Digital Scholarship in the Humanities Advance Access published June 17, 2015

A Methodology for the Semiautomatic Annotation of EPEC-RolSem, a Basque Corpus Labeled at Predicate Level following the PropBank-VerbNet Model

Ainara Estarrona, Izaskun Aldezabal, Arantza Díaz de Ilarraza and María Jesús Aranzabe
IXA NLP Group, University of the Basque Country, UPV/EHU, Spain

Abstract

In this article we describe the methodology developed for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labeled at predicate level that follows 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 with 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, FrameNet, 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.

Ainara Estarrona Ibarloza
Faculty of InformaticsP°
Manuel Lardizabal, 1 - 20018
Donostia-San Sebastián,
Spain.
E-mail:
izaskun.aldezabal@ehu.eus

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 developing for 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

doi:10.1093/llc/fqv001

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 (BDT) (Aranzabe, 2008; Aldezabal *et al.*, 2009)), and at semantic level, so far, the nouns have been tagged by means of Basque WordNet senses (Pociello *et al.*, 2011). 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 it with work conducted for other languages that also builds lexicons from tagged corpora. For instance, PropBank (Palmer *et al.*, 2005a),[3] related to the VerbNet lexicon (Kingsbury and Palmer, 2002; Kipper, 2005), or PDT (Hajic, 1998), related to the Vallex lexicon (Hajic *et al.*, 2003). Other projects head in the same direction, such as FrameNet (Baker *et al.*, 1998) for many languages, ADESSE (García-Miguel and Albertuz, 2005) for Spanish, SENSEM (Castellón *et al.*, 2006; Vázquez *et al.*, 2006) for both Catalan and Spanish, and AnCora (Aparicio *et al.*, 2008; Taulé *et al.*, 2008) also for both Catalan and Spanish, following the PropBank model. These types 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 decisions have to be made when engaging in corpus annotation: (1) what model to use for annotation, (2) what methodology and guidelines to employ in applying the model, and (3) 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,b). 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: Hindi (Bhatt *et al.*, 2009), Chinese (Palmer *et al.*, 2005b; Xue, 2008; Xue and Palmer, 2009), Korean (Palmer *et al.*, 2006), Arabic (Palmer *et al.*, 2008), Spanish (Taulé *et al.*, 2006; Aparicio, 2007), Catalonian (Civit *et al.*, 2005; Taulé *et al.*, 2006), French (Gardent and Cerisara, 2010; Van Der Plas *et al.*, 2010), and Dutch (Monachesi *et al.*, 2007) and (3) the organization of the lexicon is similar to our in-house database with syntactic/semantic subcategorization frames for Basque verbs ('EADB'), proposed in (Aldezabal, 2004) (see Section 5.1.1). 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.*, 2010c). 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 in Section 4.

This work has resulted in the development of two important resources for the computational semantic processing of Basque: (1) 'BVI', a verb lexicon that currently contains 244 verbs and their predicate information and (2) '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 in Section 3 we consider some language-specific problems when adapting it to Basque. In Section 4 we explain the semantic tag ('arg_info'[4]) used when tagging the verb complements. Section 5 explains the resources we have based our work on and the preprocess we have

applied. In Section 6 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 7 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 8, we consider some potential future lines of investigation.

2 Basic Considerations When Applying the PB-VN Model

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 deficiencies in its coverage of linguistic phenomena. Thus, even after carrying out the studies required to resolve the question of the most appropriate predicate annotating model (Agirre *et al.*, 2006; Aldezabal *et al.*, 2010a,b), we still faced the challenge of solving several model problems of our annotation task. In this section we describe our experience and the consequent decisions regarding the model-internal problems.

The PropBank model (Palmer *et al.*, 2005a) distinguishes between two independent levels: (1) the level of arguments and adjuncts, and (2) 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'), and these are linked to the VerbNet lexicon (Kipper *et al.*, 2000, 2008; Kipper, 2005), which in turn has general roles (e.g. agent, theme). 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).

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

PropBank tell.01	VerbNet tell-37.1
Arg0: speaker	Agent
Arg1: utterance	Topic
Arg2: hearer	Recipiendt

We see that PropBank and VerbNet offer complementary information, as observed by Merlo and Van der Plas (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 in Section 2.2.

2.1 Regarding Argo 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, p. 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' (Kingsbury and Palmer, 2003, p. 3).

Nevertheless, inconsistencies abound. For instance, Babko-Malaya *et al.* (2006, p. 76) report: 'In *John and Mary come* the NP *John and Mary* is a constituent in Treebank and it is also marked as

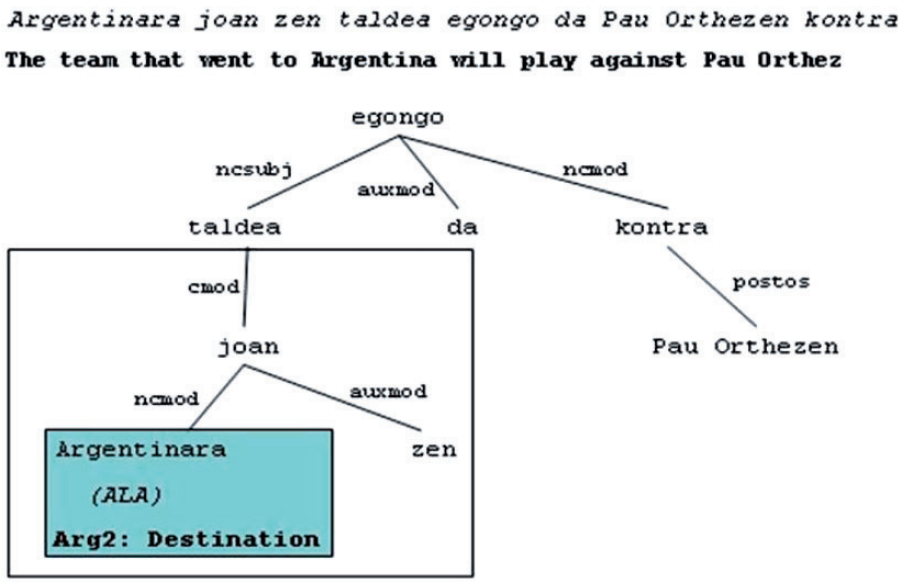

Fig. 1 The dependency tree for the sentence 'The team that went to Argentina will play against Pau Orthez'.

