HAL
open science

A Preliminary Study for Building the Basque PropBank

Eneko E. Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Pociello

To cite this version:

Eneko E. Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Pociello. A Preliminary Study for Building the Basque PropBank. . artxibo-00000094v1

HAL Id: artxibo-00000094

https://artxiker.ccsd.cnrs.fr/artxibo-00000094v1

Submitted on 6 Apr 2006 (v1), last revised 22 Jun 2006 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Preliminary Study for Building the Basque PropBank

Eneko Agirre, Izaskun Aldezabal, Jone Etxeberria, Eli Pociello*

IXA NLP Group
649 pk. 20.080 – Donostia. Basque Country
jibetagj@si.ehu.es

Abstract

This paper presents a methodology for adding a layer of semantic annotation to a syntactically annotated corpus of Basque (EPEC), in terms of semantic roles. The proposal we make here is the combination of three resources: the model used in the PropBank project (Palmer et al., 2005), an in-house database with syntactic/semantic subcategorization frames for Basque verbs (Aldezabal, 2004) and the Basque dependency treebank (Aduriz et al., 2003). In order to validate the methodology and to confirm whether the PropBank model is suitable for Basque and our treebank design, we have built lexical entries and labelled all argument and adjuncts occurring in our treebank for 3 Basque verbs. The result of this study has been very positive, and has produced a methodology adapted to the characteristics of the language and the Basque dependency treebank. Another goal of this study was to study whether semi-automatic tagging was possible. The idea is to present the human taggers a pre-tagged version of the corpus. We have seen that many arguments could be automatically tagged with high precision, given only the verbal entries for the verbs and a handful of examples.

1. Introduction

The study we present here is the continuation of work developed in the Ixa[1] research group, which involved the development of lexical databases and hand-tagged corpora with morphological information (Agirre et al., 1992, Aduriz et al., 1994), as well as syntactic information (Aduriz et al., 1998, Aranzabe et al., 2003). Our group is now moving into semantics, which is essential for many computational tasks such as syntactic disambiguation and language understanding, and applications such as question answering, machine translation and text summarization.

Our previous work on semantics has mainly focused on word senses (including the development of the Basque WordNet and Basque Semcor (Agirre et al., 2006a)), building verbal models from corpora, including selectional preferences (Agirre et al., 2003) and subcategorization frames (Aldezabal et al., 2003), as well as developing by hand a database with syntactic/semantic subcategorization frames for a number of Basque verbs (Aldezabal, 2004).

Our previous experience has persuaded us of the need to model verbal models according to semantic roles. In many cases there is a direct correspondence between an argument, a function and a semantic role, which allows continuing our work on the syntax of Basque into semantics. Our long-term goal is to unify all syntactic and semantic information into a single database coupled with hand-tagged corpora.

Our interest follows the current tendency, as shown by corpus tagging projects such us the PennTrebank2 (Marcus, 1994), PropBank (Palmer et al., 2005) and PDT (Hajic et al., 2003), and the semantic lexicons that have been developed alongside them, like VerbNet (Kingsbury et al., 2002) and Vallex (Hajic et al., 2003). Framenet (Baker et al., 1998) is also an example of the joint development of a semantic lexicon and a hand-tagged.

We chose to follow the PropBank/VerbNet model for a number of reasons:

1. The PropBank project starts from a syntactically annotated corpus, as we do.
2. The organization of the lexicon is similar to our database of verbal models.
3. Given the VerbNet lexicon and the annotations in PropBank many implicit decisions according problematic issues like the distinctions between arguments and adjuncts are settled by example, and seem therefore easy to replicate when we tag the Basque data (see example in Section 3.3)
4. Having corpora in different languages annotated following the same model allows for crosslingual studies, and hopefully enrich Basque verbal models with the richer information currently available for English.

In fact, the PropBank model is being deployed in other languages, such us, Chinese, Spanish, Catalan and Russian. Palmer and Xue (2003) describe the Chinese PropBank. Civit et al. (2005a) describe a joint project[2] to annotate comparable corpora in Spanish, Catalan and Basque. The Russian is in its preliminary stages (Civit et al., 2005b).

