Syntactic annotation in the Reference Corpus for the Processing of Basque (EPEC): Theoretical and practical issues*

IZASKUN ALDEZABAL, MARIA JESUS ARANZABE, JOSE MARI ARRIOLA, and ARANTZA DIAZ DE ILARRAZA

Abstract

In this paper, we will describe some theoretical and practical issues raised during the construction of the Basque Dependency Treebank (BDT): the syntactic annotation of EPEC (Reference Corpus for the Processing of Basque). EPEC is a 300,000 word corpus of standard written Basque whose purpose is to be a training corpus for the development and improvement of several NLP (Natural Language Processing) tools for Basque. BDT will be the first corpus for the Basque language tagged at syntactic level. We will also present the dependency-based annotation hierarchy that we have established for the syntactic tagging. Decisions made during design of the annotation hierarchy are based on the description of Basque grammar made by Euskaltzaindia (Academy for the Basque Language). When describing dependency relations, we consider lexical units as syntactic heads. This will open up a way for us to work with semantics.

Keywords: PLEASE ADD!

1. Introduction

A treebank is a text corpus in which each sentence has been annotated with its syntactic structure. The construction of a treebank is a multidisciplinary task that, although expensive, is indispensable for the development of real applications in the field of NLP. At a purely linguistic level, a Treebank is an essential database for the study of a language given that it provides analyzed/annotated examples of real language. Besides, the linguistic study produces an improvement in the quality of several applications, such as Part-Of-Speech (POS) taggers and parsers (Collins 1997, 2000; Charniak 2000), because it provides common training and testing material allowing different algorithms to be compared and improved.

Corpus Linguistics and Linguistic Theory 5–2 (2009), 245–274 1613-7027/09/0005-245
DOI 10.1515/CLLT.2009.10

Over recent years, treebank corpora such as the Penn Treebank (Marcus et al. 1993) and the Prague Dependency Treebank (Böhmová et al. 2003) have become a crucial resource for building and evaluating natural language processing tools and applications. Abeillé (2003) explained the work for Czech, German, French, Japanese, Polish, Spanish and Turkish, to name just a few. Kakkonen (2005) presents the state of the art of dependency-based treebanks.

The Basque Dependency Treebank (BDT) is currently the Reference Corpus for the Processing of Basque (EPEC), annotated at syntactic level. EPEC is a 300,000 word corpus of standard written texts which is intended to be a training corpus for the development and improvement of several NLP tools (Bengoetxea and Gojenola 2006).

In this paper, we describe the theoretical and practical issues raised during construction of the BDT, following the Dependency Grammar theory (Tesnière 1959). This is the first formalization of the syntactic tagging of Basque that follows the Dependency Model; it should be noted that we have based our work on the syntactic description of Basque grammar made in Euskaltzaindia (1991 [1985], 1987, 1990, 1994, 1999). Using dependency relations we have formalized the main syntactic structures described in this grammar. We have also made our own decisions during the design of the annotation hierarchy. This is very important, as Sampson (2003: 23–41) says with respect to the SUSSANE corpus: "In the work of my group I have chosen the opposite priority: we treat the detail, accuracy and explicitness of annotation as more important than the quantity of material annotated, with the inevitable consequence that our treebanks have to be very small".

Dependency theory is one of the most widely used methods of conceptualizing the linguistic structure of sentences. In grammars constructed using the dependency theory (Hudson 1990; Mel'cuk 1988), syntax is handled in terms of grammatical relations between pairs of individual words, such as the relation between the subject and the predicate or between a modifier and a common noun. Grammatical relations are seen as subtypes of a general, asymmetrical dependency relation: one of the words (the head) determines the syntactic and semantic features of the combination. The syntactic structure of a sentence as a whole is built up from the dependency relations between individual pairs of words. In general terms, we take as syntactic heads the lexical elements that are involved in the dependency relations. The decision to take lexical units as a basis for any dependency relation is also motivated by the previously developed syntactic approach: the dependency-oriented surface syntax (Järvinen and Tapanainen 1997). This will allow us to extend our work to semantics in the near future. A similar approach is followed by Lin (1995).

Taking into account the literature on tagging corpora for different languages and the fact that Basque syntax has been mainly developed within the generative framework of Goenaga (1991), Eguzkitza (1993), Laka (1993), Ar-

tiagoitia (2002), Trask (2003) and Zabala (2003), we decided to focus on the specification and development of the annotation scheme to build the Basque Treebank (BDT) without attempting to justify or elaborate in any depth on any theory. Our approach is intended to provide consistent annotation to facilitate automatic exploration of linguistic data. Indeed, when designing the annotation schema we do not have any linguistic theory in mind, apart from Dependency Grammar Theory, so that, depending on the phenomena we have to deal with, we determine the most accurate relation tag; for instance, in the case of elided elements such as *pro*[1] we adopt the generative approach in order to make a deeper syntactic analysis (the main peculiarities of our annotation practice are presented in Section 4).