''Arg0'' in PropBank'. But when we check it in PropBank we realize that the verb *come.01* is defined as we can see in Table 2:

Table 2 The verb 'come.01' in PropBank

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

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, p. 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, p. 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 'to 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, 2009), Korean (Palmer *et al.*, 2006), Chinese (Xue and Palmer, 2009), and Spanish (Aparicio, 2007).

In other models (for instance, in the case of Spanish, Semsen (Vázquez *et al.*, 2006), and ADESSE (García-Miguel and Albertuz, 2005)) this particular problem does not arise because these models do not use numbered arguments (Semsen) or they apply different criteria (ADESSE).

2.2 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,[5] in our current work we have decided to follow PropBank broadly, 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* ('to begin'). *Hasi* is linked to the PB-VN verbs shown in Table 3.

In Table 3 we can see that the three verbs that can be equivalents for the Basque *hasi* ('to begin') have an instrument argument in PropBank, whereas in VerbNet such an argument is not defined.[6] PropBank offers some examples for the use of Arg2_Instrument (example 1):

(1) *John started the book with a murder.*
Arg0: John
Rel: started
Arg1: the book
Arg2: with a murder

In the EADB this instrument 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. Example 2 shows this instrument argument in the EPEC corpus:

(2) *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'): Arg2_Instrument.

On the other hand, in the case of the Basque *adierazi* ('to state', 'to express'), linked to 'state.01' in PB-VN, we have followed VerbNet. Table 4 shows the description of the verb 'state.01' in PB-VN:

The Arg3 proposed in PropBank has no equivalent in VerbNet. Also, in the only example found in PropBank there is no Arg3 (example 3):

(3) *The Japanese government, Mr. Godown said, has stated that it wants 10–11% of its gross national product to come from biotechnology products.*
Arg0: The Japanese government
Rel: stated

Table 3 The verbs 'begin.01', 'start.01', and 'commence.01' in PB-VN

begin.01
Arg0: beginner, agent (vnrole: 55.1 - Agent)
Arg1: theme (creation) (vnrole: 55.1 - Theme)
Arg2: instrument
start.01
Arg0: agent (vnrole: 55.1 - Agent)
Arg1: theme (creation) (vnrole: 55.1 - Theme)
Arg2: instrument
commence.01
Arg0: beginner, Agent (vnrole: 55.1 - Agent)
Arg1: theme (creation) (vnrole: 55.1 - Theme)
Arg2: instrument

Table 4 The verb 'state.01' in PB

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

> Arg1: that it wants 10–11% 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 ('to state'), because (1) this verb has only one sense so the fourth argument does not help in distinguishing senses and (2) in the only example that appears in PropBank there is no Arg3.

In the same way, in the case of the *esan* verb ('to say', similar in sense to 'to state') we find an Arg3_Attributive in PropBank (Table 5) that does not appear in VerbNet (nor in the EADB). However, in this verb the Arg3_Attributive marks the difference between senses in Basque:

Table 5 The verb 'say.01' in PB

say.01
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

The verb *esan* has two senses in the EADB. The first one would be the equivalent of the English verb 'to say' and the second one the equivalent of the English verb 'to call':

(1) Communication action; two arguments in two syntactic variants:
   (a) experiencer [+human] (ERG[7]), theme [-concrete] (ABS)
   (b) experiencer [+human] (ERG), theme (KONP)

(2) Assigment of an attribute/quality to an entity; three arguments in a single syntactic realization:
   (c) startpoint [+human] (ERG), goal (DAT[8]), attributive (ABS)

The Arg3 proposed by PropBank for the verb 'to say' is possible in the first sense, but not in the second one. That is, although it is not a frequent argument and even when it does appear, seems to be an adjunct, unlike in the previous case ('state.01'), it distinguishes between senses. As a consequence, agreeing with PropBank, we regard it as Arg3_Attributive. Thus, we define the *esan* ('to say') verb in the PB-VN style as follows (Table 6):

Table 6 The verb *esan_1* in the BVI lexicon

Arg0: Agent, experiencer [+human] (ERG)
Arg1: Topic, theme [-concrete] (ABS/KONP)
Arg2: Recipient, - (DAT)
Arg3: Attributive, - (*-ri buruz*[9])

2.3 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' shown in Table 7:

Table 7 The verb 'see.01' in PB

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

Arg0 has associated Agent and Experiencer roles and Arg1 associated Theme and Stimulus roles. By contrast, in the EADB the verb *ikusi* ('to see') 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* ('experiencer')
- Arg1: Theme, *gaia* ('theme')

2.4 ADV role

There is an ADV role for adjuncts in the PB-VN 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.

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

2.5 Including *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* ('to pass' / 'to come by') we find examples like:

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

In the lastest version of VN (Version 3.2) there are some roles that have been changed in order to better conform the list of roles used in VN to the standard list of roles proposed by the LIRICS project (Bunt and Romary, 2002; Bunt *et al.*, 2007; Schiffrin and Bunt, 2007). Thus, VN now contains a Trajectory role equivalent to our Path role. In the same way, we have seen that some Theme1 roles have been changed to Pivot (for instance, in the verb class own-100). A move to the latest version of VN could therefore require some revisions in our judgments. At the moment, however, we mantain the same roles as they were before the changes in VN 3.2 version.

