A METHODOLOGY FOR THE SEMIAUTOMATIC ANNOTATION OF EPEC-ROLSEM, A BASQUE CORPUS LABELED AT PREDICATE LEVEL FOLLOWING THE PROPBANK-VERBNET MODEL

Izaskun Aldezabal Roteta
María Jesús Aranzabe Urruzola
Arantza Díaz de Ilarraza Sánchez
Ainara Estarrona Ibarloza

UPV/EHU / LSI / TR 01-2013

A methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labeled at predicate level following the PropBank-VerbNet model

Abstract:

In this article we describe the methodology developed for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labeled at predicate level following the PropBank-VerbNet model. The methodology presented is the product of detailed theoretical study of the semantic nature of verbs in Basque and of their similarities and differences with verbs in other languages. As part of the proposed methodology, we are creating a Basque lexicon on the PropBank-VerbNet model that we have named the *Basque Verb Index* (BVI). Our work thus dovetails the general trend toward building lexicons from tagged corpora that is clear in work conducted for other languages. EPEC-RolSem and BVI are two important resources for the computational semantic processing of Basque; as far as the authors are aware, they are also the first resources of their kind developed for Basque. In addition, each entry in BVI is linked to the corresponding verb-entry in well-known resources like PropBank, VerbNet, WordNet and Levin's Classification. We have also implemented several automatic processes to aid in creating and annotating the BVI, including processes designed to facilitate the task of manual annotation.

Keywords: predicate labeling, verb sense, valency, semantic roles, evaluation, PropBank/VerbNet

Laburpena:

Lan honetan, EPEC-RolSem corpusa etiketatzeko jarraitu dugun metodologia deskribatuko dugu. EPEC-RolSem corpusa PropBank-VerbNet ereduari jarraiki predikatu-mailan etiketatutako euskarazko corpusa da. Etiketatze-lana aurrera eramateko euskal aditzen izaera semantikoa aztertu eta ingeleseko aditzekin konparatu dugu, azterketa horren emaitza da lan honetan proposatzen dugun metodologia. Metodologiaren atal bat PropBank-VerbNet eredura sortutako euskal aditzen lexikoiaren osaketa izan da, lexikoi hau *Basque Verb Index* (BVI) deitu dugu. Gure lanak alor honetan beste hizkuntzetan dagoen joera nagusia jarraitzen du, hau da, etiketatutako corpusetatik lexikoiak sortzea. EPEC-RolSem eta BVI oso baliabide garrantzitsuak dira euskararen semantika konputazionalaren alorrean, izan ere, euskararako sortutako mota honetako lehen baliabideak dira. Honetaz guztiaz gain, BVIko sarrera bakoitza PropBank, VerbNet, WordNet, Levinen sailkapena eta FrameNet bezalako baliabide ezagunekin lotua dago. Hainbat prozesu automatiko inplementatu ditugu EPEC-RolSem corpusaren eskuzko etiketatzea laguntzeko eta baita BVI sortzeko eta osatzeko ere.

Gako-hitzak: predikatu-mailako etiketatzea, aditz-adiera, balentzia, rol semantikoak, ebaluazioa, PropBank-VerbNet.

1 Introduction and context

In this article we offer a detailed description of a methodology we have developed for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labeled at predicate level following the PropBank-VerbNet model (hereafter PB-VN). This methodology is part of a more general ongoing work the Ixa group [1] is pursuing regarding corpora-tagging frameworks. It makes use of the EPEC corpus (*Euskararen Prozesamendurako Erreferentzia Corpusa-Reference Corpus for the Processing of Basque)* (Aduriz et al. 2006), which contains 300,000 words of standard written text and is intended to function as a training corpus for the development and improvement of several NLP tools (Bengoetxea and Gojenola 2007) [2]. The EPEC corpus has previously been tagged morphologically and syntactically using a dependency grammar (Basque Dependency Treebank (Aldezabal et al. 2009)), and the aim now is to incorporate predicate information on the basis of the dependencies that are argument/adjunct candidates. Another major part of our project is the creation of a verb lexicon, tallying with work conducted for other languages that also builds lexicons from tagged corpora – for instance, the Penn Treebank (Marcus 1994); PropBank (Palmer et al. 2005a), related to the Verb-net lexicon (Kingsbury and Palmer 2002); or PDT, related to the Vallex lexicon (Hajic et al. 2003). These kinds of semantic resources are essential for many computational tasks, such as syntactic disambiguation and language understanding, as well as for advanced applications such as question answering, machine translation and text summarization.

Three basic questions have to be decided when engaging in corpus annotation: (a) what *model* to use for annotation, (b) what *methodology* and *guidelines* to employ in applying the model, and (c) what *tool* to use for tagging.

We chose the PB-VN as the *model* for predicate labeling. After conducting several analyses to find the most suitable model, we concluded that the one used by PropBank and VerbNet was appropriate for Basque (Agirre et al. 2006; Aldezabal et al. 2010a; Aldezabal et al. 2010c). This is due to three basic reasons: 1) The PropBank project starts out with a syntactically annotated corpus, exactly as we do; 2) it has been used for major projects in other languages (Palmer et al. 2005b; Xue 2008; Civit et al. 2005, among others), and 3) the organization of the lexicon is similar to our first database of Basque verbs (EADB–Data Base for Basque Verbs, proposed in Aldezabal (2004), see section 4.1). We have named the Basque lexicon defined in the PB-VNet style the *Basque Verb Index* (BVI).

We defined the first version of the *guidelines* in accordance with a preliminary *methodology* that we had planned to use for the annotation (Aldezabal et al. 2010b). However, the results obtained in an evaluation (Aldezabal et al, 2011) revealed that our preliminary methodology required modification. Those modifications and the reasons behind them form the core of the present article.

As the *tool*, we are using AbarHitz (Díaz de Ilarraza et al. 2004). AbarHitz is a tool designed in our group to help linguists in the manual annotation of the EPEC corpus at different linguistic levels. It follows the general annotation schema for representing linguistic information that we have established (Artola et al. 2009) and forms part of a general environment designed to integrate general processors and resources. AbarHitz has been adapted to facilitate the annotation at predicate level by offering the linguist new options; this feature will be described in greater detail below.

This work has resulted in the development of two important resources for the computational semantic processing of Basque: one, BVI, a verb lexicon that currently contains 246 verbs and their predicate information and two, EPEC-RolSem, a semantically tagged version of the EPEC corpus (at the time of writing, 71% of the corpus has been tagged).

The article is organized as follows: In section 2 we explain some basic considerations when applying the PB-VN model, and consider some language-specific problems when adapting it to Basque. In section 3 we explain the semantic tag (arg_info) used when tagging the verb complements. Section 4 explains the resources we have based our work on and the pre-process we have applied. In section 5 we study in depth the final methodology proposed for the best annotation of the corpus, with special attention to the required methodological alterations as compared to earlier versions. In section 6 we report on the data developed up to the present as well as offer some numbers on the work team and work time needed for its development. Finally, in section 7, we consider some potential future lines of investigation.

2 Basic considerations when applying the PB-VN model and criteria for adapting it to Basque

Adapting a predicate annotating model from one language to another is never straightforward. On the one hand, one encounters language-specific issues; on the other, the model itself may contain both questionable aspects and gaps in its coverage of linguistic phenomena. Thus, even after carrying out the several studies required to resolve the question of the most appropriate predicate annotating model (Agirre et al. 2006; Aldezabal et al. 2010a; Aldezabal et al. 2010c), we were still faced with the challenge of adapting the model to Basque. In this section we discuss our experience and the consequent decisions regarding both language-specific and model-internal problems.

The PropBank model (Palmer et al. 2005a) distinguishes between two *independent* levels: 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 roles specific to each concrete verb (e.g. buyer, thing bought, etc.), and these are linked to the VerbNet lexicon (Kipper et al. 2002), which in turn has general roles (e.g. agent, theme, etc.). VerbNet is an extensive lexicon where verbs are organized in classes following Levin's classification (Levin 1993).

Table 1 shows the PropBank roleset for the verb 'tell.01' and the corresponding VerbNet roleset with the Levin 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".

We see that PropBank and VerbNet offer complementary information, as observed by Merlo et al. (2009). PropBank provides the valency relation of each verb sense, while VerbNet gives a more class oriented role specification. These features of PropBank and VerbNet occasionally cause conflicting interpretations, which we discuss in more detail below.