The remainder of this paper is organized as follows: Section 2 presents the general features of the EPEC Corpus; in Section 3 we will set out the main decisions taken in developing the syntactic annotation scheme; Section 4 describes, using examples, the annotation decisions made; and finally, some conclusions and future work are outlined in Section 5.

2. Description of the corpus

The EPEC Corpus of written Basque is a 300,000 word collection of written standard Basque. It is intended to be a reference corpus for the development and improvement of several NLP tools for Basque. A small part of this collection has been obtained from the EEBS project (http://www.euskaracorpusa.net), and the rest from Euskaldunon Egunkaria (http://www.egunero.info), the only daily newspaper written entirely in standard Basque, published in the second half of 1999 and in 2000. The articles were chosen to cover an assorted range of topics (economics, culture, international, local, opinion, politics, sports, entertainment . . .). This corpus is being used for Natural Language Processing and although it is small, it is a strategic resource for a minority language like Basque.

The corpus has been linguistically annotated at different levels: it was first morphologically analyzed by means of MORFEUS (Alegria et al. 1997) and then manually disambiguated (Aldezabal et al. 2007a). In the manual tagging, each word form of the whole corpus was assigned its corresponding analysis at the segmentation level: part-of-speech, number, definiteness and declension case. After the morphological disambiguation, other modules within the IXATI chunker (Alegria et al. 2006; Aduriz et al. 2006), such as complex postpositions, name entities, multiword lexical units and morphosyntax, were applied. The manual dependency-based syntactic annotation started at this stage. Thus, nowadays we have a Treebank for Basque of 300,000 words completely and correctly analyzed at dependency level (Aldezabal et al. 2007b; Aranzabe 2008).

We have also developed grammars and tools for automatic disambiguation (Ezeiza et al. 1998), including the disambiguation of syntactic functions (Aduriz 2000). For this purpose, we have made use of the Constraint Grammar (CG) formalism (Karlsson et al. 1995; Tapanainen and Voutilainen 1994), and stochastic methods have been also applied (Ezeiza 2003). In all cases, the correct data (the manually disambiguated data) is used both to validate the grammar and disambiguation tools as well as to apply methods of machine learning (Bengoetxea and Gojenola 2006). The linguistic information obtained in all the processes has been represented using a general stand-off schema that uses TEI-conformant feature structures (FS) coded in XML (Artola et al. 2005).

3. Framework for syntactic annotation of the corpus

Syntactic annotation is the practice of adding syntactic information to a text by incorporating markers that give information on the syntactic structure of the sentences: e.g. labelled bracketing, or symbols indicating dependency relations between words. Although they differ in the labels and, in some cases, the function of various nodes in the tree, most annotation schemes provide a similar constituency-based representation of relations between syntactic components (Abeillé 2003). In contrast, dependency schemes (Sleator and Temperley 1993; Tapanainen and Järvinen 1997; Bunt et al. 2004) do not provide a constituency analysis but rather specify grammatical relations between elements explicitly.

3.1. Constituency-based formalism

In this type of formalism, every single constituent that makes up a syntactic constituent is tagged, including the syntactic category itself; thus, the final result derives from defining the emerging constituents and their categories (noun phrases, sentences, etc.).

The most complete and widely-used syntactically annotated corpus of English (Böhmová et al. 2003), the Penn Treebank (Marcus et al. 1993), employs this type of tagging. The following is an illustration of how a sentence would be represented in this corpus:

John tried to open the window[2]

```
(S (NP (N1 (N John_NP1)))
  (VP (V tried_VVD)
    (VP (V to_TO)
      (VP (V open_VVO)
        (NP (DT the_AT)
          (N1 (N window_NN1) ) ) ) ) ) )
```

This method has three outstanding properties:

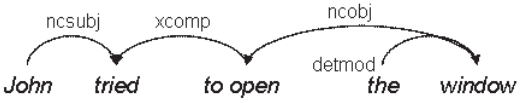
– It is based on linear word order; i.e. the order of syntactic components reflects the order in which they appear in the sentence.
– Hierarchical information is made explicit.
– The information function must be inferred.

3.2. Dependency-based formalism

Unlike the constituency-based approach, dependency-based formalism (Järvinen and Tapanainen 1997) describes the relations between the components.

