EusPropBank: Integrating Semantic Information in the Basque Dependency Treebank

Izaskun Aldezabal*, María Jesús Aranzabe*, Arantza Díaz de Ilarraza**, Ainara Estarrona** and Larraitz Uria ***
IXA NLP Group
*Basque Philology Department, **Languages and Information Systems
University of the Basque Country
*** IKER UMR 5478, University of Pau and Pays de l'Adour (UPPA), CNRS
e-mail:
{izaskun.aldezabal,maxux.aranzabe,a.diazdeilarraza,ainara.estarrona,larraitz.uria}@ehu.es

Abstract. This paper deals with theoretical problems found in the work that is being carried out for annotating semantic roles in the Basque Dependency Treebank (BDT). We will present the resources used and the way the annotation is being done. Following the model proposed in the PropBank project, we will show the problems found in the annotation process and decisions we have taken. The representation of the semantic tag has been established and detailed guidelines for the annotation process have been defined, although it is a task that needs continuous updating. Besides, we have adapted AbarHitz, a tool used in the construction of the BDT, to this task.

Key Words: Theoretical Problems in Semantic Annotation, Representation of Semantic Roles, Lexical Resources

1. Introduction

The construction of a corpus with annotation of semantic roles is an important resource for the development of advanced tools and applications such as machine translation, language learning and text summarization. We present here the work that is been carried out for annotating semantic roles in the BDT. 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 manually developing a database with syntactic/semantic subcategorization frames for a number of Basque verbs (Aldezabal, 2004).

Our interest follows the current trend, as shown by corpus tagging projects such us the Penn Treebank (Marcus, 1994), PropBank (Palmer et al., 2005) and PDT (Hajic et al., 2003), and the semantic lexicons that have been developed alongside them, suche as 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 corpus.

After a preliminary study, we chose to follow the PropBank/VerbNet model for a number of reasons:

- The PropBank project starts from a syntactically annotated corpus, just as we do.
- The organization of the lexicon is similar to our database of verbal models.
- Given the VerbNet lexicon and the annotations in PropBank, many implicit decisions on problematic issues, such as the distinctions between arguments and adjuncts have been settled and are therefore easy to replicate when we tag the Basque data.
- Having corpora in different languages annotated following the same model allows for cross-lingual studies and hopefully the enriching of Basque verbal models with the richer information currently available for English.

In fact, the PropBank model is being deployed in other languages, such as Chinese, Spanish, Catalan and Russian. Palmer and Xue (2003) and Nianwen (2008) describe the Chinese PropBank. Civit et al. (2005) describe a joint project to annotate comparable corpora in Spanish, Catalan and Basque.

The paper will be organised as follows: after a brief introduction, we will present the resources used in the semantic tagging. Section 3 explains the steps followed in the annotation, the automatic procedures defined to facilitate the task of manual annotation. In section 4, we describe the tool used for tagging (AbarHitz) while section 5 discusses theoretical problems and decisions we are facing. Finally, section 6 presents the conclusions and future work.

2. The Resources used

In this section we will present the PropBank/VerbNet model, the model followed, and the resources we have for the annotation of semantic roles. We will explain them briefly, more details can be found in Aldezabal (2007) and Agirre et al. (2006b).

2.1. PropBank/VerbNet

PropBank is a corpus that is annotated with verbal propositions and their arguments. In the PropBank model two independent levels are distinguished: the level of arguments and adjuncts, and the 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 specific 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). The lexicon provides an association between the syntactic and semantic properties of each of the described verbs.

Table 1 shows the PropBank roleset for the verb 'go.01' and the corresponding VerbNet roleset with Levin's class number (go-47.7 51.1-2).

Table 1: PropBank and VerbNet rolesets of the verb ‘go’.

PropBank go.01	VerbNet go-47.7 51.1-2
Arg1: entity in motion/goer	Theme
Arg2: extent	
Arg3: start point	Source
Arg4: end point	Destination
ArgM: medium	
ArgM: direction (usually up or down)	

A verb equivalent to the English *go* should have a similar roleset. Table 2 shows a preliminary version for the roleset of the Basque verb *joan.01* (= ‘go’) based on the roleset in table 1. VerbNet roles are more general and sometimes, as the examples show, more simple. As a first approach, we decided to use the VerbNet1.0 roles (and when the tagging task required we would add the missing ones) because it is more similar to our in-house database. We will only mention the VerbNet roles in the rest of the paper, together with the argument number.

