How the corpus-based Basque Verb Index lexicon was built

Ainara Estarrona, Izaskun Aldezabal*, Arantza Díaz de Ilarraza

IXA NLP group, University of the Basque Country, Computer Languages and Systems Department

* Basque Language and Communication Department

{ainara.estarrona}{izaskun.aldezabal}{a.diazdeilarraza}@ehu.eus

Abstract:

This article describes the method used to build the *Basque Verb Index* (BVI), a corpus-based lexicon. The BVI is the result of semiautomatic annotation of the EPEC corpus with verb predicate information, following the PropBank-VerbNet model. The method presented is the product of a deep study of the syntactic-semantic behaviour of verbs in EPEC-RolSem (the EPEC corpus tagged with verb predicate information). During the process of annotating EPEC-RolSem, we have identified and stored in the BVI lexicon the different role-patterns associated with all verbs appearing in the corpus. In addition, each entry in the BVI is linked to the corresponding verb entry in well-known resources such as PropBank, VerbNet, WordNet and FrameNet. We have also implemented a tool called *e-ROLda* to facilitate the process of looking up verb patterns in the BVI and examples in EPEC-RolSem as a basis for future studies.

Keywords: lexicon; PropBank/VerbNet; semantic roles; predicate labelling; valence

1 Introduction and context

Cognitive and generative linguistics consider language a tool for organizing thought, whereas corpus-based and corpus-driven linguistics view language as a communicative act that entails an opportunity to learn more about the meaning of language. Corpus-based linguistics has brought about a major revolution in the area of lexicography (Hanks 2012).

The present study falls within corpus-based linguistics, and its main goal is the creation of a lexicon of Basque verbs based on the information contained in a corpus tagged semiautomatically with verb predicate information. In this article we offer a detailed description of the method we have developed for building the corpus-

based *Basque Verb Index* (hereafter BVI) lexicon. The BVI lexicon is the result of the semiautomatic annotation of EPEC-RolSem, a Basque corpus labeled with predicate verb information following the PropBank-VerbNet model (hereafter PB-VN). This paper is part of a more general ongoing project about corpus-tagging frameworks which is being pursued at the IXA group[1]. This project 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. The EPEC corpus has been tagged morphologically and syntactically (EPEC-DEP, Basque Dependency Treebank (BDT), Aldezabal et al. 2009), and was recently incorporated into the Universal Dependencies (UD) initiative (Aranzabe et al. 2015). At the semantic level, the nouns have so far been tagged with Basque WordNet senses (Pociello et al. 2010). Our goal, reflecting what has been done in other languages, was to incorporate verb predicate information on the basis of the tagged dependencies that are argument/adjunct candidates. Along with the annotation of verb's predicate information, we have also created the verb lexicon BVI, again in line with the pattern for other languages in building lexicons from tagged corpora. Examples include PropBank (Palmer et al. 2005a), which is tagged on the basis of Penn Treebank (Marcus et al. 1993), related to the VerbNet lexicon (Kipper 2005); or PDT, related to the Vallex lexicon (Hajic et al. 2003). Other projects are pursuing similar approaches: for example, FrameNet (Baker et al. 1998) for several languages, ADESSE (Garcia Miguel and Albertuz 2005) for Spanish, SENSEM (Castellon et al. 2006) for Catalan and Spanish, and AnCora (Aparicio et al. 2008) also for Catalan and Spanish, following the PropBank model. The lexicon and the tagged corpus also constitute essential resources for many computational tasks such as syntactic disambiguation and language understanding, as well as for advanced applications like question answering, machine translation and text summarization.

We chose the PB-VN as the *model* for predicate labelling. 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. 2010b). We based our decision on the facts that 1) the PropBank project started out with a syntactically annotated corpus, exactly as we did; 2) PropBank 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 (Aparicio 2007, Taulé et al. 2006); Catalan (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

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

organization of the lexicon is equivalent to EADB (*Euskal Aditzen Datu Basea – Database of Basque Verbs*), our first database of Basque verbs, proposed in Aldezabal (2004).

After a substantial refinement process, we defined the *guidelines* for semantic annotation of verb predicates and then proceeded to the tagging process (Aldezabal et al. 2011; Estarrona et al. 2016). At the time of writing, 85% of the EPEC corpus has been manually tagged and the remaining 15% has been automatically tagged with an in-house SRL system implemented using machine learning techniques and trained with the manually tagged part (Salaberri et al. 2014). This annotation work has resulted in the development of the BVI which currently contains 1,211 verbs (30,740 occurrences). The BVI contains 288 verbs which include the 151 verbs that have more than 30 occurrences with their respective argument structure information based on the manually annotated corpus (85%), and 923 verbs whose argument structure has been obtained automatically by means of a module that builds new entries from the automatically tagged corpus (15%).

The EPEC-RolSem corpus has been applied to train the SRL system and in a pilot question-generating system for Basque (Aldabe et al. 2013). In addition, the BVI lexicon has been used for some qualitative experiments in machine translation that have demonstrated that the information contained in our lexicon (mainly roles and case markers) is useful for resolving some types of structures, such as transitive/intransitive structures or passive structures (Estarrona 2014).

The article is organized as follows: In Section 2 we first describe the main typological features of the Basque language to help the reader understand the examples included throughout the paper and the importance of cases in the representation of the BVI lexicon. Then, we explain the basic considerations that must be taken into account when applying the PB-VN model and the criteria for adapting the model to Basque and finally, we present some language-specific issues presented in the creation of the lexicon. In Section 3 we describe the methodology followed to create the BVI lexicon. Section 4 is dedicated to presenting the *e-ROLda* tool (http://ixa2.si.ehu.es/e-rolda/index.php) to facilitate study of role-patterns of verbs included in the BVI lexicon and looking up examples in EPEC-RolSem. Finally, in Section 5, we present some conclusions and suggest future lines of research.

2 Adapting the PB-VN model and establishing the criteria for its application to Basque

Adapting a predicate annotating model from one language to another is never straightforward. One encounters language-specific issues, and the model itself may also contain questionable aspects and gaps in its coverage of linguistic phenomena. Thus, after carrying out the studies required to verify that the model was also appropriate for Basque (Aldezabal et al. 2010a, Aldezabal et al. 2010b), we faced the challenge of adapting it to Basque.

We shall begin with a brief description of the main typological features of Basque and their effect when describing the lexical information of the Basque verbs.

2.1 Typology of Basque and its implications for the BVI lexicon

As a non-Indo-European language, indeed an isolate, Basque grammar differs considerably from that of the languages surrounding it. It is agglutinative, head-final and pro-drop. Basque is usually assumed to be a Subject-Object-Verb (SOV) type language (de Rijk 1969), but is also described as having 'free word order', meaning that the order of phrases in the sentence can vary (Laka 1996).

