Building the Basque PropBank

Izaskun Aldezabal*, María Jesús Aranzabe*, Arantza Díaz de Ilarraza and Ainara Estarrona****

IXA NLP Group
*Basque Philology Department, **Languages and Information Systems
University of the Basque Country
e-mail: {izaskun.aldezabal,maxux.aranzabe,a.diazdeilarraza,ainara.estarrona}@ehu.es

Abstract

This paper presents the work that has been carried out to annotate semantic roles in the Basque Dependency Treebank (BDT) (Aldezabal et al., 2009). In this paper we will present the resources we have used and the way the annotation of 100 verbs has been done. We have followed the model proposed in the PropBank project (Palmer et al., 2005). In addition, we have adapted AbarHitz (Díaz de Ilarraza et al., 2004), a tool used in the construction of the Basque Dependency Treebank (BDT), for the task of annotating semantic roles.

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. In this paper we present the work that has been carried out to annotate semantic roles of 100 verbs in the BDT (Basque Dependency Treebank). It is the continuation of previous annotation work developed in EPEC (Aduriz et al.,2006): a corpus that includes annotation of morphological information and several types of syntactic information, such as syntactic functions and chunks.

This paper deals with the work that has been carried out to annotate semantic roles in the BDT. 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, like VerbNet (Kingsbury et al., 2002) and Vallex (Hajic et al., 2003). FrameNet (Baker et al., 1998) is a further example of the joint development of a semantic lexicon and a hand-tagged corpus.

In this paper we explain the steps we are currently following to add a new semantic layer to BDT, in terms of semantic roles. The resources used are: an in-house database with syntactic/semantic subcategorization frames for Basque verbs (Aldezabal, 2004), an English-Basque verb mapping (Aldezabal 1998) based on Levin's classification (Levin, 1993) and the Basque Dependency Treebank (Aldezabal et al., 2009). In addition, we have adapted AbarHitz (Díaz et al., 2004), a tool used for the annotation of the BDT for the task of annotation of semantic roles.

The next section briefly reviews PropBank/VerbNet, which is the model followed, and BDT, EADB which are the resources used, and a further resource is English-Basque verb matching. 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). Finally, section 5 presents the conclusions and future work.

2. The resources used

After a preliminary study, we chose to follow the PropBank/VerbNet model. In fact, the PropBank model is being deployed in other languages, such as Chinese, Spanish, Catalan and Russian. In 2003 Palmer and Xue described the Chinese PropBank, as did Xue in 2005. Civit et al. (2005) described a joint project to annotate comparable corpora in Spanish, Catalan and Basque.

Below are some of the reasons why we chose the PropBank/VerbNet model in our study for Basque verbs:

1. The PropBank project starts from a syntactically annotated corpus, just as we do.
2. Lexicon organization is similar to our database of verbal models.
3. 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.
4. Having corpora which have been annotated in different languages following the same model allows for cross-lingual studies and hopefully the enrichment of Basque verbal models due to the more elaborate information currently available for English.

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

The corpus: BDT

We are using 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).

The EADB resource

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. 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. Typically we will call "alternation" the different syntactic structures represented by the same semantic predicate.

The 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 taking into account 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's classes comes from the fact that other resources like PropBank and VerbNet lexicon are linked to Levin's classes and contain information about semantic roles.

3. The annotation process

When constructing BDT we followed a dependency based syntactic formalism which provides a straightforward way for expressing semantic relations. So, the corpus annotated in this way constitutes a good base for tackling the next steps in the analysis-chain, such as verb valence and thematic role studies (Agirre et al., 2006b).

The process of manual annotation of semantic roles associated to verbs will begin with the tagging of 100 verbs contained in the corpus where the most frequent ones are included. The sentences of the corpus are grouped according to the verbs which belong to it.

In this preliminary study we did not want to consider light and modal verbs which will be treated in more depth later. That is the case of *egin* (='do') and *izan* (='be'), which are the two most frequent verbs in the corpus.

Once we have finished the 100 verbs, we are going to continue with the rest of the verbs.

The pre-process: 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's class so we were able to obtain automatically the PropBank/VerbNet information for each verb which was looked at from the paid data-base, based on Levin class.

However, since our mapping was done some time ago, Levin's classes in PropBank/VerbNet have been revised and consequently new classes and subclasses have been added, erased and modified. Thus, we implemented a simple algorithm to compare our previous assignment of Levin's classes and the new classes in PropBank/VerbNet. After comparing we have detected four cases.