This paper presents a methodology for adding a layer of semantic annotation in terms of semantic roles, to an existing corpus of Basque which is syntactically annotated. The proposal we make here is the combination of three resources: the model used in the PropBank project (Palmer et al., 2005), a database with syntactic/semantic subcategorization frames for Basque verbs (Aldezabal, 2004) and the Basque dependency treebank (Aduriz et al., 2003). In order to validate the methodology and to confirm whether the PropBank model is viable for Basque, we have built lexical entries and labelled all arguments and adjuncts occurring in our treebank for 3 Basque verbs

The next section briefly reviews the PropBank model. Section 3 presents our methodology, and Section 4 the analysis of the results. Section 5 presents heuristics to help speed up the semi-automatic tagging. Finally, section 6 presents the conclusions and future work.

2. The PropBank model

In the PropBank model two independent levels are distinguished: the level of arguments and adjuncts, and the

* Authors listed in alphabetical order.

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

[2] http://www.lsi.upc.edu/~mbertran/cess-ece

level of semantic roles. The elements that are regarded as arguments are numbered from *Arg0* to *Arg5*, expressing semantic proximity with respect to the verb. The lowest numbers represent the main functions (subject, object, indirect object, etc.). The adjuncts are tagged as *ArgM*.

With regard to roles, PropBank uses two kinds: roles specific to each concrete verb (e.g. buyer, thing bought, etc.), and general roles (e.g. agent, theme, etc.) linked to the VerbNet lexicon (Kipper et al., 2002). VerbNet is an extensive lexicon where verbs are organized in classes following Levin's classification (1993).

Table 1 shows the PropBank roleset for the verb 'tell.01' and the corresponding VerbNet roleset with Levin's class number (37.1).

PropBank tell.01	VerbNet tell-37.1
Arg0: Speaker	Agent
Arg1: Utterance	Topic
Arg2: Hearer	Recipient

Table 1: PropBank and VerbNet rolesets of the verb 'tell'.

A verb equivalent to the English tell, should have a similar roleset. Table 2 shows a preliminary version for the roleset of the Basque verb *esan.01* (= 'tell') based on the roleset in table 1. VerbNet roles are more general, and to simplify the examples, we will only mention the VerbNet roles in the rest of the paper, together with the argument number.

Esan.01
Arg0: Agent
Arg1: Topic
Arg2: Recipient

Table 2: Preliminary version of the lexical entry for *esan.01*.

To have a general perspective of argument numbers, VerbNet roles and syntactic function, we have elaborated a table, see table 3, that shows the roles and the syntactic functions that are usually associated with the numbered arguments and adjuncts in PropBank:

Arguments	**VerbNet roles**	**Syntactic function**
Arg0	agent, experiencer	subject
Arg1	patient, theme, attribute, extension	direct object, attribute, predicative, passive subject
Arg2	attribute, beneficiary, instrument, extension, final state	attribute, predicative, indirect object, adverbial complement
Arg3	beneficiary, instrument, attribute, cause	predicative, circumstantial complement
Arg4	destination	adverbial complement
Adjuncts		
ArgM	location, extension, destination, cause, time, manner, direction	adverbial complement

Table 3: The roles and the syntactic functions that are usually associated with the numbered arguments and adjuncts in PropBank.

Each roleset has its corresponding frameset showcasing the syntactic realization of the verb. This information is very helpful for tagging the specific sense and syntactic structure of a verb. Figure 1 shows the 4 frames that form the frameset associated with the 'tell.01' roleset. For the sake of brevity, we have only illustrated entirely the first frame, where the example shows how 'tell' is used in a ditransitive structure, and the syntactic realization of each of the arguments is marked.

Roleset tell.01 "pass along information":
Roles:
Arg0:*Speaker*
Arg1:*Utterance*
Arg2:*Hearer*

Frames:
ditransitive (-)
`The score tell you what the characters are thinking and feeling.`
Arg0: The score
REL: tell
Arg2: you
Arg1: what the characters are thinking and feeling
odd ditransitive (-)
prepositional arg2 (-)
fronted (-)

Figure 1: Part of the frameset associated with the 'tell.01' roleset.

3. Methodology

In this section we first present the five steps of our tagging methodology:

1. Choice of verbs
2. Building the preliminary entry for the verb
3. Comparison to equivalent English entries
4. Tagging of the corpus
5. Post-tagging analysis (back to step 2.)

But first we present the corpus to be tagged.