A declarative sentence in Basque contains a verb and its arguments, an aspect marker attached to the verb and a verbal inflection containing agreement morphemes, tense and modality. It can also contain other phrases such as adverbials or postpositional phrases (Laka 1996). Examples given in (1) are from Laka (1996):

(1)

a. *umea kalean erori da*
child-the-Abs[2] street-in fall-asp is
'the child fell in the street'

b. *emakumeak gizona ikusi du*
woman-the-Erg man-the-Abs seen has
'the woman has seen the man'

c. *gizonak umeari liburua eman dio*
man-the-Erg child-the-Dat book-the-Abs given has
'the man has given the book to the child'

d. *emakumea heltzen da*
woman-det-Abs arriving is
'the woman arrives'

The arguments of the verb can be identified by grammatical cases or postpositions. There are three grammatical cases in Basque: Ergative (*k* morpheme), Dative (*i* morpheme) and Absolutive (Ø morpheme) (see examples in (1)). Basque has a strong tendency to place the heads of phrases at the end of the phrase. Rather than pre-positions at the beginning of prepositional phrases, Basque has post-positions that appear at the end of postpositional phrases (2). Grammatical cases are no exception to this generalization (Laka 1996).

2 Abs: absolutive; Erg: ergative; Dat: dative.

(2)

a. *[Bilboko kale bat]-ean* (locative)
[Bilbo-of street one]-in
'in one street of Bilbo'

b. *[zazpi leiho]-tatik* (ablative)
[seven window]from
'from seven windows'

These morphosyntactic features were taken into account for creating the in-house Database of Basque Verbs (EADB, Aldezabal, 2004), which we used as a basic resource for building the BVI lexicon.

EADB is a database of 100 verbs from EPEC, including the most frequent ones. Aldezabal (2004) defined a number of syntactic-semantic frames (SSF) for each verb, which are composed of semantic roles and the case that syntactically performs each of them. When defining the SSFs for each verb, the following principles are assumed: i) The SSFs that have the same semantic roles define a coarse-grained verbal sense and are considered syntactic variants of an alternation, and ii) different sets of semantic roles reflect different senses (Aldezabal 2010, Aldezabal et al. 2010a, Estarrona et al. 2016).

In the BVI lexicon, we unify the information collected in EADB[3], in PropBank-VerbNet and in the EPEC-RolSem corpus. Table 1 shows the entry of the verb *saldu* (‘to sell’) in the BVI lexicon. As it can be seen, the first and only sense of the verb *saldu* corresponds to the ‘sell_01’ roleset in PropBank. It also shows the three arguments that PropBank defines for this roleset (Arg0, Arg1 and Arg2) and the roles that VerbNet assigns for this verb class (agent, theme and recipient). Finally, the roles provided by EADB are indicated, as well as the cases that perform each role (Source-ERG, Theme-ABS, Goal-DAT).

Table 1: The entry in the BVI lexicon for the verb *saldu* (‘to sell’).

saldu_1#sell_01
Arg0:Agent:Source**:**ERG
Arg1:Theme:Theme:ABS
Arg2:Recipient:Goal:DAT

3 Only in the case of the 100 verbs analysed in this study.

2.2 Basic considerations when applying the PB-VN model

Before starting with our basic considerations let us explain briefly the general framework of the PB-VN model.

PropBank defines semantic roles on a verb by verb basis. An individual verb's semantic arguments are numbered beginning with 0 (Palmer et al. 2005a). The elements that are regarded as arguments are numbered from Arg0 to Arg5[4], expressing semantic proximity with respect to the verb. The lower numbers represent the main functions (subject, object, indirect object, etc.). Adjuncts are tagged as ArgM. PropBank annotation scheme uses numbered arguments because of the difficulty of defining a universal set of semantic roles covering all types of predicates (Bonial et al. 2017).

PropBank adds specific roles for 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 and extending Levin's classification (Levin 1993). Table 2 shows the PropBank roleset for the verb 'tell.01' and the corresponding VerbNet roleset with the Levin class number (37.1)[5]. 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 below.

Table 2. PropBank and VerbNet rolesets of the verb 'to tell'.

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

We have to say that at present, we are using the roles as they existed before the changes in version VN 3.2 (Bonial et al., 2011). A move to this version of VN may require a revision of our decisions.

4 PropBank has recently added an Arg6 to tag nominal natural disaster Rolesets. (Bonial et al. 2017).

5 In the 3.3 version of VerbNet many changes have been implemented: path_rel semantics, initial lexical features, an many updates to verb classes, frames, and members, but full documentation about these changes is not jet available (http://verbs.colorado.edu/verbnet/). We use in this paper the data as it existed before all these changes.

In the next subsections, we will describe the three main considerations we have taken into account when applying the PB-VN model: the choice between Arg0 and Arg1, the option we choose when VerbNet has two or more classes for a single PropBank roleset, and finally, the addition of the 'path' role which is not included in VerbNet 3.1.

2.2.1 Regarding Arg0 and Arg1

As noted above, in PB the arguments are numbered from Arg0 to Arg5 and then they are linked to VN roles. In fact, however, Arg1 is always linked with the Theme (or Patient) role and Arg0 with the Agent role. No fundamental linguistic reason exists for this, though for example in Kingsbury and Palmer (2003:3) it is said:

"(...) 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"*

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 we can see in Table 3:

Table 3. The verb "come.01" in PropBank.

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

We decided 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.

Regarding intransitive verbs denoting change of location, we consider the subject to be at the same time the entity initiating the action and the entity undergoing it (in agreement 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 behaviour), 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>>…" (Bonial et al. 2015:8)

Thus, in the case of an unaccusative verb like "come.01" where only the intransitive variant is possible, we consider the entity performing the action and the entity undergoing it to be the same; thus, we tag it as Arg0 *Theme*. 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).

2.2.2 More than one VerbNet class for the same PB argument

Sometimes VerbNet has two different classes (or more) available for the given PB roleset, and consequently there are two different roles for each argument. This is the case for the verb 'see.01' (Table 4):

Table 4. The verb "see.01" in PropBank.

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)

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

By contrast, in EADB the verb *ikusi* ('to see') contains two arguments with an unique role each corresponding to 29.2 class of VerbNet:

Arg0: *esperimentatzailea* (which would be the agent)
Arg1: *gaia* (which would be the theme)

In those cases, we have decided to base our decision on EADB and to assign the VerbNet roles corresponding to class 29.2. The result in the BVI lexicon would be:

Arg0: Agent, *esperimentatzailea*

Arg1: Theme, *gaia*

2.2.3 Other roles: The Path role

It has been necessary to add some new roles, for example, the 'path' role which is not specified in VerbNet, but appears in our EADB data-base. For instance, for the verb *pasatu* ('pass' / 'come by') we find examples like:

(3) *Zure etxetik pasatu naiz gaur goizean*
you-Gen house-Abl passed-by I today morning-the-Ine
'I came by your house this morning'

As we have said before, in the 3.2 version of VN some roles have changed to make the list of roles consistent with the standard list proposed in the LIRICS project[6] (Bunt et al. 2007, Schiffrin and Bunt 2007). Thus, VN now contains a *Trajectory* role that could be equivalent to our *Path* role. In the same way, we have seen that some *Theme1* roles have been changed to *Pivot*[7].