- **Regarding Arg0 and Arg1.**

As noted above, PropBank distinguishes two independent levels (argument and roles). In fact, however, Arg1 is always labeled *Theme* and Arg0 *Agent*. No fundamental linguistic reason exists for this, though for example Kingsbury and Palmer (2003:3) offer arguments like the following:

"(...) Arg0 is very consistently assigned an "Agent"-type meaning, while Arg1 has a Patient or Theme meaning almost as consistently. There are, of course, many verbs in English for which the Patient, the entity undergoing the action of the verb, always appears in subject position. For these verbs no agent is possible. In order to maintain the consistency of Arg1 as Patient these verbs have no Arg0. A canonical example is *fall*” as seen in Figure 1:

fall.01 sense: move downward
roles:
Arg1: thing falling
Arg2: extent, distance fallen
Arg3: start point
Arg4: end point

Figure 1: the verb “fall.01” in PropBank (quoted in Kingsbury and Palmer (2003:3))

Nevertheless, inconsistencies abound. For instance, Babko-Malaya et al. (2006:76) report: *“*In *John and Mary come* the NP *John and Mary* is a constituent in Treebank and it is also marked as ‘Arg0’ in PropBank.” But when we check it in PropBank we realize that the verb “come” is defined as:

come.01
roles:
Arg1: entity in motion (theme)
Arg2: extent
Arg3: start point
Arg4: end point

Figure 2: the verb “come.01” in PropBank

Given such inconsistencies, our decision has been to maintain the independence of levels (and thus to follow the model faithfully), and consequently we have not automatically equated Arg0 and Arg1 to agent and theme, respectively.

Specifically regarding intransitive verbs denoting change of position, we consider the subject to be at the same time the entity who initiates the action and the one who undergoes it (agreeing with Vázquez et al. (2000: 183)). Therefore, we annotate the subjects of such verbs as Arg0. This decision is based on a principle taken from the PropBank guidelines (section *Choosing Arg0 versus Arg1*):

> “Whereas for many verbs, the choice between Arg0 or Arg1 does not present any difficulties, there is a class of intransitive verbs (known as verbs of variable behavior), where the argument can be tagged as either Arg0 or Arg1.
>
> (…)
>
> Arguments which are interpreted as agents should always be marked as Arg0, independent of whether they are also the ones which undergo the action.
>
> **(…)**
>
> **In general, if an argument satisfies two roles, the highest ranked argument label should be selected, where Arg0 >> Arg1 >> Arg2>>… .**”
>
> (Babko-Malaya 2005:4)

Thus, in the case of an unaccusative verb like “come.01” where only the intransitive variant is possible, we consider the entity who performs the action and the one who undergoes it to be the same; thus, we tag it as Arg0 *Theme*. In PropBank, on the other hand, the subject of these kinds of change of position verbs is also annotated as *Theme* but numbered Arg1. In our opinion, the *Agent* role is more appropriate for an entity that initiates an action oriented toward another entity. On the other hand, in causative/inchoative verbs like *break* we always annotate the *Theme* as Arg1 because we consider the *Cause* (Arg0) always to exist, even when it is not explicit in the sentence.

It should be noted that work applying the PropBank model to other languages has followed the PropBank criteria (Arg0_Agent, Arg1_Theme); examples include Arabic (Palmer et al. 2008), Hindi (Palmer et al. 2009), Korean (Palmer et al. 2006), Chinese (Xue et al. 2009) and Spanish (Aparicio 2007).

In other models – for instance, in the case of Spanish, Semsem (Vázquez et al, 2006) and Adesse (García-Miguel, JM. and Albertuz FJ., 2005) – this particular problem does not arise because these models do not use numbered arguments.

- **Disagreements between PropBank and VerbNet**

Sometimes, PropBank and VerbNet do not agree regarding the valency of arguments. Even though the EADB agrees with VerbNet in most cases where PropBank and VerbNet disagree, in our current work we have generally decided to follow PropBank, since it is the model that focuses on valency. However, there are some exceptions. Below we discuss a few examples that should clarify our decisions.

First, let us take an example that fulfills the general criterion: the Basque verb *hasi*.

Hasi is linked to the following PropBank verbs:

begin.01 , start, vncls: 55.1, framnet:
roles:
Arg0: beginner, Agent (vnrole: 55.1-Agent)
Arg1: Theme(-Creation) (vnrole: 55.1-Theme)
Arg2: Instrument

start.01 , begin, vncls: 55.1, framnet:
roles:
Arg0: Agent (vnrole: 55.1-Agent)
Arg1: Theme(-Creation) (vnrole: 55.1-Theme)
Arg2: Instrument

commence.01 , begin, vncls: 55.1, framnet:
roles:
Arg0: beginner, Agent (vnrole: 55.1-Agent)
Arg1: Theme(-Creation) (vnrole: 55.1-Theme)
Arg2: Instrument

Figure 3: the verbs "begin.01", "start.01" and "commence.01" in PropBank.

Here we can see that the three verbs that can be equivalents for the Basque *hasi* have an *Instrument* argument in PropBank, whereas in VerbNet such an argument is not defined. PropBank offers some examples for the use of "Arg2: Instrument":

(2) *John started the book with a murder*

Arg0: John
Rel: started
Arg1: the book
Arg2: with a murder

In the EADB this *Instrument* "argument" is not considered an argument; it is classified as a common modifier (denoting manner) like in any other verb. However, as the instrument argument causes no problems regarding sense distinction, we have considered it an argument and included it in our BVI lexicon as instrument. Here we have an example in the EPEC corpus:

(3) *Legebiltzar saioa "Libanoko hegoaldea askatzeko borrokan eroritakoei" eskainitako minutu bateko isilunearekin hasi zuten.* The Parliament session started with a minute of silence dedicated to the people killed in the fight for the freedom of Lebanon.

Isilunearekin (with silence) is tagged as: "Arg2, Instrument".

On the other hand, in the case of the Basque *adierazi*, linked to "state.01" in PropBank, we have followed VerbNet. Here is the description of the verb "state.01" in PropBank:

state.01 , state, say, vncls: 37.7, framnet:
roles:
Arg0: announcer (vnrole: 37.7-Agent)
Arg1: utterance (vnrole: 37.7-Topic)
Arg2: hearer (vnrole: 37.7-Recipient)
Arg3: attributive

Figure 4: the verb "state.01" in PropBank.

The Arg3 proposed in PropBank has no equivalent in VerbNet. Also, in the only example found in PropBank there is no Arg3:

(4) *The Japanese government, Mr. Godown said, has stated that it wants 10% to 11% of its gross national product to come from biotechnology products.*

Arg0: The Japanese government
Rel: stated
Arg1: that it wants 10% to 11% of its gross national product to come from biotechnology products

In this case, we decided to follow VerbNet and assigned three arguments to the *adierazi* verb, because (a) this verb has only one sense so the fourth argument does not help in distinguishing senses and (b) in the only example that appears in PropBank there is no Arg4.

In the same way, in the case of the *esan* verb (similar in sense to "state") we find an "Arg3, Attributive" in PropBank that does not appear in VerbNet (nor in the EADB). However, in this verb the "Arg3, Attributive" marks the difference between senses:

say.01 , say, vncls: 37.7, 78-1, framnet: Spelling_and_pronouncing , Text_creation , Statement

roles:
Arg0: sayer (vnrole: 37.7-agent, 78-1-cause)
Arg1: utterance (vnrole: 37.7-topic, 78-1-topic)
Arg2: hearer (vnrole: 37.7-recipient, 78-1-recipient)
Arg3: attributive

Figure 5: “say.01” verb in PropBank.

The verb *esan* has two senses in the EADB. The first one would be the equivalent of the English verb “to say”:

1- Communication action; two arguments in two syntactic variants:

1.1: experiencer [+human] (ERG [3]), theme [-concrete] (ABS)
1.2: experiencer [+human] (ERG), theme (KONP)

The second one would be the equivalent of the English verb *“to call”*.

2- Assignment of an attribute/quality to an entity; three arguments in a single syntactic realization:

2.1: startpoint [+human] (ERG), goal (DAT [4]), attributive (ABS)

The Arg3 proposed by PropBank for the verb “say” is possible in the first sense*,* but not in the second one. That is, although it is not a frequent argument and seems to resemble an adjunct even when it does appear, it does distinguish between senses, unlike in the previous case (state.01). As a consequence, agreeing with PropBank, we regard it as “Arg3, Attributive”. Thus, we define the “esan” verb in the PB/VN style as follows:

Arg0: Agent, experiencer [+human] (ERG)
Arg1: Topic, theme [-concrete] (ABS / KONP)
Arg2: Recipient, -, - (DAT)
Arg3: Attributive, -, - (-ri buruz) [5]

Figure 6: the verb “esan_say.01” in the BVI.