3 Interlingual Differences. Criteria for Applying the Model to Basque

Applying the PB-VN model to Basque is mainly a question of including the distribution of the arguments and adjuncts in a verb sense as well as the roles proposed for them. For example, in the EADB (which will be deeply explained in Section 5.1.1) the Basque verb *eskatu* ('to ask for') has two arguments, Arg0: *Esperimentatzailea* (Experiencer) and Arg1: *Gaia* (Theme). The dative complement is not included within the subcategorized cases because attending to the fact that we can 'ask for' something, in general, without saying explicitly the 'goal' as an impersonal proposition (alternation). However, the verb 'ask.02' contains three 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 which we explain below.

3.1 Arguments proposed in 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') (Table 8):

Table 8 The verb *go.01* in PB-VN

	go.01
Arg1	entity in motion / 'goer' (theme)
Arg2	**extent**
Arg3	start point (source)
Arg4	end point (destination)

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, as we have explained in Section 2.1) and the 'end point' will be Arg2.

After these changes, the resulting entry in BVI (Table 9) is the same as in the EADB (example 6).

(6) (1) affected theme_ABS; start point_ABL[10]; end point_ALA
(2) affected theme_ABS; start point [+animate]_DAT; end point_ALA

Table 9 The verb *joan_go.01* in BVI lexicon

joan_1/go.01		
Arg0	theme	affected theme (ABS)
Arg1	source	start point (ABL/DAT)
Arg2	destination	end point (ALA)

3.2 More than one PropBank verb existis 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* ('to say') 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 the first value of the *arg_info* tag (see Section 4).

In other cases, although the English verbs are the same at predicate level ('make.01', 'build.01', 'construct.01'), we annotate the concrete instances with the one we consider the most 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* and its equivalents in English. Examples are provided in (7):

(7)
- *Kanta asko egin zituen* ('He/she composed a lot of songs'): compose.02 (agent, product, beneficiary)
- *Ondoko galdera egin diote Juan Jose Ibarretxeri* ('They have asked this question to Juan Jose Ibarretxe'): ask.02 (agent, topic, recipient)
- *Boticak 27 puntu egin zituen* ('Botica had scored 27 points'): score.01 (agent, product, beneficiary)
- *Biek ere joko alaiegia egiten zuten ACBrako* ('Both of them practiced a too happy-go-lucky playing style for the ACB'): practice.01 (agent, theme, instrument)

4 The Tag for Predicate Labeling: ARG_INFO

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 (BDT (Aranzabe, 2008; Aldezabal *et al.*, 2009)), and partially at semantic level, where nouns were tagged by means of Basque WordNet senses (Pociello *et al.*, 2011). 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 *arg_info* tag, 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 1, the *arg_info* tag will be assigned to the ncsubj ('the team') and two ncmods ('to Argentina' and 'against Pau Orthez') linked to the verb (the head).

The *arg_info* tag 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 1 shows in tree format a compound sentence annotated syntactically, where semantic annotation has been added to the phrase in the allative case (ALA) linked to the verb *joan* ('to 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* ('to go'). Syntactic dependencies[11] are marked on the links, and the semantic information on the nodes. The declension case is included in the nodes as additional information.

Here we have the dependency tagging corresponding to the example in Figure 1:

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

Example 8 illustrates the *arg_info* tag that corresponds to the ncmod *Argentinara* ('to Argentina' (PP)) in Figure 1.

(8) arg_info: (go.01, joan, Argentinara, Arg2, Destination, end_location, -[12])

5 Some Basic Resources and Preprocesses

Before discussing the methodology we use, we will briefly describe the resources (Section 5.1) on which we based the project as well as the automatic procedures (Section 5.2) that we have been able to employ to facilitate the tagging task.

5.1 Basic resources

In this section we will describe two basic resources we have used to carry out the annotation task.

5.1.1 *The EADB (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 (equivalent to a frame) pertains to a verbal sense (similar to a roleset).

Aldezabal (2004) defined a specific inventory of semantic roles; the set of semantic roles associated with a verb identifies its different meanings. The semantic roles specified are: Theme, Affected Theme, Created Theme, State, Location, Time, End Location, End State, Start Location, Path, Start point, 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 ('to go'): entity in motion
  affected theme_ABS; start location/path_ABL; end location_ALA
  affected theme_ABS; start location [+animate]_DAT; end location_ALA
- joan.2 ('to go'): feature that disappears from an entity
  container_DAT; content [-animate, -concrete]_ABS
- joan.3 ('to go'): to assign a feature to an entity
  theme_DAT; feature_ABS

5.1.2 *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 focusing on the similarities in 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

Table 10 Some examples of the links between verbs in Levin's classification 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*

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 10 shows some examples of the links between verbs in (Levin, 1993) and Basque verbs.

5.2 The automatic preprocess

In this section we will describe the main automatic preprocess we have performed to facilitate the tagging task.

5.2.1 *Comparison of the Levin classes in our mapping with the PropBank database*

Drawing on the resources described above, we carried out an automatic preprocess in which two tasks were automated:

(1) 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).
(2) Some of the information contained in the EADB has been linked to the EPEC corpus.

More detailed descriptions of these two tasks follow.

(1) 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 glue' and 'to go' remained in classes 22.4 and 47.7, respectively. This category represented 74.92% of the cases.[13]
- 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 11 shows a sample of the results of the comparison between the classes in (Levin, 1993) and the classes in the current PB-VN data.

Verbs falling into the first and second categories (84.38%) were linked to PB-VN and their information displayed in the AbarHitz annotation tool.

(2) 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, while arguments with nonambiguous declension cases are automatically annotated, ambiguous cases must be

Table 11 The link between verbs in Levin's classification and Basque

Levin' verbs	Levin's classes	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

manually disambiguated by the annotator. The annotator can, however, draw on an automatically generated proposal that contains all the possible tags.