Table 2: Preliminary version of the lexical entry for *joan.01* (=‘go’).

joan.01
Arg1: Theme
Arg3: Source
Arg4: Destination

Table 3 shows the argument numbers, the VerbNet roles and the syntactic functions which are usually associated with the numbered arguments and adjuncts in PropBank:

Table 3: The argument numbers, the roles and the syntactic functions 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

We have gathered the information contained in PropBank and VerbNet (VerbNet 1.0) in a single data base. The information contained in this data base is used when applying the automatic procedure.

2.2. The BDT Corpus

For our task we will use the Basque Dependency Treebank (BDT). The Basque Dependency Treebank was built on EPEC, a corpus that contains 300,000 words of standard written texts which is intended to be a training corpus for the development and improvement of several NLP tools (Bengoetxea and Gojenola, 2007). 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 (BDT) and WordNet word senses.

2.3. The EADB Resource (Data Base for Basque Verbs)

The work done in Aldezabal (2004), which includes an in-depth study of 100 verbs for Basque from EPEC, is our starting point. 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 performs this role. The SSFs that have the same semantic roles define a coarse-grained verbal 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 defined a specific inventory of semantic roles; the set of semantic roles associated with a verb identifies the different meanings of that verb. The semantic roles specified are: Theme, Affected Theme, Created Theme, State, Location, Time, End Location, End State, Start Location, Path, Startpoint, Endpoint, Experiencer, Cause, Source, Container, Content, Feature, Activity, Measure, Manner. In addition, Aldezabal identified a detailed set of types of general predicates to facilitate the classification of verbs from a broad perspective in such a way that the meaning of the verbs is expressed from a cognitive point of view. The predicates are the following: Change of State of an Entity, Change of Location of an Entity, Change of an Entity, Creation of an Entity, Activity of an Entity, Interchange of an Entity, To contain an Entity, Assignment of a Feature to an Entity, Existence of an Entity, Location of an Entity, State of an Entity, Description of an Entity, Expression of a Suppisition.

We show an example of an EADB verb entry:

joan.1 (‘go’): entity in motion
affected theme_ABS[1]; startpoint / path_ABL; endpoint_ALA
joan.2 (‘go’): entity in motion
affected theme_ABS; startpoint [+animate]_DAT; endpoint_ALA
joan.3 (‘go’): feature that disappears from an entity
container_DAT; content [-animate, -concrate]_ABS

[1] ABS, ABL, ALA and DAT are the absolutive, ablative, adlative and dative cases respectively.

2.4. Mapping between Basque and English Verbs based on Levin's classification

In Aldezabal (1998), English and Basque verbs are compared based on Levin's alternations and classification. For this purpose, all of the verbs in Levin (1993) were translated first considering the semantic class and then paying attention to the similarity of the syntactic structure of verbs in English and Basque. The main advantage of having linked the Basque verbs to Levin classes comes from the fact that other resources like PropBank and VerbNet lexicon are linked to Levin classes and contain information about semantic roles. Verbs in a Levin class have a regular behaviour (according to diathesis alternation criteria), different from verbs belonging to other classes. Also de classes are semantically coherent and verbs belonging to one class share the same semantic roles. In Table 4, we present some examples of these links.

Table 4: the link between verbs in Levin (1993) and Basque.

glower	40.2	bekozko/kopetilun begiratu
glue	22.4	erantsi, kolatu
gnash	40.3.2	hortzak karraskatu
go	47.7	joan
go	51.1	joan
gobble	38	glu-glu egin
gobble	39.3	irentsi
goggle	30.3	liluratu moduan begiratu
gondola	51.4.1	gondolaz ibili/joan/eraman

3. The Annotation Process

When constructing BDT, we followed a Dependency Parsing Syntactic Formalism which provides a straight forward way for expressing semantic relation. The process of manual annotation of semantic roles associated to verbs will begin with the tagging of the most frequent verbs contained in the corpus (approximately 30% of all verb occurrences correspond to 10 verbs) and studied in (Aldezabal, 2004). The sentences of the corpus are grouped according to the verbs they have.

We don't annotate light and modal verbs that will be treated deeply later. That is the case of *egin* (='do') and *izan* (='be'), which are the two most frequent verbs in the corpus.