2. 3 Cross-linguistic differences and criteria adopted

Adapting 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 EADB the Basque verb *eskatu* ("ask.02"), has two arguments, Arg0: *Esperimentatzailea* (Experiencer) and Arg1: *Gaia* (theme). The dative complement is not included among the subcategorised cases because it is optional in the sense that it is considered that one can 'ask for' something, in general, without stating explicitly the 'goal' as an impersonal proposition (alternation). However, the verb "ask.02" contains three arguments in PropBank and VerbNet:

6 Linguistic InfRastructure for Interoperable resourCes and Systems (http://lirics.loria.fr).

7 See Bonial *et al*. (2011) for details on the comparison between VN and LIRICS lists of roles and the decisions taken in the 3.2 version of VN.

ask.02
Arg0: Agent
Arg1: Theme (proposition)
Arg2: Patient

Therefore, we follow the PB-VN model, tagging the DAT (dative) argument as Arg2.

Nevertheless, as we performed the verb tagging, we encountered some difficult cases. We will explain the main phenomena below.

2.3.1 Arguments proposed by PB-VN that are not present in Basque

In some verbs of change of location, PB-VN proposes types of arguments that are not possible in Basque. See the example of the verb *etorri* ("come.01") in Table 5:

Table 5. The verb "come_01" in PropBank.

come.01: motion, vncls: 51.1, framnet: Arriving
Roles:
Arg1: entity in motion / comer (vnrole: 51.1-theme)
Arg2: extent -- rare
Arg3: start point
Arg4: end point

In Basque the second argument is not possible, so we disregard it and assign its number to the next possible argument. That is, Arg1 will be the "start point" (since for us the subject of this verb is Arg0[8]) and the "end point" will be Arg2. After these changes, the resulting entry in BVI is the same as in the EADB (Table 6):

Table 6. The "etorri_come.01" verb in the BVI.

etorri_come.01
Arg0: Theme, affected theme (ABS)
Arg1: Source, start point (ABL)
Arg2: Destination, end point (ALA)

2.3.2 More than one PropBank verb exists for a Basque verb

A Basque verb can be linked to more than one PropBank verb. In such cases, we check first 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

8 See Section 2.2.1 for an explanation of the Arg0/Arg1 choice.

unquestionably be linked to both “tell.01” and “say.01”. We establish the correspondence and indicate this double equivalence by the expression “tell.01/say.01”. If the roles and arguments do not coincide with their equivalents in PB, we annotate the specific instances with the one we consider most suitable in the context. The verb *egin*[9] (‘to do’) is an example (4):

(4) a. *Kanta asko* ***egin*** *zituen* (He/she composed a lot of songs): compose.02 (agent, product, beneficiary)
b. *Ondoko galdera* ***egin*** *diote Juan Jose Ibarretxeri* (They asked Juan Jose Ibarretxe this question): ask.02 (agent, topic, recipient)
c. *Biek ere joko alaiegia* ***egiten*** *zuten ACBrako* (Both of them played a too happy-go-lucky game for the ACB): play.01 (agent, theme, instrument)

The verb *egin* (‘make/do’) in Basque can mean ‘compose’, ‘ask’ or ‘play’ depending on the context, in the sense that in Basque we ‘make songs’, we ‘make questions’ and we ‘do game’.

2.3.3 Motion Verbs

Motion verbs have been widely analysed in different languages. The concept of movement appears in all languages, but each language has its own way of expressing it, that is, of lexicalising it. Following Talmy’s typology (Talmy, 1985), Basque is a verb-framed language, as are Spanish and Turkish. However, English is a satellite-framed language, like German among others. Satellite-framed languages leave the element that marks the direction of movement outside the verb (‘out’, ‘in’, ‘up’, ‘down’, etc.). Verb-framed languages, however, indicate the direction of movement inside the verb, for example, *igo* (‘go up’) / *jaitsi* (‘go down’).

Therefore, Basque and English have different ways to lexicalise movement, and this has caused us some problems when finding an exact PB equivalent for movement verbs in Basque. These phrasal verbs are not systematically included in PB, and consequently, it has not always been easy to find an English equivalent for this type of verbs in Basque. For example, the verb *atera* (‘take out’) has a general sense which is ‘change of location’, and this sense can be expressed by the verbs ‘take out’, ‘come out’ or ‘go out’ in English. ‘Take out’ is included in PB (take.11), but not with the sense of a change of location, but with the meaning of ‘obtain’. The same happens with ‘come out’ (come.09), which has the meaning of ‘appear’ in PB,

9 The examples under (4) could suggest that this verb is mainly a light verb. Light verbs are not the focus of this research, but it has to be said that at the moment we are working on this issue to see how the light verbs and the multiword expressions created with light verbs must be included in the BVI lexicon.

and with ‘go out’ (go.17), which only has 2 arguments while the change of location in Basque needs 4 arguments. There is no specific entry in PB for these phrasal verbs, and this sometimes forces us to create an equivalent entry in PB format for this type of motion verbs in Basque. This way, we have created the equivalents “come_out.01” and “take_out.01” for the change of location sense of the verb *atera.* We have done the same with other similar motion verbs.

2.3.4 Causative/inchoative alternation

When analysing the verbs that present causative/inchoative alternation in Basque, it is not always clear whether it is a single sense with two alternations or there are two different senses. This doubt is reflected in dictionaries, which do not always act consistently in these cases. For example, the verb *hil* (‘to die’) has the intransitive alternation *hil da* (‘somebody dies’) and the transitive one *hil du* (‘somebody kills somebody’). We have the same for other verbs as *sartu DA/DU* (‘to go in’ / ‘to put in’) and *atera DA/DU* (‘to go out’ / ’to take out’).

In Basque we have a verb which has a single sense and two alternations (causative/inchoative), and in English we have one different verb for each alternation. As we are building a lexicon based on a model created for English, in this case we have had to separate each of the alternations of Basque into two different senses: one for the sense of ‘to die’ and another one for the sense of ‘to kill’, even though we think that there is a single general meaning that is ‘someone or something causes the death of someone’.

Another topic to analyse in terms of language-specific issues and their implications in the structure of the BVI lexicon would be Multiword Expressions (MWE), especially Light Verb Constructions (LVC). How do we insert them into the lexicon? Do they have to be separate entries or new senses of existing entries? When are these MWEs derived from the general predicate and when are they domain dependent? How should this be reflected in the lexicon? We believe that there is a new line of research which requires in-depth analysis, and that is why we have not included this topic in the present paper.

3 Building the lexicon: BVI version 1

The BVI (Basque Verb Index) lexicon is the first repository of syntactic-semantic information on Basque verbs. This resource is one of the results deriving from the process of semiautomatic annotation with verb predicate information of the verbs in

EPEC-DEP[10] and it is an important mechanism for the completion of the annotation process itself. It has been constructed based on the behaviour of the verbs contained in the EPEC-RolSem corpus.

