Basis for the annotation of EPEC-RolSem

Izaskun Aldezabal
IXA NLP Group
Basque Philology Department
University of the Basque Country
Donostia, Spain
izaskun.aldezabal@ehu.es

Abstract

In this paper I present the linguistic resources used when annotating the Reference Corpus for the Processing of Basque EPEC, in terms of semantic roles, argument structure and verb senses (EPEC-RolSem). When facing the annotation at any level, some crucial decisions have to be made, such as the model to be adopted and the criteria for the adaption of such model. Among other reasons, the fact of having the resources I am presenting here has led us to select the PropBank/Verbnet style model (Palmer *et al.*, 2005, Kipper 2005). Concretely, these resources are: the translation of all the verbs in Levin (1993) into Basque and an in-house database with syntactic/semantic subcategorization frames (ssf) for Basque verbs (EADB–Data Base for Basque Verbs), similar to the mentioned models. By means of the first resource and based on the Levin's class, we have linked the Basque verbs with the related PropBank/Verbnet verbs, getting all the corresponding information. On the other hand, the ssf-s of the EADB have been very useful to associate the appropriate English verb sense as well as to define an entry in the PropBank/Verbnet style for the Basque Verb.

1 Introduction

In the Ixa research group[1], the Reference Corpus for the Processing of Basque EPEC is being tagged at many linguistic levels, starting from morphosyntax to semantics, including some pragmatic features (Aduriz *et al.*, 2006). In each level, certain models and tagging manuals have to be developed for the annotators. In the case of semantics, the most difficult task is to establish criteria to define senses in a coherent and understandable way to facilitate the annotation process. Our previous work on semantics, when treating nouns, has mainly focused on Wordnet fine-grained senses (Felbaum, 1998), having as a result the Basque Wordnet (EusWordnet) (Pociello et al, forthcoming) and the Basque Semcor (EuSemcor) (Agirre et al., 2006a). Nevertheless, for the annotation of verbs, this fine-grained orientation has been questioned as some works point out (Ide and Véronis, 1998). This way, our data-base for Basque verbs (EADB) has been built with a more general perspective: although senses are defined for each specific verb, they are thought to be valid across verbs, based on Levin's (1993) methodology but mainly following Vázquez et al.'s (2000) alternation criteria. Consequently, more coarse-grained senses, similar to cognitive categories, are proposed for each verb entry. In addition, with the aim of defining alternations (either general or language specific ones), the syntactic realizations of the arguments in each sense are also taken into account (see section 2). With all this in mind, the PropBank/Verbnet style model (Palmer *et al.*, 2005, Kipper 2005) was thought to be a suitable one for adopting for Basque, as shown in Agirre et al.(2006b).

Other reasons have also persuaded us to choose the PropBank/Verbnet model:

1. The PropBank project starts from a syntactically annotated corpus, as we do.
2. Given the VerbNet lexicon and the annotations in PropBank, many implicit decisions according to problematic issues like argument/adjunct selection for distinguishing each verb senses are settled by examples, and

[1] ixa.si.ehu.es

seem therefore easier to replicate when we tag the Basque data. Moreover both (PropBank and Verbnet) resources are complementary as each one has appropriate and different linguistic information for defining verbs and for learning them automatically (Merlo and Van der Plas, 2009).

3. The PropBank model is being developed in other languages such as Chinese (Palmer and Xue, 2003), Spanish and Catalan (Civit et al., 2005a), Dutch (Monachesi et al, 2007), French (Van der Plas et al., 2010) and Russian (Civit et al., 2005b). Having corpora in different languages annotated following the same model makes it possible to carry out crosslingual studies, as it is demonstrated in Korhonen et al. (2010).
4. In the Verb Index[2], the information regarding PropBank and Verbnet is linked for many verbs. There is also information about other models such as Framenet (Baker et al, 1998), Wordnet (Fellbaum, 1998), Ontonotes (Hovy et al, 2006). This way, we could enrich Basque verbal models with the richer information currently available for English.

In this paper I present the main linguistic resources used for the predicate labelling of the EPEC corpus. In section 2, I explain the work carried out to translate the Levin's (1993) English verbs into Basque and I show the linking with the PropBank/Verbnet information. In section 3, I describe the syntactic-semantic frames (ssf) used to define verb entries in the EADB and the way of adopting the entry in the PropBank/Verbnet style. Finally, in section 4, the current situation and future work are outlined.