Once we finish the 100 verbs, we will continue with the rest of verbs, in the way we will explain in the methodology.

We carry out this work by means of the following phases:

1. The preprocessing phase: comparison of the Levin classes in our mapping and the PropBank data-base. As explained before, we have the English equivalent of a Basque verb in terms of Levin class so we were able to obtain automatically the PropBank/VerbNet information for each treated verb from the paid data-base, basing on Levin class.

However, we have to update our mappings since our mapping was done, some time ago, PropBank has changed and, consequently, new classes and subclasses have

been added, erased and modified. We performed an automatic revision of our previous mappings and distinguished the four different situations, explained below:

- **equal**: represents the case in which the identification of the class for a verb has not changed since the mapping was done. For instance, *say* and *go* continue being in the 37.7 and 47.7 classes respectively. This option represents 51% of the cases.
- **subclass**: a new subclass has been defined in PropBank. For example, the verb *go* in the 51.1 class in our mapping has been redefined as 51.1-2 in PropBank. In these cases, we directly equalized the subclass with the general class, and maintain the mapping. (6%)
- **changed**: a Levin class in PropBank has changed and there is not a direct coincidence between our mapping and the one in PropBank. For instance, the class 45.6 for the verb *increase* has been changed in PropBank (2%)
- **missing:** the verb is not included in PropBank or it has not assigned any Levin class. For instance, the verb goggle is not in PropBank (41%)

In Table 5 we present the result of this automatic comparison for some of the verbs contained in Table 4. The first column in Table 5 shows the English verb, the second column corresponds to Levin's class, the third column presents the definition of the verb in Basque and the fourth one specifies to which group the mapping belongs.

Table 5: A sample of the results of the comparison between our mapping and PropBank, regarding Levin classes.

glower	40.2	bekozko/kopetilun begiratu	MISSING
glue	22.4	erantsi, kolatu	EQUAL
glutenize	45.4		MISSING
gnash	40.3.2	hortzak karraskatu	MISSING
gnaw	39.2		MISSING
go	47.7	joan	EQUAL
go	51.1	joan	SUBCLASS
gobble	38	glu-glu egin	EQUAL
gobble	39.3	irentsi	EQUAL
goggle	30.3	liluratu moduan begiratu	MISSING
gondola	51.4.1	gondolaz ibili/joan/eraman	MISSING

We decided to deal with the first and second cases (those verbs detected as "equal" and "subclass") that cover the 46% of the EPEC corpus, leaving the rest to future study. We are refining our algorithm to see if it is possible to detect automatically more equivalences..

2. Establishing the tagging criteria. Three linguists tag 50 occurrences of the same verb for each of the verbs fixed in the first step. This step has the objective of obtaining the guidelines for the annotation.

3. Semiautomatic tagging. Again, three linguists tag 20 different occurrences of the same verb (60 occurrences in all). Once (at least) 60 occurrences of these verbs are tagged we begin with the rest of occurrences by means of automatic procedures. Throughout the process the guidelines are updated.

For the rest of the verbs, we will prepare an automatic pre-tagging process based on lexical models obtained from the tagged corpus. Features such as Verb, VNrol,

Valence and Selectional Restriction will be taken into account. In Aldezabal (2001) and Zapirain et al. (2008), we have carried out some experiments in which different methods for role inference are proposed for English verbs.

3.1. Representation of the Semantic Information (Definition of the Tag)

From the set of dependency relations associated to a clause, we will take those relations that are candidates to be arguments or adjuncts of the verb[2] We denominate the semantic tag defined "arg_info" and it is composed by the following fields (explained in the order of appearence):

- **VN** (VerbNet/PropBank verb): the English verb and its PropBank number in "VerbNet-PropBank". As it is usual to find more than one verb in the same category, we put the necessary ones separated by the slash. Example: tell_01 / say_01.
- **V** (Verb): the main verb which acts as the head of the relation.
- **Treated Element** (TE): the element depending from the head that will be the adjunct or the argument.
- **VAL** (valence): value that identifies arguments or adjuncts: arg0, arg1, arg2, arg3, arg4, argmod.
- **VNrol** (role in VerbNet): those represented in Table3.
- **EADBrol** (semantic role according to EAD roleset). We can see an enumeration of them in Table 4.
- **HM** (Selectional Restriction). Up to now we only consider [+animate], [-animate], [+count], [- count], [+hum], [-hum]

Figure 1 shows a compound sentence syntactically annotated, where a semantic annotation has been added to the phrase in adlative (ALA) linked to the verb *joan*. We can see that the sentence is divided into phrases and that each phrase has a dependency relation (e.g. ncmod for prepositional phrase) with respect to the verb (joan). Syntactic dependencies[3] are marked on the links, and the semantic information in the nodes. Declension case has been included in the nodes as additional information.