The task of tagging the EPEC-DEP corpus with the corresponding verb predicate information has been a process of continuous refinement, which is described in detail in Estarrona et al. (2016). In this section, we will mainly present the aspects that directly influence the construction of the BVI lexicon within the annotation process. As Figure 1 shows, this process was partly manual, partly semiautomatic and partly automatic.

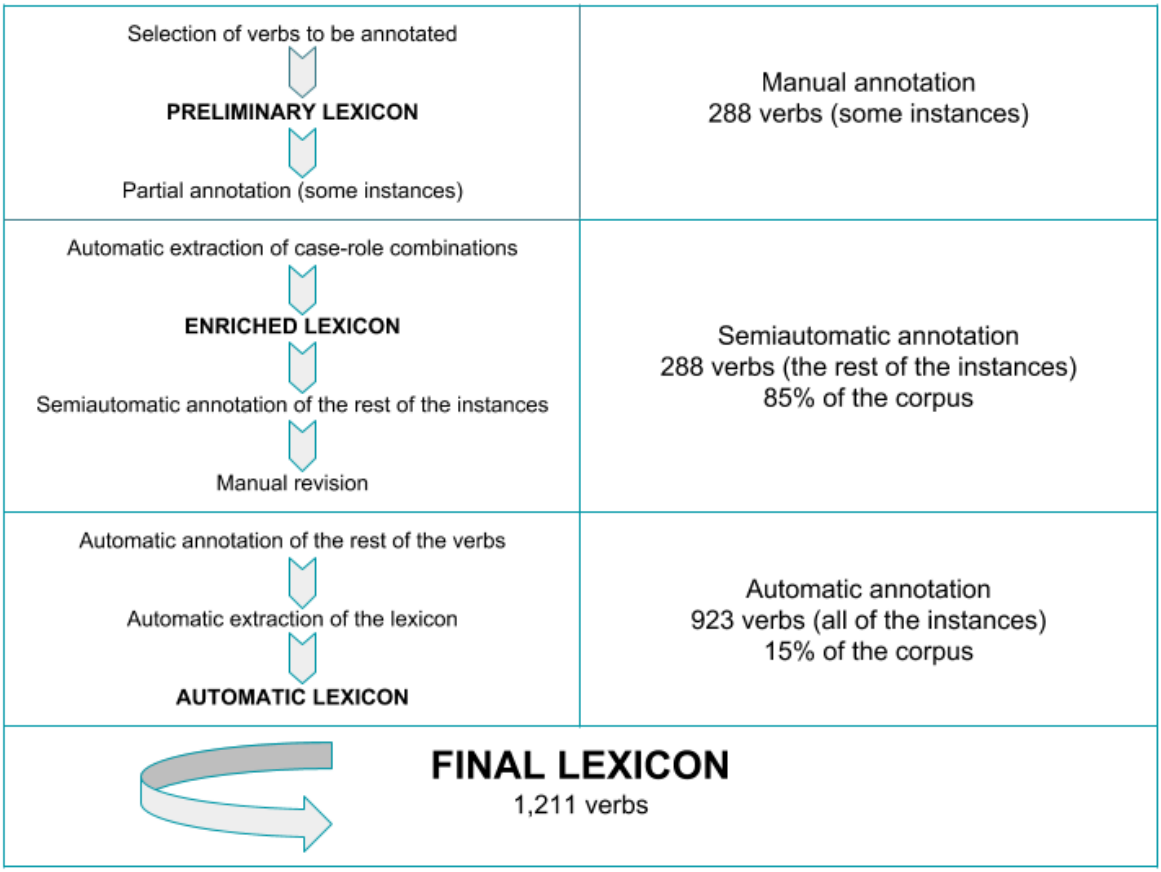

Figure 1. Steps in the final method.

It must be noted that, during the first annotation process, we carried out a manual evaluation (Aldezabal et al. 2011) to validate the proposed method and guidelines. We achieved an inter-annotator agreement of 0.80 calculated by Cohen's kappa. However, it should be said that the evaluation was only performed with 3 verbs because our main objective was not to evaluate the quality of the annotated corpus, but to validate the coverage and quality of the guidelines[11]. The main conclusion we

10 The EPEC-DEP corpus is the EPEC corpus syntactically tagged using a dependency grammar.

11 The 3 selected verbs were *adierazi* (‘to state’), *izan* (‘to be’) and *etorri* (‘to come’). We chose very different verbs to be able to draw interesting conclusions. The verb *adierazi* has a single sense and is very frequent in the corpus. The verb *izan* is the most frequent verb in the corpus (15.22%). Finally,

drew from this evaluation was that annotators did not understand the PB-VN criteria in the same way, and to ensure satisfactory results each verb entry needed to be edited completely before beginning to annotate its specific instances: one must be clear not only about the English equivalent for the sense but also about the numbered arguments and the assignment of roles. We therefore needed to define the criteria presented above in Sections 2.2 and 2.3.

3.1 The preliminary lexicon

We started our work selecting the verbs contained in EADB because we decided that this resource will be a guide in the first stages of the creation of the BVI lexicon.

Our first step was to create the preliminary entry of each selected verb combining the information included in EADB and PB-VN. We collected the information contained in both resources and we built the preliminary BVI lexicon entry for the given verb. Table 7 shows the example of the verb *jan* (‘to eat’):

Table 7: The entry in the BVI lexicon for the verb *jan* (‘to eat’).

jan_1#eat_01
Arg0:Agent:Experiencer**:**ERG**:**[+animate]
Arg1:Patient:Theme:ABS**:**[-animate]

At the end of this manual process, we had a *preliminary lexicon* for 288 verbs (including the 150 most frequent verbs in the corpus which cover 85% of the corpus).

3.2 The enriched lexicon

After creating the preliminary entry and manually annotated a sample of instances for the 288 verbs, we automatically derived the annotation of non-tagged instances from the annotated instances. We obtained the set of associated syntactic-semantic combinations (case-role combinations). Table 8 illustrates the information obtained for the verb *aldatu*: each row shows the syntactic-semantic pattern by means of its associated PropBank verb, VerbNet roles and their corresponding Basque cases.

the verb *etorri* is *a priori* a difficult verb, because it has 4 senses (not always easily distinguishable) and it is used extensively in complex expressions.

Table 8: syntactic-semantic combinations of the verb "*aldatu*_alter_01/change_01".

BasqueV	PropBankV	VerbNet roles and Basque cases
Aldatu	alter_01#change_01	Agent:ERG Patient:PAR[12] 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:ABS Patient:ABS

These case-role combinations allow us to enrich the lexicon with new cases and roles not included in the preliminary version. Compared to the preliminary lexicon, the enriched BVI contained 8.32% more roles and 23.66% more cases.

Once we had enriched the lexicon with new cases and roles, we automatically calculated the frequency of appearance of each case associated with a concrete semantic role. In this way, we obtained the information about the verb "*aldatu*_alter_01/change_01" illustrated in Table 9:

Table 9. Percentages of occurrences of Basque case/VerbNet role pairs.

