways of spelling a word**

Taking into account that Basque is a language still in course of standardization, some words in Basque can be written in more that one different ways, and both forms are deemed correct. For instance, to express the concept of *policeman*, in Basque we can use either '*polizi agente*' or '*polizia-agente*', which are alternative spellings. In these cases, we include both forms in the Basque WordNet.

- **Auto-hyponymy**

As it is well-known, nominal synsets in wordnets are structured by the hyperonymy-hyponymy relation. Auto-hyponymy refers to words that have two (or more) senses, one hyponym of the other. While the frequency of this phenomenon in the English WordNet is rare, we have found it to be quite common in the Basque WordNet. Example (1) shows some hyponyms of the synset *end_1*, where we can see that its hyponyms in English are expressed with different words (*point, pinpoint, tip*).

(1) => end (either extremity of something that has length)
=> pinpoint (the sharp point of a pin)
=> point (sharp end)
=> tip (the extreme of something)
=> …

In Basque we use a single word to refer to all these concepts *(mutur*), including the hyperonym. We have no definitive explanation for the high frequency of this phenomenon in Basque. One possibility could be that Basque words have more general meanings than English counterparts. Another explanation could arise from the fact that we take the structure and sense differentiations from the English WordNet, and Basque is probably organized differently. In most of the cases our linguists think that the most specific meanings are not really lexicalized in Basque, and only the most general term is lexicalized. For example, in (1) we would only mark the sense equivalent to 'end' as a lexicalization of *mutur*. In any case, we want to leave a trace of auto-hyponymy for further studies, so, although we do not add *mutur* to the other hyponyms of '*end*', we do mark them as non-lexicalized and add them a special label (*specific hypernym*) to differentiate them from other non-lexicalized synsets.

2.2 Tagging the Basque Semcor

As well as in the edition task, tagging has also showed the need for detailed criteria. In particular, ome occurrences cannot be tagged with a synset for a number of reasons. We devised a detailed inventory of such cases, which are tagged as **Special Cases (SC)**. We present them briefly here. Refer to (Agirre et al., 2005; Agirre et al., 2006b) for more details.

- **SC1: Word exists in WordNet but not its sense**

With this special case taggers mark those occurrences that do not match any of the synsets proposed by the editor. This mark is used to mark new senses.

- **SC2: Word does not exist in WordNet**

This special case was created to mark those words that appear in the corpus, but that have no synset in WordNet. Usually, these are words related to Basque culture, such as *ikastola* ('Basque school'), *trikitixa* ('Basque dance'), etc. This special case was devised when we were unsure about what to do with new synsets. We finally decided that the editor introduces the new words before tagging, and therefore we never used this mark.

- **SC3: Word is part of a Multiword Lexical Unit**

If a word occurrence is part of a multiword lexical unit taggers use this mark. For instance, if an occurrence of *urte* ('year') is followed by the word *berri* ('new'), it will be marked with Special Case 3, signalling that the word is part of a multiword: *urte berri* ('new year').

- **SC4: Word is (a part of) a Named Entity.**

Sometimes, an occurrence may be a named entity or part of a named entity, and taggers mark it with this special case. This is the case for *herri* ('country') when occurring as *Euskal Herri* ('Basque Country').

- **SC5: The tagger is strongly uncertain**

This special case is available for those cases where the tagger is uncertain and does not know how to tag one occurrence. It is usually used when the context is not enough to disambiguate an occurrence.

- **SC6: Word was improperly lemmatized**

Some errors can have their source in lemmatization. For instance, the noun *etxe* ('house') can get genitive-case: *etxe* + genitive-case "*-ko*" = *etxeko* ('of house'). However, this form (*etxeko*) can be used as an adjective in Basque to express 'home-made': *etxeko gazta* ('home-made cheese'). These forms are quite difficult for the lemmatizer to detect, and as a consequence, the adjective *etxeko* is lemmatized as: *etxe* (noun) + genitive-case "*-ko*". Special Case 6 is used to mark this problematic cases.