In (9) we can see an example of a nonambiguous case, *adierazi* ('to state'). The EADB includes the following information for the *adierazi* ('to state') verb:

(9) (a) experiencer_ERG; theme [-animate; -concrete]_ABS
(b) experiencer_ERG; theme [-animate; -concrete]_KONP

On the basis of the *-ela* subordinating conjunction and the ergative declension case (example 10), the preprocessing tool will prepare the *arg_info* tags for the subordinating clause 'that Israeli helicopters bombarded the Palestinian area' and for the subject 'the witnesses' that we can see in Table 12.

(10) *Israelgo helikopteroek gune palestinarrak bonbardatu zituztela adierazi zuten lekukoek.* 'The witnesses stated that Israeli helicopters bombarded the Palestinian area'.

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

By contrast, *gertatu* ('to happen', 'to be') 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:

2 *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 (see example 11 and Table 13):

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

So the annotator must decide the verb sense corresponding to the instance and consecuently, the role assigned to the argument.

Table 12 The *arg_info* of the subordinating clause and subject of *adierazi* ('to state') produced automatically on the basis of the *-ela* subordination conjunction and the *-k* ergative declension case

ccomp_obj (konpl, adierazi, bonbardatu, zituztela)
arg_info ((-, adierazi, zituztela, -, -, theme, -human/-concrete)
ncsubj (**erg**, adierazi, lekukoek, lekukoek, subj)
arg_info (-, adierazi, lekukoek, -, -, experiencer, -))

Table 13 The *arg_info* of the subject of *gertatu* ('to be'), produced automatically on the basis of the absolutive declension case

ncsubj (**abs**, gerta, babespen, egokia, subj)
arg_info (-, gerta, babespen, -, -, gaia ('theme'), -)
arg_info (-, gerta, babespen, -, -, egoera ('state'), -)

6 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

had three main steps, each composed of several subtasks:

(1) Preliminary approach.
(2) Design of the methodological basis.
(3) Final methodology and its application on the rest of the verbs.

6.1 Preliminary approach

The objective of this phase was two-fold: (1) to select the appropriate model for semantic role annotation and (2) 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 the verbs *esan* ('to say', 'to tell', 'to call'), *adierazi* ('to explain', 'to state'), and *eskatu* ('to ask for', 'to 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 *egin* 'to do' and *izan* 'to be' because they present 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 are 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) The information each verb has in the EADB database was checked. In this case the verb *esan* has two associated senses or general predicates:
   (a) '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)
   (b) 'to assign an attribute/quality to an entity' startpoint [+human] (ERG); goal (DAT); attributive (ABS)
(2) The annotators 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 in Section 5.2.1. In the case of *esan*, possible translations are: 'to say', 'to tell', 'to call'.
(3) The annotators chose from the PB-VN resource the roleset associated with the actual verb sense at hand. Table 14 shows the description of the above-mentioned verbs in PB-VN:

Table 14 The verbs 'say', 'tell', and 'call' in PB-VN

To express an idea		**To assign an attribute**
say.01	**tell.01**	**call.01**
vncls: say-37.7	vncls: tell-37.1	vncls: dub-29.3
Arg0: agent	Arg0: agent	Arg0: agent
Arg1: topic	Arg1: topic	Arg1: theme
Arg2: recipient	Arg2: recipient	Arg2: predicate
Arg3: attributive		

(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 the annotation process (but see Sections 2 and 3 for a description of some instances where we depart from the PB-VN model).

6.2 Establishing the Methodological Basis

The methodology used to tag the EPEC corpus with the corresponding predicate level information has undergone a process of continuous refinements. In this section we present the different steps we have worked on to establish the most suitable methodology.

6.2.1 *Manual creation of the BVI for the verbs contained in EADB database*

Once we selected the PB-VN model as our annotation scheme, we tagged the instances of the 100 verbs in our database (EADB) that were 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 was 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 annotating 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.*, 2010c). In addition, a complete model for the 97 verbs analyzed was manually created (7,244 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.

6.2.2 *Evaluation: results and conclusions*

The evaluation was carried out in two rounds and with three verbs: *adierazi* ('to state'), *izan* ('to be'), and *etorri* ('to come'). The aim was to use the conclusions from the first evaluation to make the necessary criteria adjustments to then use these adjusted criteria to annotate other files of the same verbs, and finally evaluate any possible improvements.

In the first step, and given that it determines the other properties, we first measured the agreement between annotators regarding selecting the English equivalent (argument role, argument number, adjunct role, etc.). Table 15 shows the Cohen's Kappa (Carletta, 1996) results:

Table 15 Cohen's Kappa on selected senses

adierazi	1.000
izan	0.939
etorri	−0.120

Table 16 Kappa measures taking into account two variables: the English equivalent and the valence

English equivalent + valence	
adierazi	1.000
izan	0.950
etorri	0.232

Table 17 Kappa measures taking into account three variables: the English equivalent, the valence and the semantic role

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

In addition, we obtained other data with Cohen's Kappa: the agreement in verb sense and valence (Table 16), and the agreement in verb sense, valence, and semantic role (Table 17).

Table 15 shows that, in the case of *adierazi* ('to state') and *izan* ('to be'), there was considerable agreement between the two annotators when selecting the sense, and, consequently, the English equivalent. But in the case of *etorri* ('to come') the Kappa was very low. Moreover, it should be noted that all cases of agreement in *etorri* ('to come') concerned the first sense; in the other two senses that appeared in the text there was no agreement. This suggested that the distinction between the two senses is not sufficiently clear.

