Improving the Basque WordNet by Corpus Annotation

Eneko Agirre and **Izaskun Aldezabal** and **Jone Etxeberria** and **Eli Izagirre** and **Karmele Mendizabal**
Eli Pociello and **Mikel Quintian***
IXA NLP Group
University of the Basque Country
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 though the nominal part of the 300.000 word corpus (roughly equivalent to a 500.000 word corpus for English).

1 Introduction

This paper presents current work on the Basque WordNet. Our team started to build the Basque WordNet following the EuroWordNet design in 2000. The Basque WordNet has been constructed with the expand approach, which means that the English synsets have been enriched with Basque variants. Besides, we also incorporate new synsets that exist for Basque but not for English. We initially linked all Base Concepts manually, and then we generated automatically Basque equivalents using bilingual dictionaries (Atserias et al., 1997). Then we performed a concept-to-concept review where the linguists focused on the correctness of the variants in the synset. 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).

This initial stage allowed building a core WordNet relatively without effort. The stress was on coverage, but we left quality enforcement for later. Regarding the human effort involved, the concept-to-concept review took 1,640 hours at approximately 16 concepts per hour.

We then turned to 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. As the stress was on quality, linguists focused on the correctness and completeness of word senses for a word, and used a number of dictionaries and terminological glossaries (Aulestia & White, 1990; Elhuyar, 1998; Morris 1998; OUP, 1994; Sarasola, 1996; UZEI, 1987 and UZEI, 1999).

This review was half way through when we decided to change our methodology and turn our attention to corpora. Fellbaum et al. (2001) pointed out that dictionaries focus more on word meanings than in the contexts that differentiates those meanings. On the other hand, corpora tell us a lot about how a word is used, but they are not explicit about the meaning of words, unless the corpus is tagged with word sense information.

We therefore decided to exploit the complementary of both kinds of resources, and turned our attention to the coordinated development of the word-to-word review of the Basque WordNet and the manual annotation of a sizeable Basque corpus. This way, we benefit from corpus data to construct, tune and improve the Basque WordNet, and we also produce a manually sense-annotated corpus for Basque (the Basque Semcor).

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 both 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 is indispensable to build word sense disambiguation programs for Basque.

This brief paper is structured as follows. We will first review the methodology used, followed by the figures regarding the current status. Lastly, the conclusions and future work are presented. Note that due to space constraints we have not included comparison to related work.

2 Methodology

Five people, graduate linguistics students, take part in this project: 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. The two taggers independently tag all the examples for the target word, and the referee reviews the disagreements between both taggers and decides which the correct synset is.

The detail of the process is the following: the editor looks up a word in the dictionary, and checks whether all the senses

* Authors listed in alphabetic order.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 287–289.

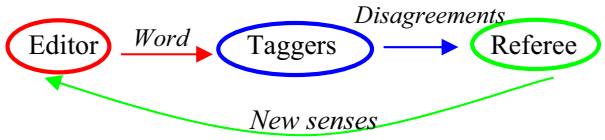
are correctly represented in the Basque WordNet. In this process, he may add new synsets or delete incorrect ones according to a sample of the target corpus and the available monolingual dictionaries[1]. In some way, we can say that the editor is the one who decides the sense inventory of a word. The word to be reviewed by the editor is chosen from a word-list arranged in descending order by their frequency in the corpora. Monosemous words are left aside at this stage.

Once the sense inventory of a word is reviewed, the editor, the two taggers and the referee meet, read the glosses and examples given in the Basque WordNet and discuss the meaning of each synset. They try to agree and clarify the meaning differences among the synsets. The number of senses of a word in the Basque WordNet might change during this meeting; that is, linguists could agree that one of the word's senses was missing, or that a synset did not fit with a word. Then, the editor would update the Basque WordNet according to those decisions before giving the taggers the final synset list.

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 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 of disagreement between the annotators (Navarro et al., 2003).

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

Below we can see the representation of this cyclic process:



The coordination of the whole team is quite complex, and we tried for all the team members to work as synchronized as possible. Incidentally we detected that taggers had some extra time, and decided that they could translate and localize the glosses of target words to Basque.

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

2.1 Special Cases

Some occurrences cannot be tagged with a synset because of some special reasons. We devised a detailed inventory of such cases, which are tagged as **Special Cases (SC)**.

2.1.1 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.

2.1.2 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.

2.1.3 SC3: Word is part of a Multiword Lexical Unit or is a lexicalized inflected form

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').

Another use of this special case is related to inflection. Some words can get a different meaning when they are inflected. In Basque some concepts are expressed in plural. For example, the Basque word *hitza* ('word') needs to be used in plural *hitza**k*** ('word**s**') to express the concept of 'lyrics' ('the text of a song').

2.1.4 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').

2.1.5 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.

2.1.6 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.

2.1.7 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*.

3 Current Data of the Basque WordNet and the Basque Semcor

Table 1 shows the current figures for the Basque WordNet.

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.

Table 1: Current figures for the Basque WordNet.

	TOT	N	V	ADJ	ADV
Word Senses	51423	41833	9450	140	0
Lemmas	25755	22492	3368	50	0
Synsets	31585	27880	3592	113	0
Basque gaps (no lex)	1439	1223	208	8	0
Proper Nouns		680			

At the time of writing the methodology has been going for one year. Up to now, we have only worked with nouns and we have already done 56% of the occurrences (including monosemous nouns and nouns not in WordNet). We estimate that the revision and tagging of the most frequent nouns (accounting for 50% of all the occurrences of polysemous nouns) will take a total of 18 months. At that stage we want to change the methodology and instead of having two taggers plus referee, we plan to use a single tagger per word, except problematic words. With a single tagger we estimate that we will need approximately 12 months to finish all nouns, including the revision of monosemous nouns and nouns not in WordNet.

4 Conclusion and Future Work

We have presented our methodology for the joint development of the Basque WordNet and the Basque Semcor. 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). 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, the use of double tagging for nouns allows for the study of confusability of senses, and the definition of coarser grained senses for nouns (Fellbaum et al. 2001).

Acknowledgements

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 Ministry (HUM2004-21127-E). Eli Pociello has a PhD grant from the Basque Government. The tagging tool was adapted from an implementation of the Universidad de Valencia.

References

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., Aldabe I., Lersundi M., Martínez D., Pociello E., Uria L. (2004) The Basque lexical-sample task. In *Proc of the ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text* (Senseval). Barcelona (Spain).

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).

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).

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.

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).

Dictionaries

Aulestia, G. & White, L. (1990) *English-Basque Dictionary*, University of Nevada Press, Reno.

Elhuyar (1998) *Hiztegi txikia*.

Morris, M. (1998) *Morris Hiztegia*.

OUP (1994) *The Oxford Spanish Dictionary*, Oxford University Press, 1994.

Sarasola, I. 1996. *Euskal Hiztegia*.

UZEI (1987) *Euskalterm.* `http://www.uzei.com/en/euskalter.htm`.

UZEI (1999) *Sinonimoen Hiztegia*.