- **equal**: represents the case in which the identification of the class for a verb has not changed since the mapping was done. For instance, "to say" and "to go" continue being in 37.7 and 47.7 classes respectively. This option represents 51% of the cases.
- **subclass**: a new subclass has been defined in PropBank. (6%)
- **changed**: a Levin class in PropBank has changed and there is not a direct coincidence between our mapping and the one in PropBank. (2%)
- **missing**: the verb is not included in PropBank or it has not been assigned to any Levin class. (41%)

Table 1 shows a sample of the results of the comparison between our mapping and PropBank regarding Levin's classes.

Levin's verbs	**Levin's classes**	**Aldezabal's work (1998)**	**Results**
burden	13.4.2	*zamatu/aspertu*	CHANGED
glom	22.3		MISSING
glue	22.4	*erantsi, kolatu*	EQUAL
glutenize	45.4		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

Table 1: the link between verbs in Levin (1993) and Basque.

This first step in annotation will deal with the first and second cases (57% of the cases) that cover 46% of the EPEC Corpus, leaving the rest to future study.

Representation of the semantic information (the 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[1] 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". Example: go_01.
- **V** (Verb): main verb,head of the relation
- **Treated Element** (TE): the element depending on 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): the roles usually associated with the numbered arguments and adjuncts in PropBank (Arg0: agent, experiencer, …).

[1] 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, …).

- **EADBrol**: semantic role according to EAD roleset (theme, state, location, experiencer, etc.)
- **HM** (Selectional Restriction). Up to now we only consider [+animate], [-animate], [+count], [- count], [+hum], [-hum]

The example illustrates the `arg_info` tag corresponding to the `ncmod` dependency between verb *joan* (”to go”) and argument in adlative case *Argentinara* (”to Argentina”) in the sentence:

“Argentinara joan zen taldea egongo da Pau Orthezen kontra”[2]

`arg_info`: (go_01, joan, *Argentinara*[3], Arg4, Destination, end_location, -[4]).

Enriching the BDT with information contained in the EADB

The sentences in the corpus containing the selected verbs are taken and the corresponding role tag is automatically created for each one of the syntactic occurrences of the arguments, according to the information contained in the EADB and based on the declension case.

In this way, arguments with non-ambiguous declension cases are automatically annotated. The ambiguous cases must be disambiguated by hand by the annotator. There is, however, an automatic proposal with all the possible tags available.

Visualizing the information of PropBank/VerbNet during the semantic annotating of the BDT

Based on the matching between Basque verbs and Levin’s classes done in Aldezabal (1998), the revision of the matching, and the BDT already built, we decided to use the information contained in VerbNet/PropBank (accessible by the Levin class) in such a way that the human tagger can easily identify the sense and the roles to be used when tagging the treated verb, without analyzing the whole database.

The tool for tagging we have developed (see more details section 4) facilitates the human annotator to visualize the information contained in PropBank/VerbNet and associate it to the verb which is being tagged.

4. AbarHitz, the tool for tagging

AbarHitz 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 AbarHitz 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.

AbarHitz communicates with the user by means of a friendly interface providing the following facilities: i) it visualizes the morphosyntactic information obtained so far; ii) it graphically visualizes the dependency-tree for each sentence and iii) it provides an environment for syntactic checking while tagging.

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 by displaying information from PropBank/VerbNet, on the right side of the window.
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 1 shows a screenshot of the tool AbarHitz. AbarHitz has been developed in Java; it follows a modular design in order to be a portable and easily maintained tool. It can be used with Microsoft Windows, Linux and Unix.

5. 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.

We tagged about 12,000 words of the corpus and we have defined general criteria for the tagging process. Structured and detailed set of guidelines for taggers and lexicon editors have been defined. However, it is a task that needs continuous updating, as new verbs are analyzed.

Our database of verbal models was a good starting point for the tagging task. We are detecting differences with English verbs regarding the status of arguments and adjuncts, due to different basic criteria, but those can be easily adjusted.

In the future we want to focus on the application of automatic methods for role tagging.

6. Acknowledgements

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

7. References

Aduriz I., Aranzabe M.J., Arriola J.M, Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R. (2006). Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing. In Andrew Wilson, Paul Rayson and Dawn Archer (eds.), *Corpus Linguistics Around*

[2] The team that went to Argentina will play against Pau Orthez

[3] to Argentina (PP)

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

the World. Book series: Language and Computers. Vol 56, 1-15. Netherlands: Rodopi.