---

[2] The relations considered are: ncsubj, ncobj, nczobj, ncmod, ncpred (non-clausal subject, object, indirect object, …), ccomp_obj, ccomp_subj, cmod (clausal finite object, subject, modifier), xcomp_obj, xcomp_subj, xcomp_zobj, xmod, xpred (clausal non-finite object, subject, indirect object, …).

[3] *cmod* is the relative clause; *auxmod* is the auxiliary verb; *ncsubj* is the noun-clause subject; and *postos* is an auxiliary tag to express a complex postposition.

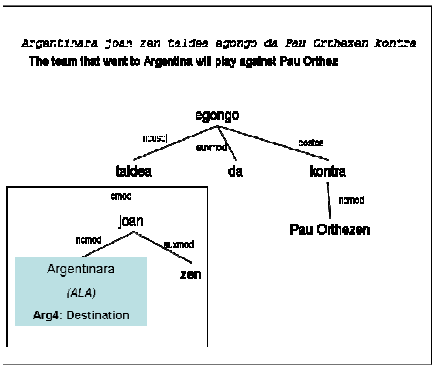

Figure 1: A syntactically and semantically annotated clause in Basque.

The example (1) illustrates the arg_info tag that corresponds to the relation highlighted in Figure 1.

(1) arg_info: (go_01, joan, *Argentinara*[4], Arg4, Destination, end_location, -[5].)

4. AbarHitz, the tool for tagging

AbarHitz (Díaz de Ilarraza et al., 2004) is a tool designed to help the linguists in the manual annotation process of the BDT. AbarHitz has been implemented to assist during the definition of dependencies among the words of the sentence.

Similar tools have been implemented with the same aim as the AbarHitz; Annotation Graph Toolkit (AGTK) (Bird et al., 2002), TREPIL Treebanking Interface (Rosén et al., 2005) are some examples. It is important to emphasize that the design of Abar-Hitz follows the general annotation schema we established for representing linguistic information and it is part of a general environment we have developed so far in which general processors and resources have been integrated.

Let us first of all describe the tool in general terms and then we will explain how it is appropriate for the semantic annotation presented here.

Abar-Hitz communicates with the user by means of a friendly interface providing the following facilities:

(1) It visualizes the morphosyntactic information obtained so far and which, for our specific corpus, have previously been manually disambiguated. The tool is able to simultaneously use outputs from several tools (a morphological parser, a POS tagger and a syntactic parser) to guide the annotator's decisions.

(2) It graphically visualizes the dependency-tree for each sentence. In addition, the tree drawn can be graphically manipulated in such a way that the user can change the tags and their fields, roll up sub-trees, remove/add nodes, remove/add connectors (dependencies) and so on.

[4] to Argentina (PP)

[5] When we are not sure of a value or we think it is not necessary to define it, we put the null mark ("-").

(3) It provides an environment for syntactic checking while tagging. We have to take into account that mistakes can be made while tagging in the number and type of slots, and the name of the tag itself. Abar-Hitz keeps away from these mistakes by showing specific pop-up menus where the only thing the linguist can do is to select the appropriate tag.

Figure 2 shows the main window of Abar-Hitz in which we can identify:

- **sentence selection area** (in the right side of the figure). In the top part the linguist specifies the verb; in the example the verb joan (to go) has been selected. Below the specification area, a list of the files containing the selected verb is given. The annotator can select one of the files to proceed with the annotation. At the side, the system also maintains a record of the status of the annotation process indicating for each sentence whether: i) the annotation has been completed or not; ii) the annotation sentence is not clear enough and some aspects must be discussed, and so on.
- **text area** (upper left). When the annotator clicks on one the files listed, the sentence is shown in the upper part of the window highlighted.
- **tagging area (left side)**. The tree visualizer is activated by clicking on the corresponding icon.