3.1. The corpus

EPEC is the Reference Corpus for the Processing of Basque. This is a corpus of standard written Basque whose aim is to be a reference corpus for the development and improvement of several NLP tools for Basque. It is a 300.000-word sample collection of written standard Basque[3]. Around one third of this collection was obtained from the Statistical Corpus of 20th Century Basque (http://www.euskaracorpusa.net). The rest was sampled from Euskaldunon Egunkaria (http://www.egunero.info) a daily newspaper. EPEC has been manually tagged at different levels (morphosyntax, syntactic phrases, syntactic dependencies and WordNet word senses) and we now want to tag the verb rolesets and semantic roles.

3.2. Choice of Basque verbs

The main criterion to select the verbs has been frequency. 29.95% of all verb occurrences correspond to 10 verbs (from a total of 622 verbs occurring in the corpus). The reason for having such a reduced number of verbs in Basque, compared to others languages such as

[3] Being Basque an agglutinative language, 300.000 word-corpus is roughly equivalent to a 500.000 word-corpus for English.

English, is that Basque lexicalizes less material in the verb, depending more on syntax (Agirre et al., 2006b). For instance, 'to bike' is translated as *bizikletaz ibili* (lit. 'to go on bike') and 'to walk' is translated as *oinez ibili* (lit. 'to go on feet'). In these two verb constructions the adverbial component (*bizikletaz* in the first one and *oinez* in the second one) is syntactically a modifier of the sentence. That is, what in English is expressed with two different verbs, the same verb with another word is needed in Basque.

In this preliminary study we did not want to consider light and modal verbs, which we will be complex to analyze. That is the case of *egin* (='do') and *izan* (='be'), which are the two most frequent verbs in the corpus. We chose 3 other verbs from the top 10 to perform the pilot study, and will showcase the methodology with *esan*. The other two verbs are *eskatu* and *adierazi*. Table 4 shows the 10 most frequent verbs.

6,10%	*egin*	do, make, …
5,95%	*izan*	be, …
4,66%	***esan***	say, tell, call, …
2,62%	***adierazi***	express, …
2,37%	***eskatu***	ask for, …
1,92%	*eman*	give, …
1,69%	*azaldu*	express, appear, …
1,56%	*hartu*	take, …
1,54%	*jo*	play, …
1,54%	*salatu*	denounce, accuse, ...

Table 4: The 10 most frequent verbs in EPEC, which account for nearly 30% of the occurrences of all verbs. The percentage with respect to all verb occurrences is shown, alongside the main English translations.

3.3. The preliminary entry for the Basque verb

The preliminary lexicon for the 3 verbs including senses and subcategorization frames is based on Aldezabal (2004), which includes an in-depth study of 100 verbs. At this step, we adapted this lexicon to the PropBank model, without examining the entries in PropBank for equivalent verbs (which is done in the next step, see Section 3.4 below).

Aldezabal defined a number of syntactic-semantic frames (SSF) for each verb. Each SSF is formed by semantic roles and the declension case that syntactically realizes this role. The SSFs that have the same semantic roles define a verbal coarse-grained sense and are considered syntactic variants of an alternation. Different sets of semantic roles reflect different senses. This is similar to the PropBank model, where each of the syntactic variants (similar to a frame) pertains to a verbal sense (similar to a roleset).

Aldezabal defines a specific inventory of semantic roles, but we only take the senses, number of arguments and corresponding declension cases for each verb. We adopt the role inventory of PropBank, and therefore use the semantic role information of Aldezabal only as auxiliary information to choose the corresponding role in PropBank.

In figure 2 we can see an example of the SSFs for the verb *esan* as given by Aldezabal. It has two senses and the first one has two syntactic variants. The first variant, *esan-1.1* realizes the first argument with the ergative case and the second argument with the absolutive case. The second variant is similar, but realizes the second argument with a completive clause. The second sense has three arguments, realized as ergative, absolutive and dative, in that order. The first sense can be translated as 'tell/say', as in 'Tell him to stop', and the second sense as 'call', as in 'What shall we call him?'

In general, our SSF lexicon displays a tendency to limit the number of arguments in comparison with the PropBank model, i.e. some of the adjuncts in our SSF lexicon would be given as arguments by PropBank. We therefore decided to adapt the preliminary entry in figure 2 to this tendency. Figure 3 shows the adapted entries for *esan*, given in PropBank format as two rolesets, and where we have added one more argument (Arg2) to *esan.01*. This new argument was listed in our SSF lexicon as an adjunct.

esan-1 (= 'tell/say'): Activity (communication) of an entity. Two arguments in two syntactic variants:
esan-1.1: argument1_ERG[4], argument2_ABS[5]
esan-1.2: argument1_ERG, argument2_COMP[6]

esan-2 (= 'call'): Assignment of an attribute. Three arguments in a single syntactic realization:
esan-2: argument1_ERG, argument2_ABS, argument3_DAT[7]

Figure 2: Syntactic-semantic frames for the verb *esan* (='tell/say/call') as given in our SSF lexicon.