Agirre E., Aldezabal I., Etxeberria J., Pociello E. (2006). 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. (2004). *Aditz-azpikategorizazioaren azterketa. 100 aditzen azterketa zehatza, Levin (1993) oinarri harturik eta metodo automatikoak baliatuz*. Leioa (Bilbao): University of Basque Country thesis.

Aldezabal I., Aranzabe M.J., Arriola J.M., Díaz de Ilarraza A. (2009). Syntactic annotation in the Reference Corpus for the processing of Basque (EPEC): Theoretical and practical issues. *Corpus Linguistics and Linguistic Theory 5-2,* 245-274. Mouton de Gruyter.

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 sintácticos-semánticos. *In XXI Congreso de la SEPLN*. Granada. Spain.

Díaz de Ilarraza A., Garmendia Aitzpea, Oronoz M. (2004). AbarHitz: 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.

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

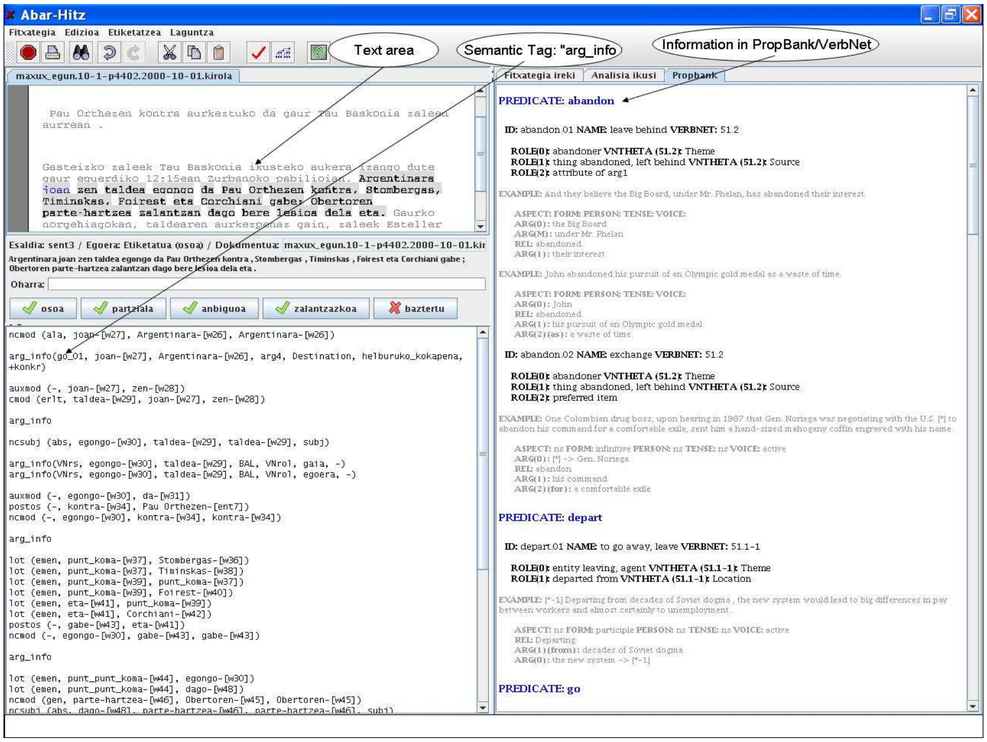
Marcus M. (1994). The Penn TreeBank: A revised corpus design for extracting predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*. Princeton, NJ.

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.

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.

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



Pau Orthezen kontra aurkeztuko da gaur Tau Baskonia zaleen aurrean .

Gasteizko zaleek Tau Baskonia ikusteko aukera izango dute gaur eguerdiko 12:15ean Zurbanoko pabilioian. Argentinara joan zen taldea egongo da Pau Orthezen kontra, Stombergas, Timinskas, Foirest eta Corchiani gabe; Obertoren parte-hartzea zalantzan dago bere lesioa dela eta. Gaurko norgehiagokan, taldearen aurkezpenaz gain, zaleek Esteller

Esaldia: sent3 / Egoera: Etiketatua (osoa) / Dokumentua: maxux_egun.10-1-p4402.2000-10-01.kir

Argentinara joan zen taldea egongo da Pau Orthezen kontra , Stombergas , Timinskas , Foirest eta Corchiani gabe ; Obertoren parte-hartzea zalantzan dago bere lesioa dela eta .