Tables 16 and 17 show that when the semantic role is taken into account, the Kappa values of *adierazi* ('to state') and *izan* ('to be') decrease slightly. Checking the results by hand, we detect 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' [has as its purpose]) 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 such as the number of the argument and the role. Sometimes one annotator followed the EADB while the other one followed PB-VN. Moreover, confusion arose when applying the criteria in the guidelines (derived both from EADB and PB-VN).

Confusion was particularly common in the case of the verb *etorri* ('to come'). For instance, in PB-VN 'come.01' contains an Arg2_Extent that is not possible in Basque (see Section 3.1). 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 PB-VN (Arg3: Start point).[14]

Other disagreements occurred when tagging Arg1. PropBank always assigns the role Theme to Arg1, but as discussed in Section 2.1, we decided not to apply this criterion, so in the unaccusative verb 'come.01' we tag the subject as Arg0_Theme. However, sometimes one of the annotators relied directly on the PB-VN information which resulted 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, so that English equivalent, the numbered arguments, and the roles assigned are absolutely clear. For instance the verb *etorri* ('to come'):

(1) Change of location

V: etorri
PB-VN: come.01
VAL: Arg0, VNrole: Theme, EADBrole: affected theme_ABS
VAL: Arg1, VNrole: Source/path, EADBrole: start location/path_ABL
VAL: Arg2, VNrole: Destination, EADBrole: end location_ALA

(2) Creation process

V: etorri
PB-VN: come.03 / come.09 (come out)
VAL: Arg0, VNrole: Theme, EADBrole: created theme_ABS, SR[15]: -concrete
VAL: Arg1, VNrole: Location, EADBrole: source_ABL, SR: -animate/_DAT, SR: +animate

(3) Containing of an entity

V: etorri
PB-VN: be.02
VAL: Arg0, VNrole: Theme, EADBrole: content_ABS, SR: -animate
VAL: Arg1, VNrole: Location, EADBrole: container_INE, SR: -animate

(4) Description of an entity

V: etorri
PB-VN: be.01
VAL: Arg0, VNrole: Topic, EADBrole: theme_ABS
VAL: Arg1, VNrole: Attributive, EADBrole: feature_ABS

After applying this principle, the results of the second step—which annotated the same verbs in a number of different files—showed a significant improvement. Tables 18–20 show the same measures after refining the criteria.

Table 18 Cohen's Kappa on selected senses

adierazi	0.854
izan	0.910
etorri	**0.781**

Table 19 Kappa measures taking into account two variables: the English equivalent and the valence

English equivalent + valence	
adierazi	0.922
izan	0.930
etorri	0.818

Table 20 Kappa measures taking into account three variables: the English equivalent, the valence and the semantic role

English equivalent + valence + role	
adierazi	0.808
izan	0.869
etorri	0.740

Table 22 Percentage of the occurrences of Basque declension case and role pair

Case/Role	ERG	ABS	ALA	ABL	denb.[16]
Agent	8 (% 88)				
Patient		35 (% 85)			
Product			1 (% 100)		
Material				1 (% 50)	
TMP					2 (% 100)
MNR		2 (% 4)			
ADV		3 (% 7)			

Table 21 Syntactic-semantic combinations of the '*aldatu*_alter.01/change.01' verb

BasqueV	PropBankV	VerbNet role	Basque declension case
aldatu	alter.01/change.01	Agent-Patient-NEG	erg-par-neg
aldatu	alter.01/change.01	Patient-NEG	abs-neg
aldatu	alter.01/change.01	Patient-TMP	abs-ine
aldatu	alter.01/change.01	Patient-ADV	abs-abs
aldatu	alter.01/change.01	Patient-MNR	abs-gen
aldatu	alter.01/change.01	Patient-LOC	abs-
aldatu	alter.01/change.01	Patient-PRP	abs-helb
aldatu	alter.01/change.01	Agent-Patient	erg-abs

After the improvements, we achieved a high 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-VN model) the guidelines 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.

6.2.3 *A semiautomatic 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 the set of associated syntactic-semantic combinations automatically for each verb (see the example in Table 21).

Once we had established the syntactic-semantic combinations, we could assign the frequency of appearance of each case associated with a concrete semantic role. In this way we obtained the information shown in Table 22 (please refer to the verb *aldatu* ('to change') in Table 21).

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, to establish, with a minimal error rate, the argument number for each case–role pair and, in some cases, the link with the PB-VN verb, we developed several heuristics that made use of the manual lexicon. This process facilitated

Table 23 The '*joan*_go.01' verb in the BVI lexicon

joan_go.01		
Arg0	theme	affected theme (ABS)
Arg1	source/path	point of depart/path (ABL)
Arg2	destination	end point (ALA/ABU)

Table 24 The '*ikasi*_learn.01/study.01' verb in the BVI

ikasi_learn.01/study.01		
Arg0	agent	experiencer [+human] (ERG)
Arg1	topic	activity (ABS/KONP/INE[3])
Arg2	source	- (ABL)

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%). Table 23 shows one example: the verb *joan* ('go.01') with its corresponding PB-VN verb, argument number, and semantic role–case association. For this type of verbs all fields are proposed automatically on the basis of a combination of the manual lexicon and automatic statistics.
- 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 Table 24 with its corresponding PB-VN 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.

Table 25 The verb *izan* in the BVI

izan			
1- izan_be.02	Arg0	theme	theme (ABS)
	Arg1	location	location (INE)
2- izan_be.01	Arg0	topic	theme (ABS/KONP)
	Arg1	Attribute	feature (ABS)
3- izan_have.03	Arg0	theme	container (ERG)
	Arg1	theme	content (ABS)

- 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 PB-VN verb and its corresponding valency and semantic role–case association. For example, in the verb *izan*, shown in Table 25, 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.
- Others. In this category we group the verbs that can not be automatically treated.

We distinguish four cases:

(1) Verbs that have 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 PB-VN equivalent. Table 26 shows one example: the verb *eskatu*, which has four senses (only the first is shown in the table), each of which has multiple equivalents in PB-VN.