Roleset esan.01 "communication activity of an entity":
Roles:
Arg0:*Agent*
Arg1:*Topic*
Arg2:*Recipient*

Frames:
- Direct object: absolutive
  Juan Maria Atutxak hori esan zuen (*Juan Maria Atutxa said that*)
  Arg0: Juan Maria Atutxak (ERG)
  Arg1: hori (ABS)
  REL: esan
  auxmod: zuen
- Direct object: completive clause
  Juan Maria Atutxak bakea nahi zuela esan zion entzulegoari (*Juan Maria Atutxa told the audience that he wanted peace*)
  Arg0: Juan Maria Atutxak (ERG)
  Arg1: bakea nahi zuela (COMP)
  REL: esan
  auxmod: zion
  Arg2: entzulegoari (DAT)

Roleset esan.02 "assignment of an attribute":
Roles:
Arg0:*Agent*
Arg1:*Theme*
Arg2:*Predicate*

Figure 3: The preliminary entry for the verb *esan* in PropBank style. Two rolesets are given, with two frames for the first sense and one frame for the second. In the

[4] ERG = ergative declension case.
[5] ABS = absolutive declension case.
[6] COMP = completive clause.
[7] DAT = dative declension case.

frame examples we specify the case to be used. Note that we give VerbNet role names in the roleset.

3.4. Comparison to equivalent English entries

Hour hypothesis is that, with respect to the Basque verb, entries for the English equivalent senses in PropBank should maintain a similar syntactic-semantic behaviour. The preliminary entry of the verb (see figure 3) is the basis for looking for the English equivalent. In this process the use of bilingual dictionaries for Basque, Spanish and English (Elhuyar[8], WordReference[9]), as well as the monolingual dictionaries for Basque and English (Cambridge[10], OneLook[11]) is essential.

Once we find the English counterparts, we make a selection based in the similarity on their syntactic structure. For example, for the roleset *esan.01* we find that in this sense there are two English equivalents: *tell.01* and *say.01*. However, the roleset of *tell.01* and *say.01* are not totally similar. The roleset of *say.01* has a fourth role (Arg3 Attributive) that the roleset of *tell.01* has not. In Basque, the verb *esan* accepts this Arg3, although its argument or adjunct status is debatable. In Basque we don't have two different verbs for the English *tell* and *say*, so we decided to take the role Arg3-Attributive as an argument for the roleset of *esan.01*[12], specifying in the frame that it can be realized as the instrumental case or a complex declension case such as *–i buruz* ('about'). As a result of this study we modify our entries, as shown briefly in table 5 for *esan.01* and table 6 for *esan.02*.

PropBank *say.01* **VerbNet *say-37.7***	**PropBank *tell.01*** **VerbNet *tell-37.1***	***esan.01***
Arg0: Agent	Arg0: Agent	Arg0: Agent *(ERG)*
Arg1: Topic	Arg1: Topic	Arg1: Topic *(ABS/COMP)*
Arg2: Recipient	Arg2: Recipient	Arg2: Recipient *(DAT)*
Arg3:Attributive		Arg3: Attributive *(INS/-i buruz)*

Table 5: The equivalent entries between Basque and English for the verb *esan.01*. Note that we are showing the brief version of the entries for Basque. Figure 3 shows a full entry.

PropBank *call.01* **VerbNet *dub-29.3***	***esan.02***
Arg0: Agent	Arg0: Agent *(ERG)*
Arg1: Theme	Arg1: Theme *(DAT)*
Arg2: Predicate	Arg2: Predicate *(ABS)*

Table 6: The equivalent entry between Basque and English for the verb *esan.02* (brief version of entries shown).

[8] www.euskadi.net/hizt_el/indice_e.htm
[9] www.wordreference.com
[10] dictionary.cambridge.org
[11] www.onelook.com
[12] It must be noted that in fact it is PropBank who has arg3 for *say*, but VerbNet does not include this argument. It seems that the argument status is also debatable for English.