- **SC7: Word is wrongly used**

Some occurrences in the corpora are wrongly used, i.e. they are misspellings or ungrammatical. This tag occurs with relatively high frequency due to the ongoing process of standardization of Basque. For instance, the corpus contains occurrences of the word *pake* which has recently been standardized as *bake*.

2.3 Multiword expressions

An important issue recurrent in the design of lexical knowledge bases is the treatment of multiword expressions[2] (MWEs). Both the boundaries for lexicalization and the types of MWEs are very difficult to draw (Contreras & Sueñer, 2004; Cowie, 1990), and this is why the task of deciding which MWE is lexicalized or not, is one of the main tasks of a wordnet builder.

In the Basque WordNet MWEs are treated as: (i) fully lexicalized, (ii) syntagmatic concepts, (iii) non-lexicalized and (iv) lexical gaps.

We consider a MWE as **fully lexicalized** when 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 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:

(2) 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:

(3) **a**. English WN {*hum*}
Basque WN {*ahopeka_kantatu*} (lit. 'sing in whispers')
b. English WN {*bike*}
Basque WN {*bizikletan_ibili*} (lit. 'move on a bike')

In order to differentiate these MWEs from the ones that are dictionary entries, they are marked with the syntagmatic concept label in the database, IXALEX.

In previous stages of the construction of the Basque WordNet, that is, before we had decided how to include MWEs in the Basque WordNet (Agirre et al., 2006), these MWEs that were not dictionary entries but were frequently used, were provisionally added in the Basque WordNet but they were marked as **non-lexicalized**. In this way, we could easily detect them for their further revision. At present, these non-lexicalized MWEs will be reviewed according to criteria explained above, and most of them will be classified as sytagmatic concepts.

[2] Note that we use *multiword expression* as a general term to denominate those constructions, either lexicalized or not, containing more than one word (*word* defined as "any string of characters between two blanks" by Fontenelle et al., 1994).

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. These are known as **lexical gaps**. For instance, in Basque, the only way to translate *forties* 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.

To put in briefly, we actually distinguish between fully lexicalized MWEs, syntagmatic concepts and lexical gaps. In the first case, we would have a normal synset with its variants. In the second case, we can include a broader range of MWEs, marked with a special flag. And in the last case, the synset would not have a counterpart in Basque.

Therefore, we have argued in favor of including non-lexicalized MWEs in order to:

i) avoid having 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.

ii) treat 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).

iii) facilitate semantic interpretation and a richer LSKB, that is to say, 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.

Therefore, this representation allows us listing the MWEs together with their lexicalization status. However, in order to reflect the inner structure and semantic relations in the MWE, we have also motivated and proposed a representation based in EuroWordNet relations.

In the future, we would like to enrich the Basque WordNet with the MWEs extracted from the Basque Semcor. As we have mentioned in section 2.2, taggers mark as Special Case 3 those occurrences that they think can be part of a MWE. However, as there is low agreement between taggers, for the moment, we have left aside their revision and these will be reviewed, edited and tagged in the next stage.

In addition, the tagging of the MWEs would be easier if the lemmatizer could detect them. In this way, the Special Case 3 would not be necessary and MWEs could be directly edited, tagged and reviewed. Therefore, we need to change the lemmatization. However, this process must be synchronized with the Basque lexical database, and it will be quite complex.

3. Current data and analysis of the methodology

In this section we give the main figures for the Basque WordNet and Semcor. In addition we present the inter-tagger agreement rates for the taggers.

Table 1 shows the current figures for the Basque WordNet. We have mainly worked on the nominal part, with nearly 28 thousand synsets for 22 thousand lemmas, and an overall ratio of 2 to 1 senses per lemma. We have also worked on the most frequent verbs, which explains the high polysemy (3 to 1 senses per lemma) for the 3 thousand lemmas and synsets. We have only worked on a small sample of adjectives, and no adverbs. We also mention the number of proper nouns and genuinely non-lexicalized synsets. In addition we have nearly 6 thousand non-lexicalized synsets which are deemed syntagmatic concepts, so we have their surface realization available.