This tagging formalism has been used for German (NEGRA) (Brants et al. 2003) and Czech (PDT) corpora[3], among others.

In this formalism, the representation of the sentence "John tried to open the window" above would be as follows:



The features of this method are:

– The relevance of word order is minimized.
– It is strongly based on hierarchical relations.
– The functional information is extremely important.

3.3. Constituency-based vs. dependency-based formalism

There is still an ongoing debate as to whether a constituency-based or a dependency-based formalism should be employed in completing the treebank. Some researchers have taken the middle-ground between these two options, as in Montemagni et al. (2003), who employ the dependency-based approach only to combine the basic components of the sentence (noun phrases, prepositional phrases and the verb), without reaching the word-level for dependency purposes.

The formalisms described above may be generally suitable, but the success and influence they may exert on applications depends to a great extent on the language under consideration. Based on a number of tests, set out in Skut et al. (1997), Tapanainen and Järvinen (1998) and Oflazer et al. (1999), to deal with the free word order displayed by Basque syntax, we have decided to follow the dependency-based procedure. The following issues also had a critical influence on our decision:

- Dependency-based formalism provides a way of expressing semantic relations that will constitute a good basse for tackling the next steps in the analysis-chain, such as verb valence and thematic role studies (Agirre et al. 2006).
- The nature of the computational tools we have used for the preprocessing of the corpus to be tagged facilitates the establishment of dependency relations.
- The rich information involved when using the dependency model would allow transformation from trees to other means of representation.
- From our viewpoint, it is more straightforward to evaluate the relation between the elements that make up a sentence than the relation between elements included in parentheses, since the latter involves the additional task of determining where the parentheses start and end.
- In our opinion, dependency-based formalism is a more accurate method for annotating empty elements, such as *pro*, long-distance dependencies, and discontinuous constructions.

3.4. Theoretical and methodological basis

The design of a methodology for creating a treebank has different objectives: (i) to demonstrate the varieties of syntactic patterns of a language exhaustively; (ii) to remain correlated with the latest linguistic theories; (iii) to create an annotation scheme that can be used extensively in later research activities and computer-assisted practical solutions.

According to Hinrichs and Simov (2005), the relationship between the practice of treebank annotation and linguistic theorizing has become an important subject of research. The advantage of assuming a particular theory is that it may solve many problems. The disadvantage, however, is that theories are unable to predict many aspects contained in the corpus. On the other hand, theory-neutral annotation schemes attempt to encode those grammatical properties that are distinguished by many, if not all, grammatical frameworks, without adhering to any particular linguistic theory. Theory-neutral annotations have the advantage of being more widely usable and of being less dependent on whatever version of a particular grammatical theory may have existed at the time when the treebank annotation scheme was determined. Since linguistic theories tend to change rapidly over time, and since treebank annotation is a labour-intensive and costly process, it is generally not feasible to update Treebank annotations as a particular linguistic theory begins to change.

Proponents of the theory-dependent Treebank point out that the notion of a theory-neutral annotation is in itself an illusion, since any annotation scheme is the result of at least implicit linguistic theorizing. These scholars also point out that grounding an annotation scheme on a linguistic theory

tends to improve the consistency of the annotations, at least if the theory provides explicit guidelines for the style of syntactic annotations. The Prague Dependency Treebank (Hajic 1998) is a prime example of a theory-dependent treebank.

With all these considerations in mind, and taking into account the literature on tagging corpora in different languages, we decided to focus on certain parameters for determining the theoretical and methodological basis we needed to build the Basque Treebank (BDT). The basic aspects addressed are as follows:

3.4.1. Do we follow any theory?

We follow Dependency Grammar model theory, so that for any dependency relation we establish as head and dependent units the corresponding lexical units. According to this model we annotate each word with its corresponding dependency relation tag (cf. Figure 1). In annotating we have not taken into account any specific linguistic theory; we determine the most appropriate dependency relation tag, depending on the phenomena we have to deal with. For instance, in the case of elided elements, such as *pro*, we adopt the generative approach in order to give a deeper linguistic analysis (the most important features of our annotation practice are set out in section 4).

In addition, the analyses are motivated by a precise, comprehensive and coherent theory of Basque grammar proposed by The Academy of the Basque Language (Euskaltzaindia 1991 [1985], 1987, 1990, 1994, 1999).

3.4.2. Which elements will be tagged?

Our object of study is the sentence; i.e. the text enclosed between two full stops (and also some other punctuation marks such as exclamation marks, question marks and colons).

As well as the explicit elements making up the sentence, we have also considered certain elided elements such as the *pro*.