3.5. First tagging

For this preliminary study we produced one file for each verb, comprising a sample of sentences containing the verb. These sentences are syntactically tagged with dependencies, which is the basis for adding the semantic layer. We are preparing a graphical user interface to help the linguist with the tagging. Figure 4 shows a syntactically annotated clause in Basque, where the semantic annotation has been added. We can see that the clause is divided in phrases and that each phrase has its dependency relation (e.g. *nc_subj* for subject) with respect to the verb. In the description of the phrases the declension case (e.g. *ERG* ergative), the lexical head (*esan*) and the inflected form (i.e. Triasek) is also available. Apart from all the information we have mentioned, the type of subordination (*COMP*: completive) is also given in the phrase description. And finally, the *auxmod* is the auxiliary verb, and which does not have a semantic role.

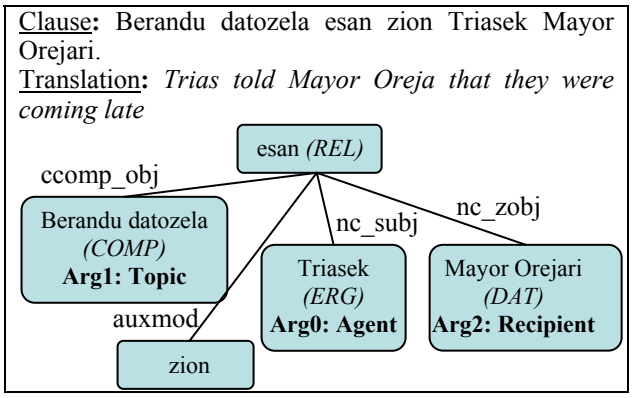

Figure 4: A syntactically and semantically annotated clause in Basque. Syntactic dependencies[13] are marked on the links, and the semantic information in the nodes. Declension case has been included in the nodes as additional information.

The taggers were asked to add the numbered argument (or modifier) and the semantic role to each dependency for the target verb. We instructed the annotators to only tag only those examples which were clear. The uncertain ones are kept apart for further discussion. This way we allow for a consistent tagging with the current lexical entry. The taggers are also asked to look out for occurrences which are not reflected in the different rolesets of the verb, and indications that a sense of the verb might be missing. All un-annotated examples and tagger comments are analyzed in the following step.

3.6. Post-tagging analysis

Considering the tagging problems, in this last step we provide the opportunity to make changes in the lexicon (e.g. add or eliminate arguments from the rolesets, add new senses, add more information to the frames, etc.) or including new criteria for further tagging. For example, in the case of the verb *esan*, 14 out of the 242 occurrences were left aside for further discussion. Most of this

[13] *ccomp_obj* is the completive clause object; the *auxmod* is the auxiliary verb; *nc_subj* is the non-clause subject; and *nc_zobj* is the non-clause second object.

uncertain cases have to do with the possible Basque syntactic realizations of the arg3, which are often complex declension cases (e.g. '*-i erreferentzia eginez*' which means more or less 'regarding to'). We therefore decided to include the complex postposition '*–i erreferentzia eginez*' in one of the frames for *esan.01* for arg3 with the attributive role. See table 7 for the definitive entry for *esan*.

Updating the lexicon might involve tagging the occurrences of the target verb again (back to previous step), as it is the case for *esan*, but it also can ask for general book-keeping. In the case of '*–i buruz*' or '*–i erreferentzia eginez*' for instance, we need to make sure that other occurrences of the complex postposition in the lexicon and corpus are coherent with the new semantic role.

esan.01	***esan.02***
Arg0: Agent *(ERG)* Arg1: Topic *(ABS/COMP)* Arg2: Recipient *(DAT)* Arg3: Attributive *(INS/-i buruz/ -i erreferentzia eginez/...)*	Arg0: Agent *(ERG)* Arg1: Theme *(DAT)* Arg2: Predicate *(ABS)*

Table 7: The definitive entry for the verb *esan* (brief version shown).

4. Analysis of the results

From this pilot study we have seen that the PropBank model is perfectly fit to properly treat Basque verbs. We have seen that the tagging can proceed smoothly from the dependency Treebank. In contrast to the English PropBank methodology, in addition to the syntactic functions in the Treebank we also use the information regarding case suffixes and postpositions (simple and complex), and specify this information in the frames.