(2) Verbs that have multiple senses in Basque and have a unique equivalent in PB-VN (4%).

(3) Verbs that have two senses in Basque and have a unique sense in PB-VN (1%).

Table 26 The first sense of the *eskatu* verb in the BVI

eskatu			
1- eskatu_ask.02	Arg0	agent	experiencer (ERG)
	Arg1	proposition	theme (ABS/KONP)
	Arg2	patient	- (DAT)
1- eskatu_order.02	Arg0	agent	experiencer (ERG)
	Arg1	theme	theme (ABS/KONP)
	Arg2	beneficiary	- (DES[18])
	Arg3	source	- (DAT)
1- eskatu_demand.01	Arg0	agent	experiencer (ERG)
	Arg1	proposition	theme (ABS/KONP)
	Arg2	patient	- (DAT)
1- eskatu_claim.01	Arg0	agent	experiencer (ERG)
	Arg1	topic	theme (ABS/KONP)
	Arg2	recipient	- (DAT)

(4) Verbs that have multiple senses in Basque and multiple equivalents in PB-VN or new senses not present in the BVI lexicon (3%).

As it is clear from the above, the semiautomatic methods (syntactic frames and lexicon) can be applied in the first three cases, resulting in 70% of verbs being processed semiautomatically and correctly. In the rest of the cases, we have used 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.

6.2.4 *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 which correspond to 143 senses. Furthermore, our use of the automatic process which proposed a tag to the annotator based on frequent association between a case and a semantic role (50% or more) substantially augmented the BVI. Compared to the manually compiled version, the enhanced BVI contained 8.32% more roles and 23.66% more cases.

6.2.5 *Tagging verbs not included in the EADB on the basis of Levin's 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 which had not been annotated previously. This was possible since we already had the Levin class of all verbs in the EPEC corpus (Aldezabal, 2010). Table 27 shows a sample of this study; each entry contains: (1) the verbs tagged in the previous phase (third column); (2) its corresponding Levin class (second column), and (3) the list of yet-unprocessed Basque equivalents to the English verbs present in that Levin class (first column).[19]

Thus, we now had a list of verbs that had not yet processed and that shared a Levin class with one or more of the first 97 tagged verbs. For example, the class that contains *jaso* ('to lift') and *eraman* ('to carry'), (11.4), also contains *irabazi* ('to carry'). 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 (the case in bold in Table 27). 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 6.3.

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. Consequently, we developed a methodology that, in our opinion, optimizes the combination of manual work with automatic methods, as described below.

Table 27 Annotated verbs and nonannotated verbs belonging to the same Levin class

Nonannotated verb (English equivalent)	Levin class	Verb annotated
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 (relinguish)	13.2	*aldatu, eman*
***ezagutu* (recognize, spot)**	**30.2**	***ikusi***
(. . .)		

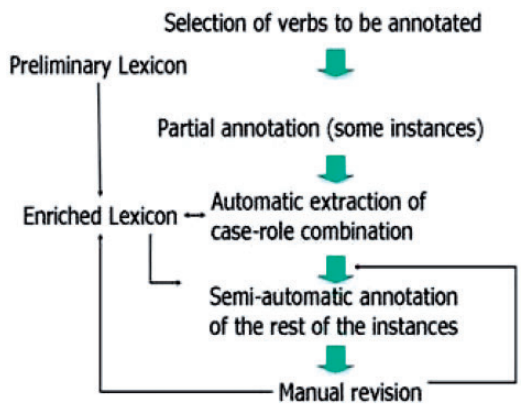

Fig. 2 Steps proposed in the final methodology.

6.3 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 to tag the remaining verbs, demonstrating: (1) the usefulness of the definition of BVI; (2) the usefulness of implementing heuristics to enrich the BVI and, (3) the need for automatic processes to facilitate the annotation task.

Concretely, the steps we propose are the following (see Figure 2):

- 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 compiled.
- Manually enrich the preliminary lexicon.
- Carry out a semiautomatic annotation of the rest of the instances, based on both the enriched lexicon and the syntactic patterns data.
- Finally, revise manually.

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

7 A Snapshot: The Work Team, Time Spent, and Data Developed

Table 28 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 (semiautomatic).

Table 28 Data related to the annotation in 05 November 2012

	Person	Verbs	Instances	Full corpus %	Tagged	Time[20]	Tagged corpus %
1.	3	3	1.007	3,18	150	11,53 h	0,47
2.	3	99	19.259	60,87	7.260	557,23 h	22,89
2.1	2	3	5.017	15,85	350	26,92 h	1,10
2.2	2	99	19.259	60,87	11.849	924,2 3 h	37,97
2.3	1	97	1.866	5,89	1.845	141,92 h	5,83
3.	1	75	5.715	18,06	1.239	95,30 h	3,91
3.1	1	1.186	4.799	15,16	0	0	0
Total		1.457	31.639	100	22.343	1.718,69 h[21]	**70,61**

Step 2.3: Verbs tagged when analyzing the usefulness of Levin classes (semiautomatic).
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.

It must be noted that the data presented in the table only show the annotation task. We do not include the time and personnel involved in earlier phases such as editing the entries, setting up the annotation criteria, creating the guidelines, or preparing the tool for the annotation task. Neither 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 2.5 years and has covered the study of the behavior of 244 verbs, the inclusion of these verbs into the BVI lexicon, and the tagging 22,343 sentences, corresponding to 70.60% of the EPEC corpus.

8 Conclusions and Future Work

We have presented a semiautomatic methodology for the predicate labeling of the EPEC corpus, a methodology that we have tested and whose efficiency in achieving our goals we have proved. 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 by means of the 'e-ROLda' tool (http://ixa2.si.ehu.es/e-rolda/index.php) which provides facilities to request information about the syntactic and semantic structure of verbs as well as examples of use.