- **VerbNet assigns two roles to the same numbered argument**

Sometimes VerbNet assigns two different roles to the same argument of a verb since, although the verb has one roleset, it is linked to two subclasses. For example, this is the case for the verb ‘see.01’:

see.01, view, vncls: 29.2 30.1
roles:
Arg0: viewer (vnrole: 29.2-Agent, 30.1-Experiencer)
Arg1: thing viewed (vnrole: 29.2-Theme, 30.1-Stimulus)

Figure: 7: the verb “see.01” in PropBank

Arg0 has associated *Agent* and *Experiencer* roles and Arg1 associated *Theme* and *Stimulus* roles.

By contrast, in the EADB the verb *ikusi* contains two arguments and one role is assigned to each argument:

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

In this ambiguous case, we have decided to base our decision on the EADB and to assign the corresponding VerbNet roles, that is, *Agent* (represented by *Experiencer* (and *Cause*) in the EADB) and *Theme*. The result would be:

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

- **The ADV role**

There is an ADV "role" in the PropBank/VerbNet role repertory whose use is not very clear. We will use it when an adverb is ambiguous as to whether it is a temporal (TMP), modal (MNR), location (LOC) or some other kind of modifier.

(5) *Houdaren familiak asko jaten du.* Houda's family eats a lot.

- **Including a *path* role**

We have found it necessary to add a path role. This role is not specified in VerbNet, but appears in our EADB. For instance, for the verb *pasatu* ("pass" / "come by") we find examples like:

(1) *Zure etxetik pasatu naiz gaur goizean*. I have come by your house this morning.

2.1 Interlingual differences. Criteria for applying the model to Basque

Applying the PB-VN model to Basque is mainly a question of including in a verb sense the distribution of the arguments and adjuncts as well as the roles proposed for them. For example, in the EADB the Basque verb *eskatu* (= "ask.02"), has two arguments, Arg0: *Esperimentatzailea* (Experiencer) and Arg1: *Gaia* (theme). The dative complement is not included within the subcategorized cases because it is optional. However, the verb "ask.02" contains 3 arguments in PropBank and VerbNet:

Arg0: Agent
Arg1: Theme (proposition)
Arg2: Patient

Therefore, we follow the PB-VN model, tagging the DAT (dative) argument as Arg2. However, as we performed the verb tagging, we encountered some difficult cases. We explain the main phenomena below.

- **Arguments proposed by PB-VN that are not possible in Basque**

In some verbs of displacement, PB-VN proposes an argument, *Extent*, that is not possible in Basque. We can illustrate this with the verb *joan* (go.01):

go.01: motion, vncls: 47.7, 51.1, framnet: Motion
Roles:
Arg1: entity in motion/goer (vnrole: 47.7-theme, 51.1-2-theme)
Arg2: extent
Arg3: start point
Arg4: end point, end state of arg1
Argm: medium
Argm: direction (usually up or down)

Figure 8: the verb "go_01" in PropBank.

In Basque the second argument is not possible; one cannot say "*lau metro joan naiz (sukaldetik gelara)*" (Lit. I have gone four meters (from the kitchen) (to the bedroom)). As a consequence, we disregard this argument and assign its number to the next possible argument. That is, Arg1 will be the "start point" (since for us in this verb the subject is Arg0) and the "end point" will be Arg2.

After these changes, the resulting entry is the same as in the EADB:

The Basque verb *joan*:

1: affected theme_ABS; start point_ABL [6]; end point_ALA
2: affected theme_ABS; start point [+animate]_DAT; end point_ALA

joan_go.01

Arg0: Theme, affected theme (ABS)
Arg1: Source, start point (ABL/DAT)
Arg2: Destination, end point (ALA)

Figure 9: the "joan_go.01" verb in the BVI.

- **More than one PropBank verb exists for a Basque verb**

Sometimes a Basque verb can be linked to more than one PropBank verb. In such cases, we check, first of all, whether the roles and arguments of the Basque verb coincide with the roles and arguments of each of its PropBank equivalents.

If they do coincide, we assign them all in each tagging instance. For example, the verb *esan* can be linked unquestionably with both "tell.01" and "say.01". We establish the correspondence and indicate this double equivalence by the expression "tell.01/say.01" as first value of arg_info tag [7].

In other cases, although the English verbs are the same at predicate level ("make.01", "build.01", "construct.01", for instance), we annotate the concrete instances with the one we consider more suitable for the context. In some cases, the roles and arguments are also different. We can find both cases (same and distinct predicate description) in the verb *egin*:

make.01 / build.01 / construct.01
Arg0: Agent, source (ERG)
Arg1: Product, created theme [-humm] (ABS)

Arg2: Material, - (INS)
Arg3: Beneficiary, - (DAT/DES)

do.02
Arg0: Agent, source (ERG)
Arg1: Product, created theme (ABS)
Arg2: Instrument, - (INS/SOZ)
Arg3: Beneficiary, - (DAT)

ask.02
Arg0: Agent, source (ERG)
Arg1: Topic, created theme [-humm] (ABS)
Arg2: Recipient, - (DAT)

compose.02
Arg0: Agent, source (ERG)
Arg1: Product, created theme (ABS)
Arg2: Beneficiary, - (DAT/DES)

practice.01
Arg0: Agent, source (ERG)
Arg1: Theme, theme (ABS)
Arg2: Instrument, - (INS/SOZ)

[...]

Figure 10: Description of the *egin* verb in the BVI.

3 The tag for predicate labeling

The EPEC-RolSem corpus we are creating takes as a basis the EPEC corpus (*Euskararen Prozesamendurako Erreferentzia Corpusa-Reference Corpus for the Processing of Basque)* (Aduriz et al., 2006). As mentioned above, the EPEC corpus has already been tagged morphologically and syntactically following the dependency grammar (Basque Dependency Treebank (Aldezabal et al., 2009), and the aim now is to incorporate predicate information on the basis of the dependencies that are argument/adjunct candidates. To accomplish this we use the semantic label *arg_info,* which is assigned to each syntactic dependent that is a candidate for the verb argument/adjunct. For instance, in the dependency tree of the sentence "The team that went to Argentina will play against Pau Orthez", shown in Figure 11, the arg_info tag will be assigned to the *ncsubj* ("the team") and two *ncmod*s ("to Argentina" and "against Pau Orthez") linked to the verb (the head).

The "arg_info" label comprises the following fields:

- **PB** (PB-VN verb): the verb in English and its PropBank number, e.g.: *go.01*
- **V** (verb): dependency-relationship head, main verb
- **Element being worked on** (TE): argument/adjunct candidate
- **VAL** (valency): the number of the arguments, and adjuncts: arg0, arg1, arg2, arg3, arg4, argM
- **VNrol** (VerbNet role): the VerbNet role assigned to the PropBank argument/adjunct. (Arg0: agent, experiencer…)
- **EADBrol:** the semantic role appearing in the EADB (Data Base for Basque Verbs)

- **HM** (Selectional restriction): at present, only the following are taken into consideration: [+animate], [-animate], [+human], [-human], [+concrete], [-concrete]

Figure 11 shows in tree format a compound sentence annotated syntactically, where semantic annotation has been added to the phrase in the adlative case (ALA) linked to the verb *joan* ('go'). 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 [8] are marked on the links, and the semantic information on the nodes. The declension case is included in the nodes as additional information.