4.1. Adapting AbarHitz to the tagging of semantic roles

A recent enhancement of AbarHitz facilitates the semantic annotation by offering the linguist new options:

(1) It provides the information associated with the verb being tagged, contained in PropBank and VerbNet. Figure 2 shows an example of this functionality, which is made explicit in two ways: i) by displaying in the right part of the window information from PropBank/VerbNet; and ii) by giving the corresponding information in the arg_info relation as seen in section 3.4.

(2) It provides new “incomplete” “arg_info” relations to be fulfilled by the annotator. We say “incomplete” because some of the arguments of the relation have been automatically obtained while others remain unspecified. Although the system doesn’t provide all the “arg_info” relation complete, the approach has been proved to be very helpful to the linguists. Figure 3 shows, on the left side, the syntactic annotation of the sentence and the semantic tag “arg_info” associated to the verb under study (*joan*) fulfilled by the annotator.

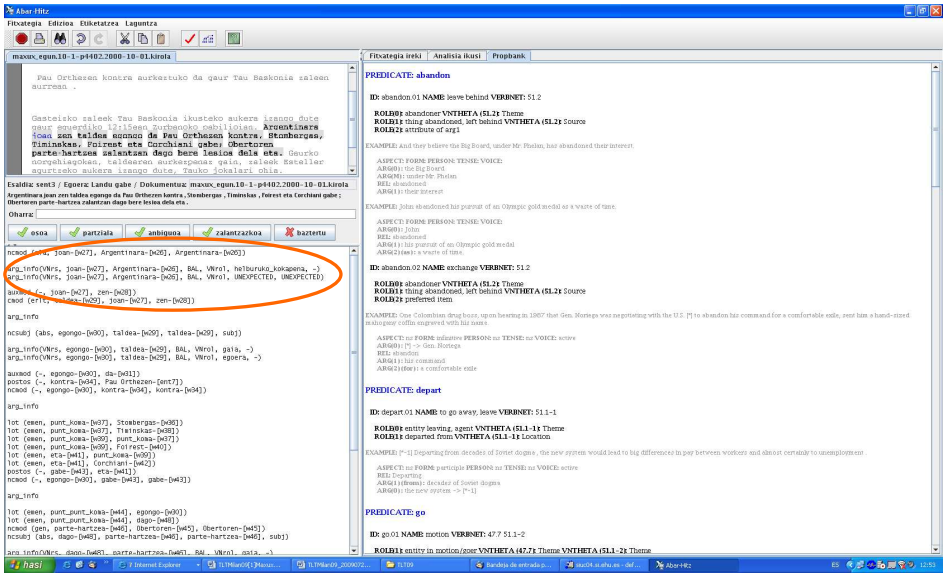

Figure 2: Visualizing the information of PropBank/VerbNet (right side) to the human annotator. On left side arg_info tag proposed to be fullfilled by the annotator

Abar-Hitz has been developed in Java; it follows a modular design in order to be a portable and easily maintainable tool. It runs under the Microsoft Windows, Linux and Unix environments.

5. Theoretical Problems and Decisions

We tagged about 37,000 words of the corpus and analyzed 32 verbs (27% of the overall corpus). We consider for tagging only some of the most frequent verbs (those which appeared in the EADB). We confirmed that the most ambiguous a verb, the more problems and criteria have to be defined.

Then, we have defined general criteria for the tagging process. Structured and detailed set of guidelines for taggers and lexicon editors have been defined (Aldezabal et al., 2010). However, it is a task that needs continuous updating, as new verbs are analyzed.

Let us mention some of the problems defined and decisions taken during this process:

- When the correspondence to the PropBank model(s) can be established automatically, it happens that this association is not always complete and consistent. A (Basque) verb can be linked to more than one PropBank verb. In such cases, we have to check, first of all, whether the rolset-number, the role and the arguments in both languages are the same or not.

In case they are equivalent, there is no doubt for tagging: we assign the corresponding verb. For example, the verb *esan* can be linked unquestionably with tell_01 and say_01. We establish the correspondence and we indicate this double equivalence by the expression tell_01/say_01 as first value of arg_info tag. If, on the

contrary, the roles and arguments are not the same, we specify the two verbs in the first field (for example: take_04/bring_01) and select the most suitable argument structure one after examining syntactic behavior of both English and Basque verbs.