At the time of writing, 70.60% of the EPEC corpus has been manually tagged and the 29.4% remaining has been automatically tagged with a SRL system implemented using machine learning techniques trained with the manually tagged subset. In the experiments the classifier that offers the best results is based on Support Vector Machines, 84.30 F1 score in identifying the PropBank semantic role for a given constituent and 82.90 F1 score in identifying the VerbNet role (Salaberri *et al.*, 2014). At present, we are doing the manual revision and evaluation of the automatically tagged sample. This annotation work has resulted in the development of the BVI which currently contains 1,211 verbs: (1) 244 verbs which include the 151 verbs that have more than 30 occurrences with their respective argument structure information (covering 70.60% of the sentences in the corpus) and (2) 967 verbs whose argument structure has been obtained automatically by means of a module that builds new entries from the corpus automatically tagged

Through the creation of the BVI, our work has also resulted in direct access to PropBank, VerbNet, WordNet, and FrameNet information for the verbs processed so far, which will significantly facilitate the use of these resources in future work.

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

First, we plan to carry out a further study of the verbs that appear in Multiword Lexical Units (MWLU) or Multiword Expression (MWE). When analyzing the verbs in the corpus, we have realized 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 worth pursuing.

Second, we would like to test the usefulness of our lexicon in specialized corpora. Again, the corpus has shown that verbs behave differently depending on the type of the text. For instance, newspaper texts 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 particularly interesting to examine these distinctive verb behaviors and to use them to enrich our lexicon and help organize it in a linguistically coherent way.

Funding

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)).

References

Aduriz, I., Aranzabe, M. J., Arriola, J. M., Atutxa, A., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Oronoz, M., Soroa, A., and 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 Wilson, A., Rayson, P., and Archer, D. (eds), *Corpus Linguistics Around the World, volume 56 of Book series: Language and Computers*. Netherlands: Rodopi, pp. 1–15.

Agirre, E., Aldezabal, I., Etxeberria, J., and 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. *Oral presentation.* UMIACS Departamental Colloquia, University of Maryland.

Aldezabal, I. (2004). *Aditz'-azpikategorizazioaren azterketa sintaxi partzialetik sintaxi osorako bidean. 100 aditzen azterketa, Levin'-en (1993) lana oinarri hartuta eta metodo automatikoa baliatuz.* Ph.D. thesis, Euskal Filologia Saila, Euskal Herriko Unibertsitatea.

Aldezabal, I. (2010). Basis for the annotation of EPEC-RolSem. In *Interdisciplinary Workshop on Verbs. The identification and Representation of Verb Features.* Universitá di Pisa, Dipartamento de Linguistica: Pisa, Italy, pp. 92–97.

Aldezabal, I., Aranzabe, M. J., Arriola, J. M., and 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**: 245–74.

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., and Estarrona, A. (2010a). Building the Basque PropBank. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odjik, J., Piperidis, S., Rosner, M., and Tapias, D. (eds), *Proceedings of the Seventh Conference on International Language Resources and Evaluation, LREC 2010.* European Language Resources Association (ELRA); Valletta, Malta, pp. 1414–17.

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., Estarrona, A., and Uria, L. (2010b). EusPropBank: Integrating semantic information in the Basque Dependency Treebank. In Gelbukh, A. (ed.), *Lecture Notes in Computer Science (LNCS) no. 6008. 11th International Conference, CICLing-2010 of Computational Linguistics and Intelligent Text Processing.* Springer, Iasi, Romania, pp. 60–73.

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., Estarrona, A., Fernández, K., and Uria, L. (2010c). 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). Technical report, UPV-EHU, Donostia, Spain.

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., and Estarrona, A. (2011). Preliminary evaluation of EPEC-RolSem, a Basque corpus labelled at predicate level. *Procesamiento del Lenguaje Natural*. Universidad de Huelva, Huelva, Spain, vol. 47, pp. 1–9.

Aparicio, J. (2007). *Clasificación semántica de los predicados en español*. Master's thesis, Universitat de Barcelona.

Aparicio, J., Taulé, M., and Martí, M. A. (2008). AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC)*. Marrakech, Maroko.

Aranzabe, M. J. (2008). *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala*. Ph.D. thesis, Euskal Filologia Saila. Euskal Herriko Unibertsitatea, Donostia.

Artola, X., Díaz de Ilarraza, A., Soroa, A., and Sologaistoa, A. (2009). Dealing with complex linguistic annotations within a Language Processing Frameword. *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, Artola 2009: ISSN:1558-7916 (Digital Publication), pp. 904–915.

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

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

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of COLING-ACL'98* Montreal, Quebec.

Bengoetxea, K. and Gojenola, K. (2007). Desarrollo de un analizador sintáctico estadístico basado en dependencia para el euskera. *Revista del procesamiento del lenguaje natural*, **39**: 5–12.

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

Bunt, H. C., Petukhova, V., and Schiffrin, A. (2007). Lirics deliverable d4.4. multilingual test suites for semantically annotated data. Technical report, http://lirics.loria.fr.

Bunt, H. C. and Romary, L. (2002). Towards multimodal content representation. *International standards of terminology and language resources management, LREC 2002*. Spain: Las Palmas.

Carletta, J. (1996). Assessing agreement on classification tasks: The Kappa statistic. *Computational Linguistics*, **22**(2): 249–54.

Castellón, I., Fernández, A., Vázquez, G., Alonso, L., and Capilla, J. A. (2006). The Sensem Corpus: a Corpus Annotated at the Syntactic and Semantic Level. In *Fifth International Conference on Language Resources and Evaluation (LREC)*, Génova, Italy.