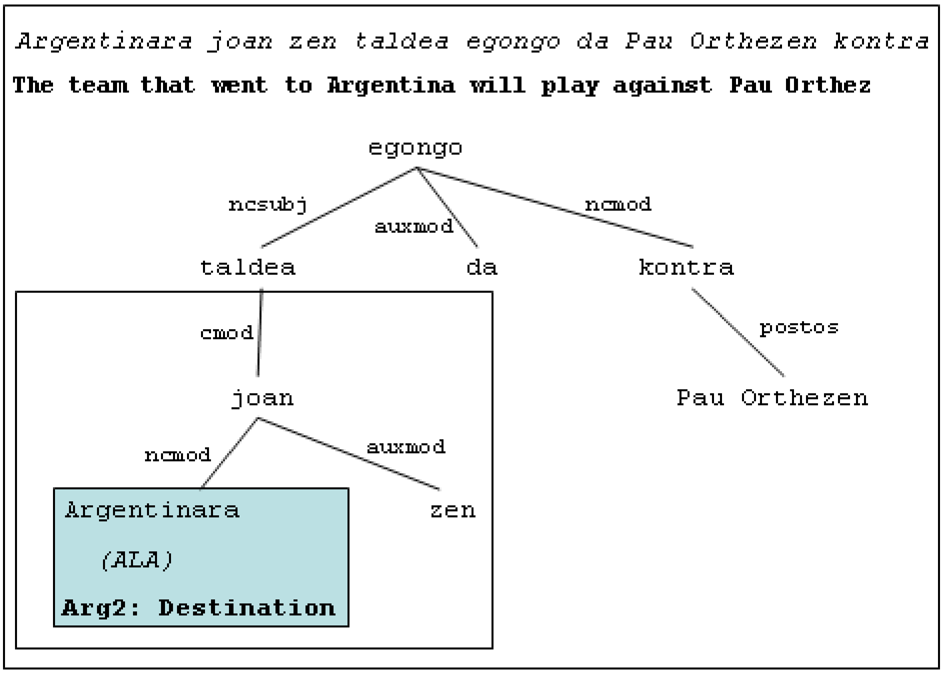
Argentinara joan zen taldea egongo da Pau Orthezen kontra
The team that went to Argentina will play against Pau Orthez
egongo
ncsubj
auxmod
ncmod
taldea
da
kontra
cmod
postos
joan
Pau Orthezen
ncmod
auxmod
Argentinara
(ALA)
Arg2: Destination
zen

Figure 11: The dependency tree for the sentence "The team that went to Argentina will play against Pau Orthez".

Here we have the dependency tagging corresponding to the example in Figure 11:

ncmod (ala, joan, Argentinara, Argentinara)
auxmod (- , joan, zen)
cmod (erlt, taldean, joan, zen)
ncsubj (abs, egongo, taldea, taldea, subj)
auxmod (- , egongo, da)
postos (gen, kontra, Pau_Orthezen)
ncmod (- , egongo, kontra, kontra)

Example (6) illustrates the arg_info tag that corresponds to the *ncmod* "*Argentinara*" ('to Argentina') in Figure 11.

(6) arg_info: (go_01, joan, *Argentinara* [9], Arg2, Destination, end_location, - [10])

4 Some basic resources and pre-processes

Before discussing the methodology we use, it may be useful to briefly describe the resources we based the project on as well as the automatic procedures that we have been able to use to facilitate the tagging task.

4.1 The EADB resource (Data Base for Basque Verbs)

Our starting point is the work carried out in (Aldezabal 2004), which involved an in-depth study of 100 verbs for Basque from EPEC and created the first version of the EADB. Aldezabal defined a number of syntactic-semantic frames (SSF) for each verb. Each SSF is composed of semantic roles and the corresponding declension case that syntactically performs each 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, Destination, 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 Supposition.

Here is an example of an EADB verb entry:

joan.1 ("go"): entity in motion
affected theme_ABS; start location / path_ABL; end location_ALA
joan.2 ("go"): entity in motion
affected theme_ABS; start location [+animate]_DAT; end location _ALA
joan.3 ("go"): feature that disappears from an entity
container_DAT; content [-animate, -concrete]_ABS

Figure 12: The entry for the *joan* verb in the EADB.

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

Aldezabal (1998) compares English and Basque verbs based on Levin's alternations and classification. For this purpose, all 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 lies in the fact that other resources like PropBank and VerbNet lexicon are also linked to Levin classes and contain information about semantic roles. Verbs in a particular Levin class display regular behavior (according to diathesis alternation criteria) that is different from verbs belonging to other classes. Also the classes are semantically coherent and verbs belonging to the same class share the same semantic roles. Table 2 shows some examples of the links between verbs in Levin (1993) and Basque verbs.

tell	37.1	*esan, erran*
tell	37.2	*esan, erran*
tense	45.4	*teinkatu, tinkatu, gogortu*

term	29.3	*deitu, izendatu, -tzat hartu/eduki*
terminate	55.1	*bukatu, amaitu*
terrify	31.1	*izutu, izuarazi*
terrorize	31.1	*izua sartu, ikaratu*
tether	22.4	*sokaz lotu*
thank	33	*eskertu, eskerrak eman*

Table 2: Some examples of the links between verbs in Levin (1993) and Basque verbs.

4.3 The pre-process: comparison of the Levin classes in our mapping with the PropBank data-base

Drawing on the resources described above, we have carried out an automatic pre-process in which two tasks have been automated:

(a): If our Basque-English mapping contains an English equivalent for a Basque verb in EPEC, the PB-VN information for that English verb has been made visible in the tagging tool AbarHitz (Díaz de Ilarraza et al. 2004).

(b): Some of the information contained in the EADB has been linked to EPEC.

More detailed descriptions of these two tasks follow.

(A) Since we already had a mapping between some Basque and English verbs in terms of the Levin class, we were able to obtain automatically the PB-VN information for each of these verbs. However, our mapping was done some time ago, and the Levin classes in PB-VN have since been revised: classes and subclasses have been added, erased and modified. Thus, we implemented a simple algorithm to compare the classes in Levin (1993), used in our mapping, and the classes in PB-VN. The results of the comparison fall into four categories:

- **equal**: the cases in which the identification of the class for a verb had not changed since the mapping was done. For instance, "to say" and "to go" remained in classes 37.7 and 47.7, respectively. This category represented 74,92% of the cases.
- **subclass**: a new subclass had been defined in PB-VN (9,46%).
- **changed**: a Levin class in PB-VN had changed and there was no direct correspondence between our mapping and the one in PB-VN (2,7%).
- **missing**: the verb was not included in PB-VN or it has not assigned a Levin class (12,8%).

Table 3 shows a sample of the results of the comparison between the classes on Levin (1993) and the classes in the current PB-VN data.

Levin's verbs	**Levin's classes 1993**	**The class in PB-VN**	**Results**
adjudicate	29.4	-	MISSING
tattoo	29.1	25.1	CHANGED
tell	37.1	37.1-1	SUBCLASS
tell	37.2	37.2-1	SUBCLASS
tense	45.4	45.4	EQUAL
term	29.3	29.3	EQUAL
terminate	55.4	55.4	EQUAL

terrify	31.1	31.1	EQUAL
terrorize	31.1	31.1	EQUAL
tether	22.4	22.4	EQUAL
thank	33	33	EQUAL

Table 3: the link between verbs in Levin (1993) and Basque.

Verbs falling into the first and second categories (84,38%) could be linked to PB-VN and their information displayed in the AbarHitz annotation tool.

(B) Adding the information contained in the EADB into EPEC.

This process involves taking the sentences in the EPEC corpus that contain EADB verbs and, with the aid of the information contained in the EADB, automatically creating a role tag for each of the syntactic occurrences of the arguments of the verb on the basis of the declension case.

In this way, arguments with non-ambiguous declension cases are automatically annotated; ambiguous cases must be manually disambiguated by the annotator. The annotator can, however, draw on an automatically generated proposal that contains all the possible tags.

Here is an example of a non-ambiguous case, *adierazi* (“to state”):

The EADB includes the following information for the *adierazi* verb:

1: *esperimentatzailea* (‘experiencer’)_ERG; *gaia* (‘theme’) *[-biz* ‘-animate’; *-konkr* ‘-concrete’]_ABS

2: *esperimentatzailea* (‘experiencer’)_ERG; *gaia* (‘theme’) *[-biz* ‘-animate’; *-konkr* ‘-concrete’]_KONP

(4): *Israelgo helikopteroek gune palestinarrak bonbardatu zituztela adierazi zuten lekukoek.*

(4)’ The witnesses stated that Israeli helicopters bombarded the Palestinian area.

On the basis of the *–ela* subordinating conjunction and the ergative declension case, the preprocessing tool will prepare the following arg_info for the subordinating clause “that Israeli helicopters bombarded the Palestinian area” and for the subject “the witnesses”:

ccomp_obj (**konpl**, adierazi-[w250], bonbardatu-[w248], zituztela-[w249])
arg_info (-, adierazi-[w250], zituztela-[w249], -, -, theme, -human/-concrete)

ncsubj (**erg**, adierazi-[w250], lekukoek-[w252], lekukoek-[w252], subj)
arg_info (-, adierazi-[w250], lekukoek-[w252], -, -, experiencer, -)

Figure 13: The arg_info of the subordinating clause and subject of *adierazi*, produced automatically on the basis of the *–ela* subordinating conjunction and the *-k* ergative declension case.