- When the correspondence to the PropBank model(s) can not be established automatically, we try to find the information in other sources (Verb-Index http://verbs.colorado.edu/verb-index/index.php), make the corresponding inference about its argument structure and roleset and update our databases.

The following example illustrates this problem: the verb *jokatu* ("to bet") is not linked because our algorithm has not established *jokatu* as an equivalent of "to bet". In this case, the steps followed will be:

1. To get the argument-structure of "to bet" in PropBank
   Roleset id: bet.01 , *wager*, vncls: 54.5 94
   Roles:
   Arg0: *better*
   Arg1: *amount of bet*
   Arg2: *basis, proposition, bet on*
   Arg3: *co-better*
2. To look at Verb-Index we can see "to bill", "to rely" and "to risk" have similar behavior
3. To look at the roles of the appropriate one, in this case, "to bill"
   Agent: [+animate / + organization]
   Asset: [+currency]
   Recipient: [+animate / +organization]
   Cause:
4. To make the corresponding inference linking argument and role
   Arg0: Agent
   Arg1: Asset
   Arg2: theme
   Arg3: recipient

Another example to illustrate the difficulty in finding the adequate correspondence can be seen when studying the Basque verb *eskatu* (= "to ask"), we find that none of the equivalents given by the system correspond to the sense we are looking for. In this case, the argument structure of the English verb doesn't agree with the one included in EADB, so, we have to specify a new sense in the EADB data-base. In the case of the verb *eskatu* (= "to ask"), *ask_02* could the appropriate equivalent but its argument structure does not match with the one specified in EADB. The verb *ask_02* in PropBank and VerbNet, contains 3 arguments: Arg0: Agent, Arg1: Theme (proposition) and Arg2: Patient.

However, the verb "eskatu" contains only 2 arguments in EADB: Arg0: *esperimentatzailea* (experiencer) and Arg1: *gaia* (theme). Besides, it is said that the DAT (dative) argument is optional although it is not included within the subcategorized cases (this argument fits with Arg2: Patient in PropBank).

We decide to follow the PropBank model and change our data base. Example (2) shows a sentence that illustrates the final annotation linked to the argument structure of *eskatu.*

Example (2):
Nemesiok, joan baino lehen, Alejandro adiskideari eskatzen dio, zaindu dezala bere "x" zakurra

(Before leaving, Nemesio asks his friend Alejandro to look after his "x" dog)

arg_info (ask_02, eskatzen, Nemesiok, arg0, Agent, ...)
arg_info (ask_02, eskatzen, lehen, argM, TMP, -, -)
arg_info (ask_02, eskatzen, adiskideari, arg2, patient, ...)
arg_info (ask_02, eskatzen, zaindu, arg1, Theme, gaia, -biz.)

We do not follow the same procedure in all cases. For example, in the case of the verb lortu ("to obtain"), the Arg2 definition of PropBank for DAT cases , will be tagged as ArgM.

- Where the value of an item of the relation is not clear or when it has not any corresponding value, we use the symbol "-".
- We do not tag verbs as part of locutions. For example we will leave the tagging process of the roles linked to the verb *joan*[6] in the expressions, *usotara doa*[7], *desarmea aurrera badoa*[8] to a subsequent step
- When VerbNet assings two different roles to the same argument, we have decided to base on EADB and to assign the corresponding roles of VerbNet roles. For example, we have found it in the case of the verb *ikusi* ("to see"). In EADB the verb *ikusi* contains two arguments and a role is assigned to each of the arguments:

  Arg0: *esperimentatzailea* (experiencer)
  Arg1: *gaia* (theme)

In PropBank/VerbNetThat assigns two roles to those arguments: Arg0 has associated "agent" and "experiencier" roles and Arg1, "theme" and "stimulus". In this ambiguous case, we use EADB information. The result would be:

Arg0: Agent, *esperimentatzailea*
Arg1: theme, *gaia*

6. Conclusions

We have presented the work being carried out on the annotation of semantic roles in the BDT, a dependency-based annotated Treebank. Some automatic and manual procedures have been developed in order to facilitate the annotation process. The idea is to present the human taggers with a pre-tagged version of the corpus.

From what we have analyzed up to now, we conclude that the PropBank model is suitable for treating Basque verbs, but, of course, cross-linguistic studies always have to cope with to difficult tasks when performing semantic mapping between verbs in different languages.