2 Translation of Levin's (1993) verbs into Basque and linking them to PropBank/Verbnet

Levin (1993) has been a reference to analyze verbs in other languages. She claims that the distinctive behavior of verb classes with respect to the diathesis alternations arises from their meaning: "once such a class is identified its members can be examined to isolate the meaning component they have in common. Thus, the diathesis alternations can be used to provide a probe into the elements entering into the lexical representation of word meaning" (Levin, 1993:14).

Many works have been carried out to compare the alternations she proposes for English with the ones existing in other languages (Jones et al., 1994; Taulé, 1995; Saint-Dizier 1995), Basque among them (Aldezabal, 1998, 2004**).** The studies carried out for Basque show that from the 80 Levin's alternations 24 are found in Basque. One reason for that is that some of the alternations in Levin are specific to a few English verbs. It has also to be pointed out that the dialectal variation was not considered, and some works reveal that, the dative alternation (or similar to it) appears in the dialect of the North part (Etxepare and Fernandez, 2011). However, those alternations that occur in specific dialects seem to be more a shade of the meaning of the sentence and they do not seem to be so useful for doing semantic classes (following Levin's methodology). Besides, in other languages also occur that many alternations do not exist. However, the alternations that are found in both languages are relevant enough for doing big classes, what we precisely do in our EADB (see section 3).

Anyway, when comparing the alternations only the verbs in the examples of the alternations were taken into account. All the verbs are found in the second part of Levin's book, where she describes the semantic classes resulting from the shared alternations. Therefore, in order to make a complete comparison, all of them were studied. The first task for that was to identify the equivalent verb in Basque, and then to ensure the differences and similarities both at alternation and class level.

2.1 Translation criteria and some examples

The translations in each class were done based on the *Morris* dictionary (Morris, 1998) and applying two general criteria:

- First of all, the meaning of the Levin's semantic class was considered.
- Then, the most syntactically similar equivalent(s) was/were selected.

In many cases, the verb in the class and the alternations involved are shared in both languages. For instance, most of the verbs in the 45.1 "*break* verbs" class with its prototypical causative/inchoative alternation can be translated without any difficulty; many of the verbs in the 9.1 "*Put* verbs" class do not either show any difficulty to be translated. Only we find the fact that for one English verb we can use more than one equivalent. For instance, the verb `break` in the 45.1 "*Break* Verbs" class is translated with the three synonyms `hautsi`, `puskatu`, `apurtu`,

[2] http://verbs.colorado.edu/verb-index/index.php

because they three mean the same (regarding the class meaning) and admit the causative/inchoative alternation without showing differences at this level. These cases are not problematic since each of the Basque verbs will be linked to the *break* verb when annotating the corpus.

However, as we went going down in the subclasses, some mismatches were found. These are illustrated in this section.

- Some of the syntactic properties are shared. For instance, the verbs `tell` and `say` which differ between them in the different behavior when admitting the dative alternation (`tell` admits (I tell sb sth / I tell sth to sb)) while `say` does not (*I say sb sth), are expressed with the same equivalent verb in Basque: `esan`. In Basque, there is not a different verb that reflects this syntactic variation, and the valence properties are the same as in the both English verbs. In those cases, both English verbs should be assigned to the Basque `esan` verb when annotating the corpus.

- A single word is used in English while two are necessary in Basque. These are mostly those verbs that have a manner or an instrument meaning incorporated, such us many of verbs into the "*funnel* verbs", "*wipe* verbs", "*spray/load* verbs", "*drive* verbs", "*poison* verbs", and "verbs of instrument of communication", among others. E.g.: `ladle`: `burduntzaliaz bota`, literally meaning 'to throw with a ladle'.

In these cases it may happen that the concept expressed by the verb that lexicalizes the manner in English is not a lexicalized concept in Basque. The example above represents that case. Therefore, these verbs should have to be considered and analyzed into the single verb (`bota throw`) class (17.1), where manner is going to be a possible adjunct. These cases are more difficult to solve when annotating: the annotator should realize that the non lexicalized adjunct + verb expression in Basque has to be annotated with the appropriate single word in English.

- The same verb is used in Basque for different verbs in different classes, but the object candidates must be specified as in English. For instance, Verbs in the 13.4.2 "*Equip* verbs" class in Basque are translated with the same verb as in 13.1 "give" or 13.2 "contribute" verb classes (for example, `charge` (13.4.2): `ardura eman` -> lit. 'to give (13.1) the charge'). That is, in Basque such distinction does not exist ('to charge somebody with a task' but not *'to give somebody with a task'). However, in order to provide the equivalent for `charge`, the object (`zeregin baten ardura`: 'charge of a task') must be equally specified in Basque. In these cases, when annotating the corpus the verb *charge* should be used when in the Basque sentence appear "*give the charge (of a task)*" and it should also be considered a multiword.