The rest of the information needs to be filled in manually.

By contrast, *gertatu* is an example of an ambiguous case. For the second sense of *gertatu* (state of an entity “to be, to end up”), the EADB offers the following information:

1- *gaia* (“theme”)_ABS; *egoera* (“state”)_ABS

As can be seen, the two arguments are syntactically realized with the same declension case (ABS). As a consequence, the automatic system creates two labels for each which need to then be manually disambiguated:

(5): *Espezieen babespen egokia gerta dadin, habitat bera babestu egin behar da.*
(5)’ For the best protection of the species, their habitat must be protected.

ncsubj (**abs**, gerta-[w1942], babespen-[w1940], egokia-[w1941], subj)
arg_info (-, gerta-[w1942], babespen-[w1940], -, -, gaia, -)
arg_info (-, gerta-[w1942], babespen-[w1940], -, -, egoera, -)

Figure 14: The arg_info of the subject of *gertatu*, produced automatically on the basis of the absolutive declension case.

5 The development of the methodology

In this section we will describe the methodology used to tag the EPEC corpus with the corresponding predicate level information. The methodology used was established in three main steps, each composed of several subtasks.

5.1 Preliminary approach

The objective of this phase was twofold: to select the appropriate model for semantic role annotation and to create general annotation guidelines that could serve as the basis for annotating the EPEC corpus.

With this aim three annotators processed 50 instances each of each of the verbs *esan* (“say”, “tell”, “call”), *adierazi* (“explain”) and *eskatu* (“ask for”, “demand”), testing how well they could be modeled by the PB-VN models. These verbs were selected because they appear frequently in the corpus but do not present a high level of complexity in terms of ambiguity (we set aside the analysis of verbs like “to do” and “to be” because they present such a high level of ambiguity and usually appear integrated into complex expressions).

This preliminary work resulted in a set of general guidelines on predicate level labeling for Basque verbs. The guidelines will be constantly updated during the annotation process.

We will use the verb *esan* as an example to illustrate the process the three annotators carried out.

1. They checked the information each verb has in the EADB database. In this case the verb *esan* has associated with it two senses or general predicates:

   1: “to tell somebody to do something“, “to express an idea”, “to narrate or give a detailed account of”,
   experiencer [+human] (ERG); theme [-concrete] (ABS/KONP)

   2: “to assign an attribute/quality to an entity“
   startpoint [+human] (ERG); goal (DAT); attributive (ABS)

2. They found the equivalent verb in English for each sense; here, they could use the mapping we built between Basque and English verbs on the basis of Levin’s

classification, discussed above in section 4.2. In the case of *esan*, possible translations are: “to say”, “ to tell”, “to call”.

3. They chose from the PB-VN resource the roleset associated with the verb sense at hand. Figure 15 shows the description of the above-mentioned verbs in PB-VN:

PB-VN say.01/say-37.7
Arg0: Agent
Arg1: Topic
Arg2: Recipient
Arg3: Attributive

PB-VN tell.01/tell-37.1
Arg0: Agent
Arg1: Topic
Arg2: Recipient

PB-VN call.01/dub-29.3
Arg0: Agent
Arg1: Theme
Arg2: Predicate

Figure 15: the verbs “say”, “tell” and “call” in PB-VN.

4. They annotated the instances based on the information found in PB-VN.

Our experience with this first annotation round validates our previous decision to use the PB-VN model in our annotation process (but see section 2 for a description of some instances where we depart from the PB-VN model).

5.2 Establishing the methodological basis

Manual creation of the Basque Verb Index (BVI) for the verbs contained in EADB database

Once we had selected the PB-VN model as our annotation scheme, we proceeded by tagging the instances of the 100 verbs in our database (EADB) that are examined in depth in (Aldezabal 2004). Our aim was to improve and refine our understanding of the behavior of Basque verbs. In addition, we adapted our tool in such a way that the human annotator would be provided with part of the information contained in the EADB by means of an automatic process.

The goal of this step was to have three human annotators annotate manually a sample set of instances of 97 verbs, leaving the completion of the task to a future automatic process. As a first step, about 120 instances of the verbs were selected and distributed among the annotators; thus, each annotator tagged 40 instances of each verb under study. After the complete annotation of 120 instances of the first 22 verbs, we decided to reduce the number of instances to 20 (about 60 instances in total, since there were three annotators).

This step resulted in a complete set of annotation guidelines (Aldezabal et al. 2010b). In addition, a complete model for the 97 verbs analyzed was manually created (7244 occurrences).

Before proceeding to the annotation task, we wanted to ensure the quality of both the annotations and the guidelines. For that purpose, we carried out an evaluation of the performed task. The next section summarizes the work done (Aldezabal et al. 2011) regarding the evaluation task, emphasizing the main conclusions.

Evaluation: results and conclusions

The evaluation was carried out in two rounds. The aim was to use the conclusions from the first evaluation to make the necessary criteria adjustments, then use these adjusted criteria to annotate other files of the same verbs, and finally evaluate any possible improvements.

In the first step, we first measured the agreement between annotators regarding selecting the English equivalent, because it determines the other properties (argument role, argument number, adjunct role, etc.). Table 4 shows the Cohen's Kappa (Carletta 1996) results:

adierazi	1.000
izan	0.939
etorri	-0.120

Table 4: Cohen's Kappa on selected senses.

In addition, we obtained other data with Cohen's Kappa: the agreement in verb sense and valence (Table 5), and the agreement in verb sense, valence and semantic role (Table 6).

English equivalent + valence	
adierazi	1.000
izan	0.950
etorri	0.232

Table 5: Kappa measures taking into account two variables: the English equivalent and the valence.

English equivalent + valence + role	
adierazi	0.783
izan	0.846
etorri	0.231

Table 6: Kappa measures taking into account three variables: the English equivalent, the valence and the semantic role.

Table 4 shows that, in the case of *adierazi* and *izan*, there was considerable agreement between the two annotators when selecting the sense, and, consequently, the English equivalent. But in the case of *etorri* the Kappa was very low. Moreover, it should be noted that all cases of agreement in *etorri* concerned the first sense; in the other two senses that appeared in the text there was no agreement. This suggested to us that the distinction between the two senses is not clear enough.

Tables 5 and 6 show that when the semantic role is taken into account, the Kappa values of *adierazi* and *izan* decrease slightly. Checking the results by hand, we were able to see that the disagreements occur when assigning a role to the adjuncts.

One conclusion regarding the coverage of the guidelines, then, was that the criteria for assigning a role to the modifier needed to be refined. (Some disagreements, of course, are unavoidable. For instance: in *hitzaldian adierazi* ("express in a speech")*,* one annotator might regard the INE (inessive) phrase as *time* and the other one as *place*).

Multi-lexical units (MLU) were also a source of disagreements. We do not tag verbs as parts of locutions, but this is not always evident. For instance, in the example *Sharonen jarrera probokatzailea zertara datorren galdetu zuen Mubarakek* (Lit. Mubarak asked what Sharon's provocative attitude comes for [has as its purpose]), one annotator considered *zertara etorri* ("come for what") as MLU and the other one did not.

However, the main problem was that although the annotators agreed when selecting the English equivalent, disagreements appeared when tagging other features like the number of the argument and the role. Sometimes one annotator followed the EADB while the other one followed PropBank. Moreover, confusion arose in applying the criteria in the guidelines (derived both from EADB and PropBank).

Confusion was particularly common in the case of the *etorri* verb. For instance, in PropBank "come_01" contains an *Extent* Arg2 that is not possible in Basque. Although the role does not exist for this verb, one annotator continued using the numbered Arg2 for a different role (Arg2: Start point), while the other annotator left aside the argument numbered 2, maintaining the argument-role link of PropBank (Arg3: Start point) [11]. (For more on these types of phenomena, see section 2).

Other disagreements occurred when tagging Arg1. PropBank always assigns the role *Theme* to Arg1, but as discussed in section 2, we have decided not to apply this criterion, so in the unaccusative verb "come.01" we tag the subject as "Arg0, Theme/Cause". However, sometimes one of the annotators relied directly on the PropBank information, resulting in discrepancies between the annotators.