	TOT	N	V	ADJ
Word Senses	51,423	41,833	9,450	140
Lemmas	25,755	22,492	3,368	50
Synsets	31,585	27,880	3,592	113
Proper Nouns		680		
Basque gaps (no lex)	1,439	1,223	208	8
MWE (no lex)	5,730	2,935	2,439	0
Syntagmatic concepts	356	79	273	4

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

The corpus under annotation was compiled with samples from a balanced corpus and a newspaper corpus. It comprises 300,000 words in total. 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.

At the time of writing, the methodology has been going for 18 months. Up to now, we have only worked with nouns and we have already done 52% of the occurrences of polysemous 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 need a new sense in EWN. These will be edited and tagged in the next stage, but we already started to mark those that are genuinely monosemous. The words not in EWN are mainly proper nouns, but the list needs to be revised, in order to find common nouns that do need to be included in EWN, and tagged accordingly. We also left them for the next stage.

	Done		To be done		Total	
	words	occ.	words	occ.	words	occ.
Polysemous	256 (7%)	36,345 (%52)	2,956	32,473	3,212	68,818
Monosemous	448 (28.4%)	9,214 (70%)	1,127	3,827	1,575	13,041
Not in EWN	-	-	3,995	20,909	3,995	20,909
Total	704 (8%)	45,402 (44%)	8,078	57,366	8782	102,768

Table 2: Current figures for the nouns in the Basque Semcor.

Table 2 show the current figures. We can see that we have done 256 polysemous nouns, but given their high frequency, they account for 36,345 occurrences (53% of the occurrences of all polysemous nouns). We have considered as 'done' both nouns tagged with senses (184 nouns, 24,188 occ.) and nouns left untagged (72 nouns, 12,157 occ.), which correspond to lemmatization errors or words which normally do not function semantically as nouns (being for instance part of complex postpositions

like). Table 2 also lists the 448 nouns from the monosemous list which have been deemed to be really monosemouns (accounting for 71% of all monosemous occurrences), and the words not in EWN.

At this stage of the tagging we have changed the methodology. For the rest of the corpus, instead of having two taggers plus referee, we plan to use a single tagger per word, except some problematic words. The rationale is that less frequent words should be easier to tag, with a few exceptions. With a single tagger we estimate that we will need approximately 12 months to finish all nouns, including the revision of the rest of monosemous nouns and all the nouns not in 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.

We computed inter-tagger agreement (ITA) 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 3 and 4 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 3: 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.n	1.00	1.00	5	102
kontratu.n	1.00	1.00	3	52
hiri.n	1.00	1.00	4	87
partidu.n	1.00	1.00	5	465
anaia.n	1.00	1.00	3	44

Table 4: 5 words with best kappa (respectively *noth*, *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 will be much higher.

Another reason for the low agreement is that the team would need more time to prepare each of the words. Sense-tagging, in contrast to other hand-tagging tasks like PoS tagging or treebanking, has the peculiarity that each word is in fact a different task. Knowing and interiorizing the sense boundaries can be very time-consuming task, 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.

4. Related work

WordNet has already been used for corpora annotation (Fellbaum et al., 2001; Navarro et al., 2003; Pianta et al., 2005). However, few wordnets have been developed together with tagged corpora. For instance, Navarro et al. (2003) use a frozen version of the Spanish WordNet for semantic annotation, and they don't update it. Pianta et al. (2005) translate the English to Italian, and then port the English annotations to Italian via corpus alignment. They mention briefly that they plan to use the information in the alignment to enrich WordNet.

5. Conclusions and future work

We have presented our methodology for the joint development of the Basque WordNet and the Basque Semcor, consisting if editing WordNet, double-blind tagging of Semcor with a referee for adjudication, and a farther editing-tagging cycle when required. We are satisfied for the results so far: even if the cost of developing both resources jointly is higher than doing it separately, 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 annotation of the corpus serves to have a robust Basque WordNet, which we are confident now that can be used to treat real corpora after the process is finished.