2.2 Linking to PropBank/Verbnet

Taking into account all these phenomena, we are able to say *a priory* that when linking the Propbank/Verbnet equivalent to the Basque verb, the argument structures (at least at valence level) of English and Basque verbs are not going to be the same in some cases, and, as a consequence, neither the alternations involved on them.

Moreover, when the concepts are not lexicalized in Basque, there will be an element that will be appearing as an apart adjunct (and not as an argument) in Basque, while in English there will not be a syntactic counterpart (but it will be incorporated in the verb).

In any case, the information obtained from the linking regarding the sense and rolesets will be very helpful in the process of building the Basque verb entry with the PropBank/Verbnet scheme (although classes are not shared). In table 1 a list of some of the verbs after the linking based on Levin's classes is shown.[3]

glue	22.4	*erantsi, kolatu*
go	47.7	*joan*
go	51.1	*joan*
gobble	38	*glu-glu egin*
gobble	39.3	*irentsi*

Table 1: the link between the PropBank and Basque verbs based on Levin's (1993) class.

3 The EADB: data-base of syntactic-semantic frames (ssf) for the Basque verbs

Following the methodology that Levin suggests in her work, the crucial task is to detect those alternations that are semantically sensitive and then find the semantic components that would be in the lexical representation of the verbs.

For this task, and taking a revised point of view of the alternation concept which is also

[3] It has to be noted that the Levin's classes have been revised in PropBank/Verbnet. Consequently some verbs remained without any link (Aldezabal et al, 2010).

adopted in other works (Vázquez et al., 2000; Rebolledo, 2002), I studied 100 Basque verbs basing on real corpus examples (Aldezabal, 2004).

I concluded that each verb has one or more prototypical frames to express any of the general semantic values appearing when analyzing verbs in general. These semantic values are not senses in the way they appear in the dictionaries, but basic cognitive categories or general predicate types which can serve to propose big classes at semantic level (such as *change of state*, *change of position*, *activity of an entity*, *creation of an entity*, *assignment of an attribute*, *exchange of an entity*, *situation of an entity* and so on). This semantic information is expressed by general semantic roles (or semantic components) coherently combined (that is, for a verb to express the general predicate *change of state*, at least an *affected theme* must be contain; or for a verb to express the general predicate *creation of an entity*, a *created theme* must be contain, and so on)[4]. This way, some verbs share the capability to represent the same general predicate. However, it does not mean neither they should have the same syntactic frames (although it happens in many cases), nor they share the same alternations (although it also happens in many cases).

Based on that assumption, I defined a number of syntactic-semantic frames (ssf) for each verb. Each ssf is formed by semantic roles and the declension case that syntactically realizes this role. The ssf-s that have the same semantic roles define a verbal coarse-grained sense and are considered syntactic variants of an alternation. Different sets of semantic roles reflect different senses. This is similar to the PropBank model, where each of the syntactic variants (similar to a frame) pertains to a verbal sense (similar to a roleset).

In Table 2 we can see an example of the ssf-s for the verb esan. It has two senses and the first one contains two syntactic variants. The first sense can be translated as 'tell/say' as in Levin's 37 "Verbs of communication" class, and the second sense as 'call', as in Levin's 29 "Verbs with Predicative Complements" class.

esan-1 (= 'tell/say'): Activity (communication) of an entity. Two arguments in two syntactic variants:
esan-1.1: arg1_ERG[5], arg2_ABS
esan-1.2: arg1_ERG, arg2_COMP
esan-2 (= 'call'): Assignment of an attribute. Three arguments in a single syntactic realization:
esan-2: arg1_ERG, arg2_ABS, arg3_DAT

Table 2. Syntactic-semantic frames for the verb esan (='tell/say/call') as provided by the EADB lexicon.

These ssf-s together with the information obtained from the link to PropBank/Verbnet are a robust basis to define the new lexical entry with the PropBank/Verbnet scheme and to go on tagging the EPEC corpus in such framework.

Table 3 shows the adopted PropBank/Verbnet entry for the verb *esan*.

Basque verb: esan	
say.01/tell.01	***call.01***
Arg0: Agent *(ERG)* Arg1: Topic *(ABS/COMP)* Arg2: Recipient *(DAT)* Arg3: Attributive *(INS[6]/-i buruz/ -i erreferentzia eginez[7]/...)*	Arg0: Agent *(ERG)* Arg1: Theme *(DAT)* Arg2: Predicate *(ABS)*

Table 3: The PropBank/Verbnet style entry for the verb *esan*.