The main conclusion we drew from these problems was that it is crucial to *edit the verb entry completely* before beginning to annotate: one must be clear not only about the English equivalent for the sense but also about the numbered arguments and the assignment of the role. For instance:

1- Change of location

V: *etorri*
VN: come.01
VAL: Arg0, VNrol: Theme, EADBrol: affected theme_ABS
VAL: Arg1, VNrol: Source/path, EADBrol: start location/path_ABL
VAL: Arg2, VNrol: Destination, EADBrol: end location_ALA

2- Creation process

V: *etorri*
VN: come.03 / come.09 (*come out*)
VAL: Arg0, VNrol: Theme, EADBrol: created theme_ABS, SR [12]: -concrete
VAL: Arg1, VNrol: Location, EADBrol: source_ABL, SR: -animate/_DAT, SR: +animate

3- Containing of an entity

V: *etorri*
VN: be.02
VAL: Arg0, VNrol: Theme, EADBrol: content_ABS, SR: -animate
VAL: Arg1, VNrol: Location, EADBrol: container_INE, SR: -animate

4- Description of an entity

V: *etorri*
VN: be.01
VAL: Arg0, VNrol: Topic, EADBrol: theme_ABS
VAL: Arg1, VNrol: Attributive, EADBrol: feature_ABS

Table 7: The *etorri* verb in the BVI.

After applying this principle, the results for the second step – which annotated the same verbs in a number of different files – were much better. Tables 8, 9 and 10 show the same measures after refining the criteria.

adierazi	0.854
izan	0.910
etorri	0.781

Table 8: Cohen's Kappa on selected senses.

English equivalent + valence	
adierazi	0.922
izan	0.930
etorri	0.818

Table 9: Kappa measures taking into account two variables: the English equivalent and the valence.

English equivalent + valence + role	
adierazi	0.808
izan	0.869
etorri	0.704

Table 10: Kappa measures taking into account three variables: the English equivalent, the valence and the role.

After the improvements, then, we achieved a high level of agreement. We can therefore affirm, first, that the PB-VN model serves our purposes, even if we needed to make some adaptations to it, and second, that after applying the improvements made on the basis of the first evaluation (better definition of adjunct role assignment and adjustment of the criteria for applying the PB-V model) the guidelines now have a satisfactory coverage and quality. Furthermore, we conclude that to secure satisfactory results, an essential step in the methodology is to edit each verb entry completely before beginning to annotate its specific instances.

A semi-automatic annotation process applied to the remaining instances of the EADB verbs

The evaluation that we performed corroborated the quality of our manual annotation. Our next step was to annotate automatically the remaining instances of the verbs, drawing on the manually created lexicon and the manual tagging performed on a smaller sample.

We obtained automatically for each verb the set of associated syntactic combinations (see the example in figure 16).

BasqueV, **PropBankV**, **VerbNet role**, **Basque declension case**

```
aldatu:alter_01#change_01 Agent:erg Patient:par NEG:neg
aldatu:alter_01#change_01 Patient:abs NEG:neg
aldatu:alter_01#change_01 Patient:abs TMP:ine
aldatu:alter_01#change_01 Patient:abs ADV:abs
aldatu:alter_01#change_01 Patient:abs MNR:gen
aldatu:alter_01#change_01 Patient:abs LOC:-
aldatu:alter_01#change_01 Patient:abs PRP:helb
aldatu:alter_01#change_01 Agent:erg Patient:abs
```

Figure 16: Syntactic combinations of the “aldatu_alter.01/change.01” verb.

Once we had established the syntactic combinations we could assign the frequency of appearance of each case associated with a concrete semantic role. In this way we obtained the following information (please refer to the verb *aldatu* in figure 16).

Basque declension case, **VerbNet role**, **percentage of occurrences**

Abl	Product	50%
Abl	Material	50%
Abs	Patient	85%
Abs	ADV	7%
Abs	MNR	4%
Abs	TMP	2%
Ala	Product	100%
Aurk	TMP	100%
Bald	DIS	100%
Denb	TMP	100%
Erg	Agent	88%
(…)		

Figure 17: Percentage of the occurrences of Basque declension case and role pair.

The annotation tool was adapted so that for the 100 verbs, the tool automatically offers information about the instances not annotated manually. The tag corresponding to an association between a case and a semantic role was proposed to the human annotators only if that association had a frequency greater than or equal to 50%. In order to facilitate the work of the human annotators, it was also necessary to assign the argument number to each case-role association. Therefore, we developed some heuristics that made use of the manual lexicon to allow us to establish, with a minimal error rate, the argument number for each case-role pair and, in some cases, the link with the PropBank verb. This process facilitated the annotation work substantially: in 70% of the cases the tagging proposed was completely correct, while in the remaining 30%, the annotation, while useful, required some type of correction. The heuristics implemented drew on the results of the manual classification work in which different sets of verbs were identified. Each set is associated with an automatic procedure depending on its semantic features. During the partial manual tagging process, we distinguished four groups of verbs:

- Verbs that have a unique sense and unique equivalent in PB-VN (41%). Figure 18 shows one example: the verb *joan* “go_01” with its corresponding PropBank verb, argument number and semantic role-case association. For verbs of this kind, when annotating the corpus, all fields are proposed automatically on the basis of a combination of the manual lexicon and automatic statistics.

joan_go.01
Arg0: Theme, affected them (ABS)
Arg1: Source/path, point of depart/path (ABL)
Arg2: Destination, end point (ALA)

Figure 18: The "joan_go_01" verb in the BVI.

- Verbs that have a unique sense but multiple equivalents in PB-VN (13%). One example of such verbs is the verb *ikasi* "learn.01 / study.01", shown in Figure 19 with its corresponding PropBank verb, argument number and semantic role-case association. For these verbs, the annotation tool offers all possible equivalents in the first field and the verb is then disambiguated manually based on the sentence context. The remaining fields are assigned automatically on the basis of the manual lexicon.

ikasi_learn.01 / study.01
Arg0: Agent, experiencer [+human] (ERG)
Arg1: Topic, activity (ABS/KONP/INE [13])
Arg2: Source, -, - (ABL)

Figure 19: The "ikasi_learn.01/study.01" verb in the BVI.

- Verbs that have multiple senses, each of which is associated with a unique equivalent (16%). Their treatment is not straightforward. Based on the distinctive declension cases each sense presents, the annotation tool proposes a PropBank verb and its corresponding valency and semantic role-case association. For example, in the verb *izan*, shown in figure 20, the presence of the inessive case in a non-tagged instance of the verb prompts the automatic assignment of the be.02 sense to that instance; in the same way, the case KONP prompts the selection of the be.01 sense and hence also its corresponding PB-VN information.

1)
be.02
Arg1: Theme, theme (ABS)
Arg2: Location, loacation (INE)

be.01
Arg1: Topic, theme (ABS / KONP)
Arg2: Attributive, feature (ABS)

have.03
Arg0: Theme, container (ERG)
Arg1: Theme, content (ABS)

Figure 20: The verb "izan_be.01/be.02/have.03" in the BVI.

- Others. In this category we group the verbs that can not be automatically treated. We distinguish four cases:
  A) Verbs that have a multiple senses, each of which has multiple equivalents in PropBank (10%). Such cases are difficult to treat automatically, and therefore their remaining instances have been tagged manually, with a human annotator

deciding the sense and the PropBank equivalent. Figure 21 shows one example: the verb *eskatu*, which has four senses (only two are shown in the figure), each of which has multiple equivalents in PropBank.

1)
ask.02
Arg0: Agent, esperimentatzailea (ERG)
Arg1: Proposition, gaia (ABS/ELA_KONP)
Arg2: Patient, - (DAT)

order.02
Arg0: Agent, esperimentatzailea(ERG)
Arg1: Theme, gaia (ABS/ELA_KONP)
Arg2: Bneficiary, - (DES)
Arg3: Source, - (DAT)

demand.01
Arg0: Agent, esperimentatzailea (ERG)
Arg1: Proposition, gaia (ABS/ELA_KONP)
Arg2: Patient, - (DAT)

claim.01
Arg0: Agent, esperimentatzailea (ERG)
Arg1: Topic, gaia (ABS/ELA_KONP)
Arg2: Recipient, - (DAT)

2)
require.01
Arg0: Theme, - (ERG)
Arg1: Theme, - (ABS)
Arg2: Source, - (INE)
[...]

Figure 21: The *eskatu* verb in the BVI.

B) Verbs that have multiple senses in Basque and have a unique equivalent in PropBank (4%).
C) Verbs that have two senses in Basque and have a unique sense in PropBank (1%).
D) Verbs that have multiple senses in Basque and multiple equivalents in PropBank or new senses not present in the BVI lexicon (3%).