Oharra:

```
ncmod (ala, joan-[w27], Argentinara-[w26], Argentinara-[w26])

arg_info(go_01, joan-[w27], Argentinara-[w26], arg4, Destination, helburuko_kokapena,
+konkr)

auxmod (-, joan-[w27], zen-[w28])
cmod (erlt, taldea-[w29], joan-[w27], zen-[w28])

arg_info

ncsubj (abs, egongo-[w30], taldea-[w29], taldea-[w29], subj)

arg_info(VNrs, egongo-[w30], taldea-[w29], BAL, VNrol, gaia, -)
arg_info(VNrs, egongo-[w30], taldea-[w29], BAL, VNrol, egoera, -)

auxmod (-, egongo-[w30], da-[w31])
postos (-, kontra-[w34], Pau Orthezen-[ent7])
ncmod (-, egongo-[w30], kontra-[w34], kontra-[w34])

arg_info

lot (emen, punt_koma-[w37], Stombergas-[w36])
lot (emen, punt_koma-[w37], Timinskas-[w38])
lot (emen, punt_koma-[w39], punt_koma-[w37])
lot (emen, punt_koma-[w39], Foirest-[w40])
lot (emen, eta-[w41], punt_koma-[w39])
lot (emen, eta-[w41], Corchiani-[w42])
postos (-, gabe-[w43], eta-[w41])
ncmod (-, egongo-[w30], gabe-[w43], gabe-[w43])

arg_info

lot (emen, punt_punt_koma-[w44], egongo-[w30])
lot (emen, punt_punt_koma-[w44], dago-[w48])
ncmod (gen, parte-hartzea-[w46], Obertoren-[w45], Obertoren-[w45])
ncsubj (abs, dago-[w48], parte-hartzea-[w46], parte-hartzea-[w46], subj)
```

PREDICATE: abandon

ID: abandon.01 **NAME:** leave behind **VERBNET:** 51.2

ROLE(0): abandoner **VNTHETA (51.2):** Theme
ROLE(1): thing abandoned, left behind **VNTHETA (51.2):** Source
ROLE(2): attribute of arg1

EXAMPLE: And they believe the Big Board, under Mr. Phelan, has abandoned their interest.

ASPECT: FORM: PERSON: TENSE: VOICE:
ARG(0): the Big Board
ARG(M): under Mr. Phelan
REL: abandoned
ARG(1): their interest

EXAMPLE: John abandoned his pursuit of an Olympic gold medal as a waste of time.

ASPECT: FORM: PERSON: TENSE: VOICE:
ARG(0): John
REL: abandoned
ARG(1): his pursuit of an Olympic gold medal
ARG(2) (as): a waste of time.

ID: abandon.02 **NAME:** exchange **VERBNET:** 51.2

ROLE(0): abandoner **VNTHETA (51.2):** Theme
ROLE(1): thing abandoned, left behind **VNTHETA (51.2):** Source
ROLE(2): preferred item

EXAMPLE: One Colombian drug boss, upon hearing in 1987 that Gen. Noriega was negotiating with the U.S. [*] to abandon his command for a comfortable exile, sent him a hand-sized mahogany coffin engraved with his name

ASPECT: ns **FORM:** infinitive **PERSON:** ns **TENSE:** ns **VOICE:** active
ARG(0): [*] -> Gen. Noriega
REL: abandon
ARG(1): his command
ARG(2) (for): a comfortable exile

PREDICATE: depart

ID: depart.01 **NAME:** to go away, leave **VERBNET:** 51.1-1

ROLE(0): entity leaving, agent **VNTHETA (51.1-1):** Theme
ROLE(1): departed from **VNTHETA (51.1-1):** Location

EXAMPLE: [*-1] Departing from decades of Soviet dogma , the new system would lead to big differences in pay between workers and almost certainly to unemployment .

ASPECT: ns **FORM:** participle **PERSON:** ns **TENSE:** ns **VOICE:** active
REL: Departing
ARG(1) (from): decades of Soviet dogma
ARG(0): the new system -> [*-1]

PREDICATE: go

Figure 1: An example of AbarHitz proposed to the human annotator