4 Current situation and future lines

We have already annotated a sample of sentences for each of the 100 verbs including in the EADB. During the tagging process some adjustments had to be made, because of differences both at multiword level and at valence level. For instance, in some verbs of motion an *extend* argument is defined for the English verb while in Basque it does not exist.

Besides, the annotation has been evaluated and one of the most significant conclusions has been that before annotating, taggers must clearly understand the entries that have been adapted to the PropBank/Verbnet model. In addition, it must be also taken into account that multiword expressions are problematic and that it is necessary to decide what to do with those cases. Moreover, in order to avoid confusions with modifiers, it is important to provide some information or guidelines, although we know that some things will remain unsolved since they are subjective.

At present, we are planning to automatize the annotation-process taking into account the lexi-

[4] I propose 13 general predicates and 21 semantic roles in total.

[5] ERG: ergative declension case; ABS: absolutive declension case; COMP: completive clause; DAT: dative declension case.

[6] The instrumental declension case

[7] These are complex prepositions meaning 'regarding', 'with respect to'…

con resulting from the annotated corpus. As a first step, we will detect the univocal role_case pairs, and then we will automatically annotate the occurrences of the corpus, including its corresponding verb sense. For the automatic annotation of new verbs, class based cross studies will be carried out.

References

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

Agirre E., Aldezabal I., Etxeberria J., Iruskieta M., Izagirre E., Mendizabal K., Pociello E. 2006. (Agirre et al 2006a). A methodology for the joint development of the Basque WordNet and Semcor. *Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC)*. Genoa (Italy).

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

Aldezabal I., Aranzabe M., Díaz de Ilarraza A., Estarrona A. 2010. 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). European Language Resources Association (ELRA). Valletta.

Aldezabal I. 2004. *Aditz-azpikategorizazioaren azterketa. 100 aditzen azterketa zehatza, Levin (1993) oinarri harturik eta metodo automatikoak baliatuz*. PhD thesis. Basque Philology Department Leioa. UPV-EHU.

Aldezabal I. 1998. Levin's verb classes and Basque. A comparative approach. Colloquium series of the Department of Computer Science at UMIACS (University of Maryland Institute for Advanced Computer Studies).

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

Etxepare, R. and Fernandez B. (to appear in 2011). Microparameters in Dative Constructions. Oxford Studies in Comparative Syntax. Oxford University Press. Under contract.

Fellbaum C. 1998. *Wordnet, an electronic lexical database*. MIT Press.

Hovy E., Marcus M., Palmer M., Ramshaw L. and Weischedel R. 2006. OntoNotes: The 90% Solution. *Proceedings of HLT/NAACL*. New York.

Ide N. and Véronis J. 1998. "Word Sense Disambiguation: State of the art". *Computational Linguistics*, 1998, 24 (1).

Jones D., Berwick R., Cho F., Khan Z., Kohl K., Nomura N, Radhakrishnan A., Sauerland U. & Ulicny B. 1994. *Verb Classes and Alternations in Bangla, German, English, and Korean*. Massachussetts Institute of Technology center for Biological and Computational Learning and the Artificial Intelligence Laboratory.

Kipper, K. 2005. Verbnet: *A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvnia.

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

Korhonen A., Sun L, Poibeau T. and Messiant C. 2010. Investigating the cross-linguistic potential of Verbnet-style classification. *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010).* Beijing.

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

Monachesi G., Stevens G., Trapman J. 2007. adding semantic role annotation to a corpus of written Dutch. *Proceedings of the Lingusitcs Annotation Workshop (LAW)*. Prague, Czech repuplic.

Morris M.. 1998. Morris Hiztegia.

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

Pociello E., Agirre E. and Aldezabal I. (to appear in Language Resources and Evaluation (LRE)). The Basque Lexical Knowledge Base: the Basque Wordnet.

Rebolledo M. 2002. *Estructura sintáctica y significado verbal*. Dissertation. University of Santiago de Compostela.

Saint-Dizier P. 1995. A semantic classification of French verbs based on B. Levin's approach. Research report. IRIT.

Taulé M. 1995. Representación verbal en una Base de Conocimiento Léxico. PhD. Barcelona

Van der Plas L., Samardzic T. and Merlo P. Cross-lingual Validity of PropBank in the Manual Annotation of French. *Proceedings of the Fouth Lingusitic Annotation Workshop, ACL 2010*. Uppsala, Sweden.

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