As is clear from the above, the semi-automatic methods (syntactic frames and lexicon) can be applied in the first three cases, resulting in 70% of verbs being processed semi-automatically and precisely. In the rest of the cases, we have made use of frequent syntactic patterns: if a case / semantic role pair appears in more than 50% of instances, that case / semantic role pair has been automatically assigned and then manually disambiguated.

Enrichment of the BVI by means of automatic tagging

The work described above has resulted in the enrichment of the information present in the EADB as well as in the creation of a lexicon derived from the tagging of the first instances.

In particular, our work has resulted in the addition of new senses and new correspondences to PB-VN to these resources. In total, we have processed 97 verbs representing 143 senses. Furthermore, our use of the automatic process which proposed a tag to the annotator based on frequent (50% or more) association between a case and a semantic role substantially augmented the BVI. Compared to the manually compiled version, the enhanced BVI contained 8,32% more roles and 23,66% more cases.

Tagging Verbs not contained in the EADB on the basis of Levin´s (1993) classification

To assist in the annotation of verbs present in the EPEC corpus but not studied previously, we decided to implement several automatic programs. First, we decided to make use of Levin's classification (Levin 1993). Starting with the idea that verbs belonging to the same Levin class would behave similarly in relation to valency and semantic role-case pairs, we associated verbs annotated in the previous step with verbs belonging to the same Levin class but not annotated previously. This was possible since we already had the Levin class of all verbs in the EPEC corpus (Aldezabal 2010). Figure 22 shows a sample of this study; each entry contains: i) the verbs tagged in the previous phase (third column); ii) its corresponding Levin class (second column) and, iii) the list of yet-unprocessed Basque equivalents to the English verbs present in that Levin class (first column).

irabazi (carry)	11.4	*jaso, eraman*
irabazi (earn, win)	13.5.1	*eskatu, lortu, iritsi, topatu, eraman, jaso, ulertu, hartu, hautatu, ekarri, aurkitu*
jakin (know)	29.5	*adierazi, asmatu, onartu*
utzi (accept)	13.5.2	*eskatu, atera, jaso, hartu, hautatu, onartu*
utzi (admit, allow)	29.5	*adierazi, asmatu, onartu*
utzi (cease)	55.1	*amaitu, hasi*
utzi (leave)	13.4.1	*eman, hornitu*
utzi (leave)	13.5.1	*eskatu, lortu, iritsi, topatu, eraman, jaso, ulertu, hartu, hautatu, ekarri, aurkitu*
utzi (leave)	13.3	*egokitu, atera, eman, eskaini, hautatu, onartu*
utzi (relinquish)	13.2	*aldatu, eman*
ezagutu (recognize, spot)	30.2	*ikusi*
(...)		

Figure 22: Annotated verbs and non-annotated verbs belonging to the same Levin class [14].

Thus, we now had a list of verbs that were not yet processed and that shared a Levin class with one or more of the first 100 tagged verbs. For example, the class that contains *jaso* and *eraman* also contains that *irabazi*, "carry", (11.24). In this way we identified 97 verbs.

We analyzed these verbs and decided to apply automatic processing to those verbs that had only one sense in the tagged part and were associated with a unique PB-VN model. In such cases the model of the tagged verb was automatically assigned to all the instances of the untagged verb based on the BVI and the results automatically obtained. In this way 27 verbs were automatically tagged (28%).

The rest of the verbs were annotated manually following the final methodology, discussed in section 5.3.

This experiment led us to conclude that the Levin's classification we have for Basque is too limited to offer automatic procedures for annotating new verbs and corpora. As a consequence, we developed a methodology that, we find, optimally combines manual work with automatic methods, as described below.

5.3 The final methodology and its application to the rest of the verbs

The methodology for annotation applied so far give us a number of cues as to how to proceed with tagging the remaining verbs, demonstrating: i) the usefulness of the definition of BVI; ii) the usefulness of implementing heuristics to enrich the BVI and, iii) the need for automatic processes to facilitate the annotation task.

Concretely, the steps we propose are the following (See Figure 23):

- Select the verbs to be annotated.
- Define a preliminary lexicon in the PB/VN style.
- Manually annotate some instances of the selected verbs.
- Derive syntactic/semantic patterns from the annotated corpora thus compiled.
- Manually enrich the preliminary lexicon.
- Carry out a semi-automatic annotation of the rest of the instances, based on both the enriched lexicon and the syntactic patterns data.
- Finally, revise manually.

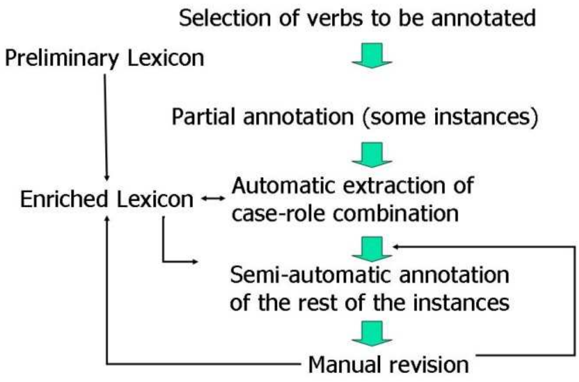
Selection of verbs to be annotated
Preliminary Lexicon
Partial annotation (some instances)
Enriched Lexicon
Automatic extraction of case-role combination
Semi-automatic annotation of the rest of the instances
Manual revision

Figure 23. Steps proposed in the final methodology.

We will apply the methodology described above to the annotation of the remaining verbs, proceeding from the most frequent verbs to rarer ones.

6 A snapshot: the work team, time spent, and data developed

The table below (table 11) shows the data developed up to the present, the time employed, and the people involved, step by step:

Step 1: Verbs tagged in the preliminary approach.
Step 2: Verbs tagged when setting the methodology basis (manually).
Step 2.1: Verbs tagged when setting the methodology basis (evaluation).
Step 2.2: Verbs tagged when setting the methodology basis (semi-automatic).
Step 2.3: Verbs tagged when analyzing the usefulness of Levin classes (semi-automatic).
Step 3.1: Verbs in process of tagging at present with more than 30 occurrences.
Step 3.2: Untagged verbs with less that 30 occurrences.

	Person	Instances	Verbs	Full Corpus %	Tagged	Time [15]	Tagged Corpus %
Step 1	3	1007	3	3,18	150	11,53h	0,47
Step 2	3	19259	97	60,87	7244	557,23h	22,89
Step 2.1	2	5017	3	15,85	350	26,92h	1,10
Step 2.2	2	19259	97	60,87	12015	924,23h	37,97
Step 2.3	1	1866	97	5,89	1845	141,92h	5,83
Step 3.1	1	5715	76	18,06	1239	95,30h	3,91
Step 3.2	1	4799	1187	%15,16	0	0	0
Total		31639	1457	%100	**22343**	1718,69h	**70,60**

Table 11: Data related to the annotation in 05/11/2012.

It must be noted that the data presented in the table only concern the annotation task. We do not include the time and personnel involved in earlier phases like editing the entries, setting up the annotation criteria, creating the guidelines or preparing the tool for the annotation task. Nor do we include the time spent in carrying out all the automatic processes or in reediting the verb's entries. The project has required a minimum of one linguist supervising all linguistic tasks and one computer scientist carrying out all technical aspects. In total, the work carried out up to the present has taken two and a half years. That work has covered studying the behavior of 246 verbs, including these verbs in the BVI lexicon, and tagging 22343 sentences, corresponding to 70,60% of the EPEC corpus.

7 Conclusions and future work

We have presented a semi-automatic methodology for the predicate labeling of EPEC corpus, a methodology that we have tested and whose efficiency in achieving our goals we have ascertained. In parallel with developing this methodology, we have also created two important resources for the computational semantic processing of Basque (BVI and EPEC-RolSem); these resources can be consulted at:
http://ixa2.si.ehu.es/~sisfetek/rolsem_lexicon/galdera.php?adi_eu=&adi_en=&alda=&submit=Bilatu.

The 246 verbs we have processed correspond to those verbs that occur more than 30 times in the EPEC corpus; 70,60% of the sentences in the corpus include one of these verbs. In other words, we have achieved substantial coverage and completed the main portion of the work.

Through the creation of the Basque Verb Index (BVI), our work has also resulted in direct access to PropBank, VerbNet, WordNet and FrameNet information for the verbs processed so far; this will significantly facilitate work on those resources.