Structured and detailed set of guidelines for taggers and lexicon editors have been defined. However, it is a task that needs continuous updating.

Our database of verbal models was a good starting point for the tagging task. We 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.

---

[6] In general "to go"
[7] to go to hunt pigeons
[8] If disarmament goes on

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 but not lacking in problems. We have also shown the method for integrating our pre-existing resources into this new framework

In the future we want to focus on the application of automatic methods for role tagging. We have seen that once a verb is tagged with a certain number of appearances, the resulting lexicon can be used to automatically tag the rest of the appearances. Previous experimentation (Aldezabal et al., 2003) shows us that, in some cases, we can automatically tag up to 82% of the occurrences of a verb and leave a small proportion of occurrences for manual tagging.

However, we want to stress that the automatic tagging is not a substitute for manual tagging. We plan to review all occurrences, regardless of whether they remain ambiguous or no.

Acknowledgments

This work has been partially funded by the Education Department of the Spanish Government (EPEC-RS project, HUM2004-21127-E) and (IMLT, TIN2007-63173)

References

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*. 12-19. Czech Republic.

Agirre E., Aldezabal I., Etxeberria J., Izagirre I., Mendizabal 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., Etxeberria J., Pociello E. (2006b). A Preliminary Study for Building the Basque PropBank. *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*. Genoa, Italy.

Aldezabal I. (1998). Levin's verb classes and Basque. A comparative approach. UMIACS Departmental Colloquia. University of Maryland.

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Sarasola K., Goenaga P. (2001). Extracción masiva de información sobre subcategorización verbal vasca a partir de corpus. *Actas del XVII Congreso de la SEPLN. Nº. 27. 29-36.* Universidad de Jaen, Spain.

Aldezabal, I., Aranzabe, M.J., 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, 39-48. Cambridge University Press.

Aldezabal, I. (2004). *Aditz-azpikategorizazioaren azterketa. 100 aditzen azterketa zehatza, Levin (1993) oinarri harturik eta metodo automatikoak baliatuz.* Leioa (Bilbao): University of Basque Country thesis.

Aldezabal (2007). Estudio preliminar para la creación de Euskal PropBank. In Irene Castellón and Ana Fernández (eds.), *Perspectivas de análisis de la unidad verbal*. SERES. Universitat de Barcelona. Spain.

Aldezabal, I., Aranzabe, M.J., Díaz de Ilarraza, A., Estarrona, A., Fernández, K., Uria, L. (2010). EPEC-RS: EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) rol semantikoekin etiketatzeko eskuliburua [Guidelines to tag semantic roles in the EPEC corpus (the Reference Corpus for the Processing of Basque)]. *Internal Report, UPV / EHU / LSI / TR 02-2010*.

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

Bengoetxea K., Gojenola K. (2007). Desarrollo de un analizador sintáctico-estadístico basado en dependencias para el euskera [Development of a statistical parser for Basque]. *Procesamiento del Lenguaje Natural 39*, 5-12.

Bird S., Maeda K., Ma X., Lee H., Randall B., Zayat S. (2002). TreeTrans: Diverse Tools Built on The Annotation Graph Toolkit. *Third International Conference on Language Resources and Evaluation*, 29-31, Las Palmas, Canary Islands, Spain.

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

Díaz de Ilarraza A., Garmendia Aitzpea, Oronoz M. (2004). Abar-Hitz: An annotation tool for the Basque Dependency Treebank. Paper presented at the International Conference on Language Resources and Evaluation. Lisbon, Portugal.

Hajic J., Panevová J., Urešová Z., Bémová A., Kolárová 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*, 57–68. Sweden.

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.

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.

Nianwen Xue. 2008. Labeling Chinese predicates with semantic roles. Computational Linguistics, 34(2): 225-255

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.

Rosén V., Smedt K.D., Dyvik H., Meurer P. (2005). TREPIL: Developing Methods and Tools for Multilevel Treebank Construction. In Civit M., Küber S. and Martí M (eds.), *Proceeding of the Fourth Workshop on Trebank and Linguistics Theories*, 161-172, Universitat de Barcelona, Spain.

Zapirain B., Agirre E., Màrquez L. (2008). Robustness and Generalization of Role Sets: PropBank vs. VerbNet. *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics, ACL-08: HLT*, 550-558, Columbus, Ohio.