Our database of verbal models has been a good starting point to produce the preliminary lexicon for the verbs. We have detected some differences with English verbs regarding the status of arguments and adjuncts, due to different basic criteria, but those can be easily adjusted. Our database is stricter on arguments, while PropBank has a wider perspective. In the 3 verbs under scrutiny it has not been a problem, because there were no interferences regarding the senses. We want to note that we worked on three of the top-five from the most frequent verbs of our corpus, which are usually the most problematic, which further stresses the validity of our approach.

Another goal of this study was to study whether semi-automatic tagging was possible. The idea is to present the human taggers a pre-tagged version of the corpus. We have seen that many arguments could be automatically tagged with high precision, given only the verbal entries for the verbs and a handful of examples. The heuristics can be general (for all verbs) or specific for each individual verb). We will explain them in more detail in the next section.

5. Semiautomatic tagging

The idea of the semiautomatic tagging is to speed up the process of tagging using hand-made heuristics based on manual analysis of the corpora. At this stage we have only analyzed three verbs, but some regularities are already arising. We will continue to study and tag more verbs in order to improve, add, or cancel the heuristics.

Some of the heuristics would apply to all verbs. For instance, the ergative case gets Arg0 in for all instances of the studied verbs. The absolutive case is more variable, as shown by the two rolesets of *esan*, where it gets Arg1 for *esan.01*, and Arg2 for *esan.02*. We therefore plan to tag all ergatives as Arg0.

The other kind of heuristics is specific for each verb, and can be derived automatically from the preliminary entries as built in Section 3.4. We will detail here the case for the three studied verbs.

In the case of the verb *esan*, it has two competing rolesets (see table 7). If we list the possible interpretations for each declension case (see table 8), we can see that the completive and the instrumental (including several other complex declension cases) are unambiguous, as they mark the sense (01) and the role (Arg1 and Arg3, respectively). Using this information, we can automatically tag the occurrences of the verb *esan* and its constituents. Those occurrences with a COMP or INS constituent, can be fully disambiguated (sense and roles), while the rest will get ambiguous tags for the sense and roles. For current occurrences of *esan*, this implies that 80% of the occurrences are disambiguated by COMP and 3% by INS (or other complex declension case). Together 82% of the occurrences will be fully disambiguated and tagged, and only 18% will be left ambiguous for the human tagger to disambiguate.

declension case	**Roles**	**Sense of *esan***
ERG	Arg0: Agent	01/02
ABS	Arg1: Topic / Arg2: Predicate	01/02
COMP	Arg1: Topic	01
DAT	Arg2: Recipient / Arg1: Theme	01/02
INS/*-i buruz…*	Arg3: Attributive	01

Table 8: The ambiguous and unambiguous declension cases of the verb *esan*.

The other two verbs have a single roleset, and none of the declension cases is ambiguous regarding the role. The occurrences of these verbs can be tagged unambiguously.

The accuracy of these verb-to-verb heuristics is close to 100%. This is not totally representative until we apply the heuristic to unseen data, but we don't expect much deviation. Still, we want to stress that the automatic tagging is no substitute for the manual tagging. We plan to review all occurrences, regardless whether they are left ambiguous or not. The automatic tagging does not consider the adjuncts yet, but a table of possible interpretations for each adjunct would help the work of the tagger.

6. Conclusions and future work

Our study confirms that building a lexicon and tagging a Basque corpus with verbal sense and semantic role information following the VerbNet/PropBank model of PropBank is feasible. We have also shown the method to integrate our pre-existing resources (Basque dependency treebank and a database with syntactic/semantic subcategorization frames) into this new framework. We have extended the representation of the entries with information about the declension cases that realize the arguments.

The methodology drawn here is being structured in a detailed set of guidelines for taggers and lexicon editors. Both teams work together in order to build a coherent lexicon and tagged corpus. We plan to update the guidelines as we continue to work on more verbs.

The analysis of the results has also shown that the preliminary entries can be used to tag automatically the verbal senses and roles with promising accuracy and disambiguation rates.

As future work, we are now starting to extend the tagging to the 300.000-word EPEC corpus, which is being entirely hand-annotated with morphologic and syntactic tags, as well as Basque WordNet word senses.

We are also interested in using the EuroWordNet model as a pivot to link the framesets across different languages. A similar study was performed in (Lersundi, 2005) for prepositions, and we plan to extend it to verbs. Comparable corpora is being tagged for Spanish and Catalan (Civit et al., 2005a), which will allow for further crosslingual studies.