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

Díaz de Ilarraza, A., Garmendia, A., and Oronoz, M. (2004). Abar-hitz: An annotation tool for the basque dependency treebank. In *Proceedings of the fourth international conference on Language Resources and Evaluation, LREC 2004*. Lisbon, Portugal, pages 251–254.

García-Miguel, J. and Albertuz, F. J. (2005). Verbs, Semantic Classes and Semantic Roles in the ADESSE Project. In *Proceedings of the Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, Saarbrüken.

Gardent, C. and Cerisara, C. (2010). Semi-Automatic Propbanking for French. In *Proceedings of the ninth international workshop on Treebanks and Linguistic Theories*, Tartu, Estonia.

Hajic, J. (1998). *Building a Syntactically Annotated Corpus: The Prague Dependency Treebank* Charles University Press, pp. 106–132.

Hajic, J., Panevová, J., Uresová, Z., Bémová, A., Kolárová, V., and Pajas, P. (2003). PDT-VALLEX: Creating a Large-coverage Valency Lexicon for Treebank Annotation. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories*. Vaxjo, Sweden.

Kingsbury, P. and Palmer, M. (2002). From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)*, Las Palmas Canary Islands, Spain.

Kingsbury, P. and Palmer, M. (2003). PropBank: The next level of Treebank. In *Proceedings of Treebanks and Lexical Theories*, Vaxjo, Sweden.

Kipper, K. (2005). *VerbNet: A Broad-coverage, Comprehensive Verb Lexicon.* Ph.D. thesis, University of Pennsylvainia.

Kipper, K., Dang, H., and Palmer, M. (2000). Class-based construction of a verb lexicon. In *Seventeenth National Conference on Artificial Intelligence*.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation (LREV)*, **1**: 21–40.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. Chicago eta Londres: The University of Chicago Press.

Marcus, M. P. (1994). The Penn Treebank: A revised corpus design for extracting predicate argument structure. In *Proceedings of the ARPA Human Language Technology Workshop*, Plainsboro, New Jersey.

Merlo, P. and Van der Plas, L. (2009). Abstraction and Generalisation in Semantic Role Labels: PropBank, VerbNet or both? In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, Suntec, Singapore.

Monachesi, P., Stevens, G., and Trapman, J. (2007). Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of the Linguistic Annotation Workshop*, Praga, pp. 77–84.

Palmer, M. (2009). SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, Pisa, Italia.

Palmer, M., Babko-Malaya, O., Bies, A., Diab, M., Maamouri, M., Mansouri, A., and Zaghouani, W. (2008). A pilot arabic propbank. In *Proceedings of LREC 2008*, Marrakech, Morocco.

Palmer, M., Gildea, A., and Kingsbury, P. (2005a). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics*, Ann Arbor, Michigan, **31**(1): 71–106.

Palmer, M., Nianwen, X., Babko-Malaya, O., Chen, J., and Snyder, B. (2005b). A parallel proposition bank ii for Chinese and English. In *Frontiers in Corpus Annotation, Workshop in conjunction with ACL*.

Palmer, M., Ryu, S., Choi, J., Yoon, S., and Jeon, Y. (2006). *Korean PropBank*. Philadelphia: Linguistic Data Consortium.

Pociello, E., Agirre, E., and Aldezabal, I. (2011). Methodology and Construction of the Basque WordNet. *Language Resources and Evaluation (LRE)*, **45**(2): 121–42.

Salaberri, H., Arregi, O., and Zapirain, B. (2014). First approach toward Semantic Role Labeling for Basque. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland.

Schiffrin, A. and Bunt, H. C. (2007). Lirics deliverable d4.3. document compilation of semantic data categories. Technical report. http://lirics.loria.fr.

Taulé, M., Castellví, J., Martí, M. A., and Aparicio, J. (2006). Fundamentos teóricos y metodológicos para el etiquetado semántico de CESS-CAT y CESS-ESP. In XXII Congreso de la SEPLN, Zaragoza, Spain.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of 6th International Conference on Language Resources and Evaluation*, Marrakech, Maroko.

Van Der Plas, L., Samardžić, T., and Merlo, P. (2010). Cross-lingual validity of PropBank in the manual annotation of French. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV '10)*, Uppsala, Suedia, pp. 113–17.

Vázquez, G., Alonso, L., Capilla, J., Castellón, I., and Fernández, A. (2006). Sensem: Sentidos verbales, semántica oracional y anotación de corpus. *Procesamiento del Lenguaje Natural*, **37**: 113–20.

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

Xue, N. (2008). Labeling Chinese predicates with semantic roles. *Computational Linguistics*, **34**(2): 225–55.

Xue, N. and Palmer, M. (2009). Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, **15**(1): 143–72.

Notes

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 As it is known, PropBank is tagged on the basis of Penn Treebank (Marcus, 1994).

4 *arg_info*: argument information.

5 It should be noted that nowadays VN assigns more arguments (roles) to the verbs in order to better conform to the valency proposed in PB. In this article we

present data collected in 2012, so it could be possible that some data presented here have changed lately.

6 Nowadays VN assigns this Instrument argument to the 'begin-51.-1' class, so we can say that VN has made the same decision that we have made in order to better conform to PB.

7 ERG: ergative declension case; ABS: absolutive declension case; KONP: completive clause.

8 DAT: dative declension case.

9 *-ri buruz*: a complex declension case ('about')

10 ABL: ablative declension case; ALA: allative declension case.

11 *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.

12 We mark cases where the value is either too ambiguous or unnecessary to define with the null mark ('-').

13 We have made this mapping with data collected in 2012.

14 It should be noted that the Extent argumenti is marked 'rare' in PropBank, indicating that it is not a common argumenti in English either.

15 SR: Selectional restriction.

16 Denb: temporal clause.

17 INE: inesive declension case.

18 DES: destinative declension case.

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

20 Thirteen instances per hour are tagged.

21 Two hundred forty-six days.