Basque case	VerbNet role[13]	Percentage of occurrences
ABL[14]	Product	50%
ABL	Material	50%
ABS	Patient	85%
ABS	ADV	7%
ABS	MNR	4%
ABS	TMP	2%
ALA	Product	100%
BALD	DIS	100%
DENB	TMP	100%
ERG	Agent	88%
(...)		

On the basis of this automatic study, we adapted our annotation tool so it could be of assistance in the manual annotation of the instances that have not been tagged yet. The tool automatically offers information about the instances to be annotated and proposes to the human annotator an association between a case and a semantic role where the combination has a frequency greater than or equal to 50 percent. Thus, in the example in Table 9, we did not consider the ABS/ADV, ABS/MNR and ABS/TMP combinations, because their frequencies of appearance were lower than

12 PAR: partitive case; INE: inessive case; GEN: genitive case; NEG: negative particle; HELB: purpose clause.; TMP: temporal; PRP: purpose.

13 Some of the roles that we have added to the VerbNet role list are not VerbNet roles but roles for adjuncts in PropBank, for instance ADV, TMP, DIS, MNR…

14 ABL: ablative case; ALA: allative case; BALD: conditional clause; DENB: temporal clause.

50 percent. This process facilitated the annotation work substantially: in 70% of cases the tagging proposed was completely correct, while in the remaining 30% the proposal, while useful, required some kind of correction (Estarrona et al. 2016).

Throughout all this labelling process, 3 human annotators have participated. A single expert annotator has edited the preliminary lexicon entries first. Then, the expert annotator has trained 2 annotators for manually tagging a sample of occurrences of 288 verbs, and finally, this expert annotator has revised the result of the semiautomatic process for the non manually tagged occurrences of these 288 verbs.

At this stage of the lexicon, it so far had the syntactic-semantic patterns of 288 Basque verbs (with 461 different senses) defined manually (accounting for 85% of the whole corpus).

Each entry in the BVI contains the following information:

1. The Basque verb senses and its PropBank equivalents
2. A set of elements consisting of: i) number of argument, ii) semantic role in VN-PB, iii) semantic role stored in EADB, iv) case and v) (optional) selectional restriction regarding animate/inanimate, human/non-human, concrete/non-concrete semantic features.

A total of 26,028 occurrences have been labelled and, taking into account that we tag, on average, 13 occurrences per hour, the manual work of labelling has lasted just over a year[15].

3.3 Automatic lexicon

The next step was to add the remaining 923 verbs (15% of the corpus). Being aware that performing such work manually is beyond our reach, we decided to do all the process (including the annotation of the corpus and the extraction of the lexicon) automatically.

3.3.1 Corpus tagging by means of an SRL system

In order to annotate the corpus we have used the in-house SRL system described in Salaberri *et al.* (2014). Typically, the role labelling task consists of identifying the arguments of each predicate, verbal predicate in our case (argument identification) and labelling them with semantic roles (argument classification). However, in order to identify and classify these arguments, SRL systems have to identify the predicate first (predicate identification) and then assign a sense to it (predicate classification).

15 We do not include the time and personnel involved in earlier phases such as setting up the annotation criteria, creating the guidelines, or preparing the tool for the annotation task.

The semantic role labeller we have used has just focused on the argument identification and classification by making use of the manually identified and classified verbs in the EPEC-DEP corpus. The system was implemented using machine learning techniques and trained with the manually tagged part. It was evaluated with the manually annotated part of the corpus and scored 84.30 F1 in identifying the PropBank semantic role for a given constituent and 82.90 F1 in identifying the VerbNet role. This system establishes the baseline for basque SRL (Salaberri et al. 2014) .

In order to check the reliability of the system with verbs that we had not annotated manually in the previous step, we carried out a manual evaluation. We have created a gold standard consisting of an 800 occurrences sample of 24 new verbs and then, we have compared it with the same sample tagged by the SRL system. The results were quite adequate: 85.90 F1 identifying the PB argument and 78.20 identifying the VN role.

3.3.2 Automatic extraction of the lexicon

All the verbs in the lexicon have their PB equivalent, so the first objective when automatically creating the lexicon was to assign a PB equivalent to each Basque verb, for which the bilingual Elhuyar Basque-English was used[16]. In this step we assumed that the verbs have a single sense, because our SRL system does not assign a sense to them.

Once the PB equivalent for the Basque verb had been inserted in the first field of the ARG_INFO tag[17] and tagging had been carried out by the SRL system, we executed automatic procedure as explained in previous section (3.2) to obtain the syntactic-semantic patterns of the new verbs. This automatic procedure extracts from the ARG_INFO tag the information contained in the first, 4th and 5th fields; the case is inherited from the dependency tagging. After the results are obtained, we group the arguments, roles and cases automatically and build the canonical entry for the verb. This automatic grouping includes i) all the different arguments that appeared in the syntactic-semantic patterns (Arg0, Arg1, Arg2…), ii) the most frequent argument-role pair, and iii) all the cases for each argument-role pair.

In the output of the automatic processing, a variety of cases are found as illustrated in the following examples, while showing how we create the canonical lexicon entry from these syntactic-semantic patterns.

16 https://hiztegiak.elhuyar.eus/eu_en

17 ARG_INFO tag is the semantic label we have created to annotate verb predicate information. For more details about this label see Estarrona *et al.* 216.

A) In some cases we find the **same syntactic-semantic frame** in all occurrences in the corpus of a given verb. For instance, all the instances of the verb *poztu* (‘to delight’) in the corpus have an Arg1_*Theme* argument that appears with the absolutive (abs) case[18]. Therefore, we take this Arg1_*Theme* argument in the absolutive case to build the BVI entry for this verb (Table 10):

Table 10. The entry for *poztu* in the BVI lexicon.

poztu_1#delight_01
Arg1: Theme: ABS

B) In other cases, the syntactic-semantic frames are the same, but in some occurrences one or more **arguments are elided**. In these cases we group the patterns and build the canonical entry of the verb. Table 11 shows the syntactic-semantic patterns for the verb *aholkatu* (‘to advise’):

Table 11. Syntactic-semantic patterns in the corpus for *aholkatu* (‘to advise’).

AHOLKATU	
advise_01/arg0/Agent/erg#advise_01/arg1/Theme/abs	1/3
advise_01/arg1/Theme/abs	1/3
advise_01/arg1/Theme/konpl#advise_01/arg0/Agent/erg #advise_01/arg2/Recipient/dat	1/3

The first pattern has two arguments (Arg0_*Agent*_erg and Arg1_*Theme*_abs), the second has only one (Arg1_*Theme*_abs) and the third has three arguments (Arg1_*Theme*_konpl, Arg0_*Agent*_erg and Arg2_*Recipient*_dat). The Arg1_*Theme* argument-role pair appears in the absolutive case and with a complement clause (*konpl*), so we include both in the canonical entry as in Table 12:

Table 12: Syntactic-semantic patterns in the corpus for *aholkatu* (‘to advise’).