7. Acknowledgments

The work has been partially funded by the Education Department of the Spanish Government (CESS-ECE project, HUM2004-21127-E). Eli Pociello has a PhD grant from the Basque Government.

References

Aduriz, I., Alegria, I., Arriola, J., Artola, X., Zubillaga, X., Díaz de Ilarraza, A., Ezeiza, N. (1994). EUSLEM: un lematizador/etiquetador de textos en euskera. *Actas del X congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN)*. Grupo: 6 Análisis de corpus. Córdoba.

Aduriz, I., Agirre, E., Aldezabal, I., Alegria, I., Ansa, O., Arregi, X., Arriola, J., Artola, X., Zubillaga, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M., Oronoz, M., Sarasola, K., Soroa, A., Urizar, R. (1998) A framework for the automatic processing of Basque. In *Proceedings of Workshop on Lexical Resources for Minority Languages.* Granada.

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., Alegria, I., Arregi, X., Artola, X., Zubillaga, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K. (1992). Xuxen: A Spelling Checker/Corrector for Basque based in Two-Level Morphology. In *Proceedings of ANLP'92*, 119-125. Povo Trento.

Agirre E., Aldezabal I., Pociello E. (2003). A pilot study of English Selectional Preferences and their Cross-Lingual Compatibility with Basque. *International Conference on Text Speech and Dialogue*. Czech Republic, pp. 12-19.

Agirre, E., Aldezabal, I., Etxeberria, J., Izagirre, I., Mendizábal, K., Pociello, E., Quintian, M. (2006a). A methodology for the joint development of the Basque WordNet and Semcor. In *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*. Genoa, Italy.

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

Aldezabal, I., Aranzabe, M., Atutxa, A., Gojenola, K., Oronoz, M., Sarasola, K. (2003). Application of finite-state transducers to the acquisition of verb subcategorization information. *Natural Language Engineering*, Volume 9, pp 39-48, ISSN: 1351-3249. Cambridge University Press.

Aldezabal, I. (2004). Aditz-azpikategorizazioaren azterketa. 100 aditzen azterketa zehatza, Levin (1993) oinarri harturik eta metodo automatikoak baliatuz. Doktorego txostena. *Euskal Filologia saila*. Leioa.

Aranzabe, M., Arriola, J., Atutxa, A., Balza, I., Uria, L. (2003). Guía para la anotación sintáctica manual de Eus3LB (corpus del euskera anotado a nivel sintáctico, semántico y pragmático). *UPV/EHU/LSI/TR 13-2003*.

Baker, C.F., Fillmore, C.J., Lowe, J.B. (1998). The Berkeley FrameNet project. In *Proceedings of the COLING-ACL*. Montreal, Canada.

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

Civit, M., Castelvi j., Morante, R., Oliver, A., Aparicio, J. (2005b). 4LEX: a Multilingual Lexical Resource. In *Croos-Language Knowledge Induction Workshop*. EuroLAN Summer School. Babes,-Bolyai University, Cluj-Napoca, Romania.

Haji˘c, J., Panevová, J., Urešová, Z., Bémová, A., Kolá˘rová, V., Pajas, P. (2003). PDT-VALLEX: Creating a Largecoverage Valency Lexicon for Treebank Annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*. Sweden, pp. 57–68.

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.

Kipper, K., Palmer, M., Rambow, O. (2002). Extending PropBank with VerbNet Semantic Predicates. In *Workshop on Applied Interlinguas, held in conjunction with AMTA-2002.* Tiburon, CA.

Lersundi, M. (2005). Ezagutza-base lexikala eraikitzeko Euskal Hiztegiko definizioen azterketa sintaktiko-semantikoa. Hitzen arteko erlazio lexiko-semantikoak: definizio-patroiak, eratorpena eta postposizioak. *Euskal Filologia Saila.* Leioa.

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

Marcus, M. (1994). The Penn TreeBank: A revised corpus design for extracting predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*. Princeton, NJ.

Palmer, M., Xue, N. (2003). Annotating the Propositions in the Penn Chinese Treebank. In *Proceedings of the Second Sighan Workshop*, Sapporo, Japan.

Palmer, M., Gildea, D., Kingsbury, P. (2005). The Proposition Bank: A Corpus Annotated with Semantic Roles. In *Computational Linguistics Journal*. 31:1.