Even if the average agreement rate is in the fringe of well-defined problems, we think that it is acceptable fo a demanding task such as word sense disambiguation, where succinct glosses are used to define all uses of a word. In fact we found that most disagreements where systematic, and very easy to assign to some sense or the other by the referee. After improving WordNet and complementing it with the tagged examples, we are sure that the agreement will be much higher.

We have gone halfway through the process of tagging nouns. After tagging the most frequent nouns with this

methodology we are planning to use a single tagger for the rest of the corpus, assuming that less frequent nouns should be less polysemous and easier to tag.

For the future, we are doing pilot studies for the annotation of the corpus with semantic roles in the style of PropBank (Civit et al., 2005; Agirre et al., 2006c). We are also evaluating the possibility of using coarse grained distinctions, coarser than synsets, for the annotation of the senses in the verbal part of WordNet. In the same sense, we would like to use the agreement information to study the confusability of senses, and the definition of coarser grained senses for nouns (Fellbaum et al., 2001).

6. 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.

7. References

Aduriz I., Aranzabe M., Arriola J., Atutxa A., Díaz de Ilarraza A., Garmendia A. & Oronoz M. (2003). Construction of a Basque Dependency Treebank. In *TLT 2003. Second Workshop on Treebanks and Linguistic Theories*, Vaxjo, Sweden, November 14-15.

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

Agirre E., Aldezabal I., Etxeberria J., Izagirre I., Mendizabal K., Pociello E. & Quintian M. (2005). EUSEMCOR: euskarako corpusa semantikoki etiketatzeko eskuliburua; editatze-, etiketatze- eta epaitze-lanak. Internal report.

Agirre E., Aldezabal I. & Pociello E. (2006a). Lexicalization and multiword expressions in the Basque WordNet. In *Proceedings of Third International WordNet Conference*. Jeju Island (Korea).

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

Agirre E., Aldezabal I., Etxeberria J. & Pociello E. (2006c). A preliminary study for building the Basque PropBank. In Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC). Genoa, Italy.

Agirre E. & Lersundi M. (2001). Extracción de relaciones léxico-semánticas a partir de palabras derivadas usando patrones de definición. In *Proceedings of SEPLN 2001*. Jaén (Spain). 157-165.

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

Atserias J., Climent S., Farreras J., Rigau G. & Rodríguez H. (1997). Combining Multiple Methods for the Automatic Construction of Multilingual WordNets. In *Proceedings of Conference on Recent Advances on NLP. (RANLP'97)*. Tzigov Chark (Bulgaria).

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

Bentivogli L. & Pianta E. (2004). Extending wordnet with syntagmatic information. In *Proceedings of Second Global WordNet Conference*. Brno (Czech Republic).

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

Civit M., Aldezabal I., Pociello E., Taulé M., Aparicio J. & Màrquez L. (2005). 3LB-LEX: léxico verbal con frames sintáctico-semánticos. In *XXI Congreso de la SEPLN*. Granada (Spain).

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

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

Fellbaum C., Palmer M., Dang H., Delfs L. & Wolff S. (2001). Manual and Automatic Semantic Annotation with WordNet. In *NAACL-2001 Workshop on WordNet and Other Lexical Resources*, Pittsburgh PA.

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

Kilgarriff A. (1998). Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. In *Computer Speech and Language. Special Use on Evaluation* 12(4), pp 453-472.

Navarro B., Civit M., Martí M., Marcos R. & Fernández B. (2003). Syntactic, Semantic and Pragmatic Annotation in Cast3LB. In *Computational Linguistics 2003 Workshop on Shallow Procesing of Large Corpora. UCREL Technical Report.* Lancaster (UK).

Bentivogli L. & Pianta E. (2005). Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus. In *Natural Language Engineering*, Volume 11, Issue 03, pp 247-261.

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 ComputationalLinguistics (CICLING 2002)*, pages 1–15. Mexico City (Mexico).

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

- **Dictionaries**

Elhuyar (2000). *Euskal Hiztegi Modernoa*.
Euskaltzaindia (2000). *Hiztegi Batua.*
Sarasola I. (1996). *Euskal Hiztegia*.
UZEI (1987). *Euskalterm.* www.uzei.com/en/euskalterm.htm.