Aholkatu_1#advise_01
Arg0: Agent: ERG
Arg1: Theme: ABS/KONPL
Arg2: Recipient: DAT

C) Finally, we have **different syntactic-semantic frames** in each occurrence in the corpus. An example is seen in Table 13 with the verb *ailegatu* (‘to arrive’):

Table 13. syntactic-semantic patterns in the corpus for *ailegatu* (‘to arrive’).

AILEGATU	
arrive_01/arg0/Agent/erg	1/7
arrive_01/arg1/Theme/abs#arrive_01/arg2/Destination/ala	2/7

18 We do not take into consideration adjuncts (ArgM) when building lexicon entries.

arrive_01/arg2/Destination/abu#arrive_01/arg2/LOC/ine 1/7
arrive_01/arg2/Destination/ala 2/7
arrive_01/arg2/LOC/ala 1/7

This verb has three different arguments (Arg0, Arg1 and Arg2) and 4 different argument-role pairs: Arg0_*Agent*, Arg1_*Theme*, Arg2_*Destination* and Arg2_*LOC*, but the pair Arg2_*Destination* is the most frequent. Furthermore, there are two different cases for the Arg2_*Destination* argument-role pair. In this way we build the canonical lexicon entry as in Table 14:

Table 14: Syntactic-semantic patterns in the corpus for the verb *ailegatu* ('to arrive').

Ailegatu_1#arrive_01
Arg0: Agent: ERG
Arg1: Theme: ABS
Arg2: Destination: ALA/ABU

All these entries that are obtained automatically will be marked with a green bullet in the lexicon (Figure 2)[19] to tell the user that the verb entry has been created automatically through automatic tagging of the corpus.

Figure 2. The entry of the verb *ailegatu* ('to arrive') marked with a green bullet to warn the user that it is an automatically analysed verb.

3.4 First edition of the BVI lexicon

This first edition of the BVI lexicon contains syntactic-semantic information about all the verbs in the EPEC-DEP corpus; it is formatted in XML. Of the 1,211 verbs in the corpus, 288 entries (covering 85% of the corpus) had been created manually and the remaining 923 (covering 15% of the corpus) automatically. The imbalance between the number of verbs and their frequencies in the corpus is usual in other languages as well. For example, in the PropBank corpus there are 3,101 different

19 This figure is taken from our e-ROLda tool that we will present in detail in Section 4.

verbs of which only 485 have more than 30 occurrences. In the case of EPEC-DEP the percentage of verbs that have more than 30 occurrences is 12.71%, while in PB it is 15.64%. This distribution of the frequency of verbs follows Zipf's Law (Zipf, 1949)[20]. This distribution is known as the "long tail" and is quite a usual phenomenon in any language's corpus (Estarrona, 2014).

4 e-ROLda: A tool for retrieving information in the BVI and EPEC-RolSem

In this section, we present *e-ROLda,* the tool that allows us to view the information contained in the BVI lexicon and the EPEC-RolSem corpus[21].

When entering the *e-ROLda* system (http://ixa2.si.ehu.es/e-rolda/index.php), information is given about the tool itself and the search features offered. Searches can be performed with regard to different general features: i) the Basque verb, ii) a concrete sense of a Basque verb, and iii) the PB-VN English verb. The tool also has a private section allowing an authorized linguist to edit the BVI lexicon; it is implemented in PHP. In addition, we use a *MySQL* database to store the data in the EPEC-RolSem corpus and links to other resources.

When looking up a verb, the system usually returns:

1) Senses recorded for the verb in the BVI Lexicon.
2) Links to PropBank (PB), FrameNet (FN) and Basque WordNet (BWN) (Laparra and Rigau 2010).
3) Examples of sentences in EPEC-RolSem for the given verb, grouped by senses. For each example, the system indicates: i) the file; ii) the number of the sentence in the file; iii) the verb and sense; iv) the equivalent in PB-VN; v) the number of arguments present in the example, taking as a reference the number of arguments in the pattern (this is relevant information for future studies about elided constituents); vi) examples in the corpus and a link ("check the analysis") to their corresponding analysis (listing dependency relations and the semantic information in the ARG_INFO tag.

The system allows more advanced searches where the user may ask for sentences that contain a verb with a certain: a) argument (Arg0, Arg1, ...); b) semantic role in

20 "These verbs are arrayed in a classic Zipfian distribution, with a few verbs occurring very often (*say*, for example, is the most common verb, with over 10,000 instances in its various inflectional forms), and most verbs occurring two or fewer times" (Palmer et al., 2005a: 13).

21 At the time of writing, we are working on a Basque NOMLEX and including the information of this new resource in the e-ROLda tool. Given the fact that the work is ongoing, the data is still tentative and incomplete at this stage.

PB-VN; c) semantic role in our EADB database; d) case and e) selectional restriction associated with the semantic role. These detailed searches may be useful for studying linguistic phenomena such as ellipsis in Basque: which argument tends to be elided most with a particular verb? Or is there a role that tends to be elided more than others, say in motion verbs? By means of the *e-ROLda* tool we can analyze real examples from the corpus and try to answer these questions

The *e-ROLda* tool is similar to tools used in a number of projects noted above, such as ADESSE[22], SenSem[23] and AnCora[24], which all contain an environment for exploring both the corpus and the lexicon in their corresponding languages. For English there is the *Unified Verb Index* where the information stored in the resources PropBank, VerbNet, FrameNet, WordNet and OntoNotes (Pradhan et al. 2007) is shown for each verb. However, there is no option that would permit detailed searching in the various components of the verbs.

5 Conclusions and future work

This paper has two objectives: i) to present the methodology used in the creation of the BVI lexicon derived from the annotation of verb predicate information in the EPEC-DEP corpus, and ii) to present *e-ROLda,* a tool built to extract information associated with verb argument structure information and examples of the annotated verbs. In as much as the process of semantic tagging relied on previously performed syntactic annotation, the BVI Lexicon was built on the foundation of the syntactic-semantic structure of the verbs in the corpus.

As far as we know, this is the first resource built for Basque that applies the main ideas used in similar resources for other languages. 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 facilitate work utilizing those resources significantly. In addition BVI (and consequently *e-ROLda* tool) is included in the list of VerbNet-based lexicons on its website (“VerbNets in other languages”: http://verbs.colorado.edu/verbnet/).

The BVI lexicon resulting from the annotated corpus (EPEC-RolSem) can be consulted by means of the *e-ROLda* tool. This tool provides facilities for requesting information about the syntactic and semantic structure of verbs as well as examples of use.

22 http://adesse.uvigo.es/

23 http://grial.uab.es/projectes/SenSem.php

24 http://clic.ub.edu/corpus/en/ancora

The BVI lexicon stores information about all the verbs in the EPEC-DEP corpus. It contains syntactic-semantic information about 1,211 verbs. The entries for the 288 most frequent verbs (covering 85% of the sentences in the corpus) have been defined manually and the information for the other 923 verbs has been obtained automatically by an SRL system trained with the tagged part (Salaberri et al., 2014).