The annotation of the EPEC corpus and the creation of BVI verb lexicon opens up some new lines of investigation on related themes.

First, we plan to carry out further study of the verbs appearing in Multiword Lexical Units (MWLU) or Multiword Expression (MWE). When analyzing the verbs in the corpus, we have noted that they display special behavior when they are part of a MWLU or MWE. While verbs can usually express one or more general predicates, the sense or the syntactic behavior of verbs incorporated in a MWLU or MWE changes regarding these general predicates. The study of the changes in the roles in such cases is an interesting issue.

Second, we would like to test the usefulness of our lexicon in specialized corpora. Again, the corpus has shown us that verbs behave differently depending on the type of the text. For instance, newspaper text may only include a particular sense of a verb, or to exhibit special uses or senses of a verb (in, say, sports reporting). It would be very interesting to examine these distinctive verb behaviors and to use them to enrich our lexicon and help organize it in a linguistically coherent way.

Acknowledgements

This research has been supported by the University of the Basque Country (GIU09/19), the Basque Government (IXA group, Research Group of type A (2010-2015) (IT344-10), EUS-SRL project (S-PE11UN098) and Berbatek project (IE09-262)) and The Ministry of Science and Innovation of the Spanish Government (EPEC-RolSem project (FFI2008-02805-E/FILO) (Complementary Action) and AncoraNet project (FFI2009-06497-E)).

Endnotes

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

[2] 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.

[3] ERG: ergative declension case; ABS: absolutive declension case; KONP: completive clause.

[4] DAT: dative declension case.

[5] -ri buruz: a complex declension case (‘about’).

[6] ABL: ablative declension case; ALA: allative declension case.

[7] In Basque the verb *esan* admits all the arguments and syntactic variations of both “tell” and “say” verbs.

[8] *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.

[9] To Argentina (PP)

[10] We mark cases where the value is either too ambiguous or unnecessary to define with the null mark (“-”).

[11] It should be noted that the *Extent* argument is marked “rare” in PropBank, indicating that it is not a common argument in English either.

[12] SR: Selectional Restriction.

[13] INE: Inessive declension case.

[14] We have to take into account that we only possess a reference to the equivalent verb, not to the specific sense of that verb.

[15] 13 occurrences per hour are tagged.

References

Aduriz, I., M. J. Aranzabe, J. M. Arriola, A. Atutxa, A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, M. Oronoz, A. Soroa and R. Urizar. 2006. Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In Andrew Wilson, Paul Rayson and Dawn Archer (eds.), *Corpus Linguistics Around the World*. Book series: Language and Computers. Vol. 56, 1-15. Rodopi (Netherlands).

Agirre, E., I. Aldezabal, J. Etxeberria and E. Pociello. 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. 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., M. J. Aranzabe, J. M. Arriola and A. Díaz de Ilarraza. 2009. Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): Theoretical and practical issues. *Corpus Linguistics and Linguistic Theory* 5-2, 241-269. Mouton de Gruyter. Berlin-New York.

Aldezabal, I., M. J. Aranzabe, A. Díaz de Ilarraza and A. Estarrona. 2010a. Building the Basque PropBank. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.), *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, 1414-1417, European Language Resources Association (ELRA), Valletta (Malta).

Aldezabal, I., M. J. Aranzabe, A. Díaz de Ilarraza, A. Estarrona, K. Fernández and L. Uria. 2010b. 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.

Aldezabal, I., M. J. Aranzabe, A. Díaz de Ilarraza, A. Estarrona and L. Uria. 2010c. EusPropBank: Integrating Semantic Information in the Basque Dependency Treebank. In Alexander Gelbukh (ed.), *Lecture Notes in Computer Science (LNCS) nº 6008, Computational Linguistics and Intelligent Text Processing*, 60-73, Springer, Berlin-Heidelberg-New York.

Aldezabal, I. 2010. Basis for the annotation of EPEC-RolSem. *Interdisciplinary WorkShop on Verbs. The identification and Representation of Verb Features.* Scuola Normale Superiore - Laboratori di Linguistica. pp. 92-97. Universitá di Pisa, Dipartamento di Linguistica. Pisa (Italia).

Aldezabal I., M. Aranzabe, A. Díaz de Ilarraza, A. Estarrona. 2011 Preliminary evaluation of EPEC-RolSem, a Basque corpus labelled at predicate level *SEPLN 2011. ISSN: 1135-5948. Nº 47.*

Aparicio, J. 2007. *Clasificación semántica de los predicados en español.* Masters thesis, Universitat de Barcelona.

Artola, X., A. Díaz de Ilarraza, A. Soroa, A. Sologaistoa. 2009. Dealing with Complex Linguistic Annotations within a Language Processing Frameword. *IEEE Transactions on Audio, Speech, and Language Processing. Vol 17, number 5. Pages 904-915. ISSN: 1558-7916*

Babko-Malaya, O. 2005. PropBank Annotation Guidelines. In http://verbs.colorado.edu/~mpalmer/projects/ace/PBguidelines.pdf

Babko-Malaya, O., A. Bies, A. Taylor, S. Yi, M. Palmer, M. Marcus, S. Kulick and L. Shen. 2006. *Issues in Synchronizing the English Treebank and PropBank* Frontiers in Linguistically Annotated Corpora, A Merged Workshop with 7th International Workshop on Linguistically Interpreted Corpora (LINC-2006) and Frontiers in Corpus Annotation III, Coling/ACL. Sydney, Australia.

Begoetxea K., Gojenola K. 2007. Desarrollo de un analizador sintáctico estadístico basado en dependencia para el euskera. *Revista del procesamiento del lenguaje natural*, nº 39; ISSN: 1135-5948. Pages 5-12.

Bhatt R., B. Narasimhan, M. Palmer, O. Rambow, D. Sharma and F. Xia. 2009. A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu, *In the Proceedings of the Third Linguistic Annotation Workshop*, held in conjunction with ACL-IJCNLP 2009, Singapore.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2): 249–254.

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

Díaz de Ilarraza, A., A. Garmendia and M. Oronoz. 2004. Abar-Hitz An Annotation Tool for the Basque Dependency Treebank, *Language Resources and Evaluation Conference (LREC 2004),* Lisgon, Portugal.

García-Miguel, JM. and FJ. Albertuz. 2005. Verbs, Semantic Classes and Semantic Roles in the ADESSE project. in Erk, Katrin; Alissa Melinger & Sabine Schulte im Walde (eds): *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes* , Saarbrücken.

Hajic, J., J. Panevová, Z. Urešová, A. Bémová, V. Kolárová and P. Pajas. 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. and M. Palmer. 2002. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002).* Las Palmas, Spain.

Kingsbury, P. and M. Palmer. 2003. PropBank: The Next Level of Treebank. In *Proceedings of Treebanks and Lexical Theories*. Växjö, Sweden.

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

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.

Merlo P. and L. & Van der Plas. 2009. Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, pages 288–296, Suntec, Singapore, 2-7 ACL and AFNLP

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

Palmer, M., X. Nianwen, O. Babko-Malaya, J. Chen, B. Snyder. 2005b. A Parallel Proposition Bank II for Chinese and English. *Frontiers in Corpus Annotation, Workshop in conjunction with ACL-05.* Ann Arbor, MI: 2005.

Palmer, M., S. Ryu, J. Choi, S. Yoon and Y. Jeon. 2006. *Korean PropBank.* Linguistic Data Consortium, Philadelphia. LDC2006T03. ISBN: 1-58563-374-7.

Palmer, M., O. Babko-Malaya, A. Bies, M. Diab, M. Maamouri, A. Mansouri, W. Zaghouani. 2008. A Pilot Arabic PropBank. In *Proccedings of LREC 2008.* Marrakech, Morocco.

Pociello, E., E. Agirre and I. Aldezabal. 2010. Methodology and Construction of the Basque WordNet. *Language Resources and Evaluation (LRE)* Journal.

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

Xue, N. and M. Palmer. 2009. Adding semantic roles to the Chinese Trrebank. *Natural Language Engineering* 15, 1 (Jan.), 143-172.

Vázquez G., A. Fernández and M.A. Martí. 2000. *Clasificación verbal. Alternancias de diátesis.* Quaderns de Sintagma 3. Edicions de la Universitat de Lleida.

Vázquez, G., L. Alonso, J.A. Capilla, I. Castellón, A. Fernández (2006). "SenSem: sentidos verbales, semántica oracional y anotación de corpus", *Procesamiento del Lenguaje Natural*, 37, pp. 113-120.