The annotation of the EPEC-RolSem corpus and the creation of the BVI verb lexicon opens new lines of research: i) study of the verbs appearing in Multiword Lexical Units (MWLU) or Multiword Expressions (MWE) and consequently, also in the field of light verbs; ii) study of verb patterns in specialized corpora.

The data stored in the BVI will further be used to analyse features associated with the Basque language, such as the most typical sequences of case-roles in verb patterns or the phenomenon of ellipsis.

Acknowledgements

This research has been supported by the Basque Government: (IXA group, (IT344-10), the Ministry of Science and Innovation of the Spanish Government (PROSA-MED (TIN2016-77820-C3-1-R)) and MINECO: TUNER (TIN 2015-65308-C5-1-R).

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 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). ISBN: 90-420-1836-4.

Agirre, E., Aldezabal, I., Etxeberria, J. and Pociello, E. (2006). A Preliminary Study for Building the Basque PropBank. *Proc. of the 5th International Conference on Language Resources and Evaluations (LREC'06)*, 981-986. Genoa, Italy. ISBN: 2-9517408-2-4.

Aldabe, I., Gonzáles-Dios, I., López-Gazpio, I., Madrazo, J. and Maritxalar, M. (2013). Two Approaches to Generate Questions in Basque. *Procesamiento del Lenguaje Natural, 51,* 101-108. Print ISSN: 1135-5948. Online ISSN: 1989-7553.

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

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, Dipartamente di Linguistica. Pisa (Italy).

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* 5-2, 241-269. Mouton de Gruyter. Berlin-New York. Print ISSN: 1613-7027. Online ISSN: 1613-7035.

Aldezabal, I., Aranzabe, M. J., Díaz de Ilarraza, A., Estarrona, A. and Uria, L. (2010a). 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. ISSN: 0302-9743, ISBN-10: 3-642-12115-2.

Aldezabal, I., Aranzabe, M.J., Díaz de Ilarraza, A., Estarrona, A. (2010b). Building the Basque PropBank. 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)*, pp. 1414-1417, European Language Resources Association (ELRA), ISBN: 2-9517408-6-7. LREC 2010, Valletta (Malta), May 19-21, 2010.

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. *XXVII Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2011).* Universidad de Huelva.

Aparicio, J., Taulé, M. and Martí, M.A. (2008). AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias (eds.). *Proc. of 6th International Conference on Language Resources and Evaluation (LREC'08)*, 797-802. ELRA. ISBN: 2-9517408-4-0.

Aranzabe, M.J., Atutxa, A., Bengoetxea, K., Díaz de Ilarraza, A., Goenaga, I., Gojenola, K. and Uria, L. (2015). Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. Markus Dickinsons, Erhard Hinrichs, Agnieszka Patejuk, Adam Przepiórkowski (eds), *Proceedings of the Fourteenth International Workshop on Treebanks an Linguistic Theories (TLT14)*, 233-241. Institute of Computer Science of the Polish Academy of Sciences, Warszawa, Poland. ISBN: 978-83-63159-18-4

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. *Proc. of the Workshop on Frontiers in Linguistically Annotated Corpora, A Merged Workshop with 7th Int. Workshop on Linguistically Interpreted Corpora (LINC-2006) and Frontier in Corpus Annotation III (Coling/ACL 2006),* 70-77. Association for Computational Linguistics (ACL). Sydney, Australia. ISBN: 1-932432-78-7.

Baker, C.F., Fillmore, C.J., and Lowe, J.B. (1998). The Berkeley FrameNet Project. *Proceedings of COLING-ACL'98,* 86-90. *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (COLING-ACL'98).* Montréal, Quebec, Canada. Morgan Kaufmann Publishers / ACL.

Bhatt, R., Narasimhan, B., Palmer, M., Rambow, O., Sharma, D. and Xia, F. (2009). A Multi-Representational and Multi-Layered Treebank for Hindi/Urdu, *Proceedings of the Third Linguistic Annotation Workshop, ACL-IJCNLP 2009,* 186-189. Association for Computational Linguistics (ACL). Suntec, Singapore.

Bonial, C., Corvey, W., Palmer, M., Petukhova, V. and Bunt, H.C. (2011). A Hierarchical Unification of LIRICS and VerbNet Semantic Roles. *Proc. of the Workshop on Semantic Annotation for Computational Ling. Resources (SACL-ICSC 2011),* 483-489. IEEE. Palo Alto, Californa, USA. ISBN: 978-1-4577-1648-5.

Bonial, C., Bonn J., Conger K., Hwang J., Palmer M. and Reese N. (2015). *English PropBank Annotation Guidelines.* http://propbank.github.io/

Bonial, C., Conger, K., Hwang, J.D., Mansouri, A., Aseri, Y., Bonn, J., O'Gorman, T. and Palmer, M. Current Directions in English and Arabic PropBank. N. Ide and J. Pustejovsky (eds.). *Handbook of Linguistic Annotation,* 737-769. Springer Netherlands.

Buitelaar, P. (1998). CoreLex: An Ontology of Systematic Polysemous Classes. *Formal Ontology in Information Systems: Proceedings of the 1st International Conference (FOIS'98), Trento, Italy*. IOS Press Amsterdam. The Netherlands. ISBN: 9051993994.

Bunt, H.C., Petukhova, V. and Schiffrin, A. (2007). LIRICS Deliverable D4.4. Multilingual test suites for semantically annotated data. http://lirics.loria.fr

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. *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06),* 355-358. European Language Resources Association (ELRA). Genoa, Italy. ISBN: 2-9517408-2-4.

Civit, M., Aldezabal, I., Pociello, E., Taulé, M., Aparicio, J. and Márquez, L. (2005). 3LBLEX: léxico verbal con frames sintáctico-semánticos. *Procesamiento del Lenguaje Natural*, 35, 367-373. Print ISSN: 1135-5948. Online ISSN: 1989-7553.

Dorr, B.J. (2001). *LSC Verb Database, Online Software Database of Lexical Conceptual Structures and Documentation*. University of Maryland.

Estarrona, A. (2014). *EPEC corpusa predikatu-mailan etiketatzeko oinarriak: EPEC-RolSem, BVI eta e-ROLda.* Basque Language and Communication, Basque Country University (UPV-EHU), Donostia. PhD Thesis.

Estarrona A., Aldezabal I., Díaz de Ilarraza A. and Aranzabe M.J. (2016). Methodology for the semiautomatic annotation of EPEC-RolSem, a Basque corpus labelled at predicate level following the PropBank/VerbNet model. Edward Vanhoutte (ed.) *Digital Scholarship in the Humanities* (2016) 31 (3): 470-492. DOI: http://dx.doi.org/10.1093/llc/fqv010. First published online: 17 June 2015 (23 pages). Published by Oxford University Press on behalf of EADH: The European Association for Digital Humanities (Online ISSN 2055-768X - Print ISSN 2055-7671). https://academic.oup.com/dsh/article/31/3/470/1745349

Fellbaum, C. (1998). *WordNet, An Electronic Lexical Database.* MIT Press. Cambridge. ISBN: 0-262-06197-X.

García-Miguel, J. and Albertuz, F.J. (2005). Verbs, Semantic Classes and Semantic Roles in the ADESSE Project. In Katrin Erk, Alissa Melinger and Sabine Schulte im Walde (eds.). *Proc. of Workshop on the Identification and Representation of Verb Features and Verb Classes,* 50-55. Saarbrücken, Germany.

Gardent, C. and Cerisara, C. (2010). Semi-Automatic Propbanking for French. *Proceedings of the Ninth International Workshop on Treebanks and Linguistic Theories,* 67-78. Northern European Association for Language Technology (NEALT). Tartu, Estonia. Print ISSN: 1736-8197. Online ISSN: 1736-6305.

Grishman ,F., Macleod , C. and Meyers, A. (1994). Comlex Syntax: building a computational lexicon. *Proceedings of the 15th conference on Computational linguistics (COLING'94),* 1, 268-272. Association for Computational Linguistics (ACL). Kyoto, Japan.

Hajic, J., Panevová, J., Urešová, Z., Bémová, A., Kolárová, V. and Pajas, P. (2003). PDT-VALLEX: Creating a Largecoverage Valency Lexicon for Treebank Annotation. In Nivre, J. and Hinrichs, E. (eds.). *Proc. of the Second Workshop on Treebanks and Linguistic Theories*, 57–68. ISBN: 9176363945 9789176363942.

Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25(4): 398–436. Oxford University Press. Print ISSN: 0950-3846. Online ISSN: 1477-4577.

Kingsbury, P. and Palmer, M. (2003). PropBank: The Next Level of Treebank. *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003),* 3. ISBN: 9176363945 9789176363942.

Kipper, K. (2005). *VerbNet: A Broad-coverage, Comprehensive Verb lexicon.* U. of Pennsylvania. PhD Thesis.

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

Laka, I. (1996). *A Brief Grammar of Euskara, the Basque Language*. University of the Basque Country. ISBN: 84-8373-850-3. http://www.ehu.es/grammar

Laparra, E. and Rigau, G. (2010). eXtended WordFrameNet. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias (eds.). *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10),* 1214-1219. European Language Resources Association (ELRA). ISBN: 2-9517408-6-7.

Levin, B. (1993). *English Verb Classes and Alternations. A preliminary Investigation.* The University of Chicago Press. Chicago and London. ISBN: 0-226-47533-6.

Marcus, M., Santorini, B. and Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics Journal*, 19:2, 313-330. MIT Press Journals. ISSN: 0891-2017.

Merlo, P. and Van der Plas, L. (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*, 288–296. Association for Computational Linguistics (ACL). Suntec, Singapore.

Monachesi, P., Stevens, G. and Trapman, J. (2007). Adding semantic role annotation to a corpus of written Dutch. *Proceedings of the Linguistic Annotation Workshop (LAW'07),* 77-84. Association for Computational Linguistics (ACL). Prague, Czech Republic.

Palmer, M., Gildea, D. and Kingsbury, P. (2005a). The Proposition Bank: A Corpus Annotated with Semantic Roles. *Computational Linguistics Journal*, 31:1, 71-106. MIT Press Journals. ISSN: 0891-2017.

Palmer, M., Nianwen, X., Babko-Malaya, O., Chen, J. and Snyder, B. (2005b). A Parallel Proposition Bank II for Chinese and English. *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, 61-67. Association for Computational Linguistics (ACL). Ann Arbor, Michigan, USA.

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

Palmer, M., Babko-Malaya, O., Bies, A., Diab, M., Maamouri, M., Mansouri, A. and Zaghouani, W. (2008). A Pilot Arabic PropBank. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis and Daniel Tapias. *Proceedings of the Sixth Conference on International Language Resources and Evaluation (LREC'08)*, 3467-3471. European Language Resources Association (ELRA). Marrakech, Morocco. ISBN: 2-9517408-4-0.

Pociello, E., Agirre, E. and Aldezabal, I. (2010). Methodology and Construction of the Basque WordNet. *Language Resources and Evaluation Journal*, 45:2, 121-142. Springer. Print ISSN: 1574-020X. Online ISSN: 1574-0218.

Pradhan, S., Hovy, E., Marcus, M.P., Palmer, M., Ramshaw, L.A. and Weischedel, R.M. (2007). OntoNotes: A Unified Relational Semantic Representation. *International Journal of Semantic Computing.* 1:4, 405-419. World Scientific. Print ISSN: 1793-351X. Online ISSN: 1793-7108.

Pustejovsky, J. (1995). *The Generative Lexicon.* Cambridge, London: MIT Press. ISBN: 9780262661409.

de Rijk, R. (1969). Is Basque an SOV language? *Fontes Linguae Vasconum 1*, 319-351. Gobierno de Navarra. Institución Príncipe de Viana. ISSN: 0046-435X.

Salaberri, H., Arregi, O. and Zapirain, B. (2014). First approach toward Semantic Role Labeling for Basque. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)*, 1387-1393. European Language Resources Association (ELRA). Reykjavik, Iceland. ISBN: 978-2-9517408-8-4.

Schiffrin, A. and Bunt, H.C. (2007). LIRICS Deliverable D4.3. Document compilation of semantic data categories. http://lirics.loria.fr

Talmy, L. (1985). Lexicalization patterns: Semantic structure in lexical forms. In Shopen T. (ed.), *Language Typology and Syntactic Description: Vol 3. Grammatical Categories and the Lexicon*, 3, 36-149. Cambridge University Press, Cambridge.

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. Procesamiento del Lenguaje Natural, 37, 75-82. Print ISSN: 1135-5948. Online ISSN: 1989-7553.

Van Der Plas, L., Samardžić, T. and Merlo, P. (2010). Cross-lingual validity of PropBank in the manual annotation of French. *Proceedings of the 4th Linguistic Annotation Workshop (LAW IV '10)*, 113-117. Association for Computational Linguistics (ACL). ISBN 978-1-932432-72-5 / 1-932432-72-8.

Vázquez, G. and Fernández, A. and Martí, M.A. (2000). *Clasificación Verbal. Alternancias de Diátesis.* Quaderns de Sintagma 3. Edicions de la Universitat de Lleida. Lleida. ISBN: 84-8409-067-1.

Xue, N. and Palmer, M. (2009). Adding semantic roles to the Chinese Trrebank. *Natural Language Engineering*, 15:1, 143-172. Cambridge University Press. ISSN: 1351-3249. EISSN: 1469-8110.

Zipf, G. (1949). *Human behavior and the principle of least effort.* Addison-Wesley Press. ISBN-13: 978-1614273127. ISBN-10: 161427312X