In addition, long-distance dependencies and discontinuous constructions are also annotated; that is, multiword lexical units (e.g. *bat egin* in (1)), name-entities (e.g. *Henriette Aire* in (2)) and complex postpositions (e.g.. *kartelen artetik* in (3)), obtained by IXATI as analysis units.

(1) *Proposamen-arekin bat egin zuen Espilondo-k*
Proposal-SC.COM To join AUX-PST-3SG-SG Espilondo-ERG
'Espilondo joins the proposal.'

(2) *Henrietta Aire-k* olerki *unibertsal-ari buruzko bere*
Henrietta Aire-ERG poetry universal-DAT about her-POS
gogoeta-k azaldu-ko ditu
thought-PL explain-FUT AUX-3PL-3SG
'Henriette Aire will explain her throughts about universal poetry.'

(3) *Leiho-ko* *kartel-en* *arte-tik* *begiratzen* *du*
Window-GENLOC poster-GEN through- look AUX-3SG-3SG
'He/she looks through the posters affixed to the window.'

Furthermore, we have defined some auxiliary labels to tag units as multiword when the previous IXATI module does not treat them as such (see 4.1).

3.4.3. Which component will be the head in a dependency relation?

The criteria for establishing dependency relations, and for distinguishing the head and the dependent in such relations, are clearly of central importance for dependency grammar. Such criteria have been discussed not only in the dependency grammar tradition, but also within other frameworks where the notion of syntactic head plays an important role, including all constituency-based frameworks that subscribe to some version of *X* theory (Chomsky 1970; Jackendoff 1977). Here are some of the criteria that have been proposed for identifying a syntactic relation between a head (H) and a dependent (D) in a construction (C) (Zwicky 1985; Hudson 1990):

- H determines the syntactic category of C and can often replace C;
- H determines the semantic category of C; D gives semantic specification;
- H is obligatory; D may be optional;
- H selects D and determines whether D is obligatory or optional;
- The form of D depends on H (agreement or government);
- The linear position of D is specified with reference to H.

It is clear that this list contains a mix of different criteria, some syntactic and some semantic, and one may ask whether there is a single coherent notion of dependency corresponding to all of them. Taking into account some of these criteria, we take morphological and syntactic features, such as POS, position of Dependent and Head, Agreement, and so on to link the head and its corresponding dependent. The linguistic principles followed can be found in (Aranzabe 2008).

For instance, in the case of noun phrases (NP) and prepositional phrases (PP) the noun will be the head of such structure. In this approach we differ from the generativists who consider the determiner as the head of the NPs and the postposition as the head of the PPs.

In summary, it was decided to take lexical units as the basis of any dependency relation for the following reasons:

- The previously developed syntactic approach was the dependency-oriented surface syntax.
- In the next step we will address semantics. In our opinion, considering lexical units as syntactic heads will fit better with our semantic work.

3.4.4. The annotation scheme employed

In order to define the tagging system we have assumed the hierarchy proposed in Carroll et al. (1998). They propose an annotation scheme in which each sentence in the corpus is marked up with a set of grammatical relations, specifying the syntactic dependency which holds between each head and its dependent(s). Following this approach we have developed a tagset based on hierarchies of grammatical relations (see Figure 1). In this paper we will explain the use of this tagset.

4. The syntactic annotation

In this section we will present, using examples, the most significant features of the manual annotation process. Before explaining the use of the tagset, we will speak about the way we have annotated multiword expressions not identified by the previous computational process. We will go on to explain the annotation of a noun, heads and their dependents in a clause; and we will then give an overview of annotation in subordinate clauses.

Next, we will show the annotation in constructions such as appositions, predicative clauses, coordination and we will give some examples of the empty category *pro*.

Finally, we will present Abar-Hitz (Díaz de Ilarraza et al. 2004), the application implemented to help the annotators in their work.

4.1. Discontinuous constituents: Multiword expressions

Multiword expressions are a problematic issue both in NLP tasks and in theoretic studies. First of all, there is no clear delimitation when defining different types of multiword expressions; and even delimited, the list of multiwords is not a complete and closed one, as is the case with simple words. Thus, when tagging real corpora, it is quite common to find two or more simple words that should be treated as a single unit; that is, words that the module for the multiword treatment has not detected as such. This is the case with multiword lexical units, entities, complex postpositions and complex subordinating conjunctions.

With regard to multiword lexical units, entities and complex-postpositions, we should say that, although our automatic processing offers good precision and recall, some of them remain unidentified; and at the same time, no multiwords associated with subordinating conjunctions are detected, since there is no specific treatment for them (e.g. *hori egin bitartean* /'while doing that': a verb with a complex conjunction).

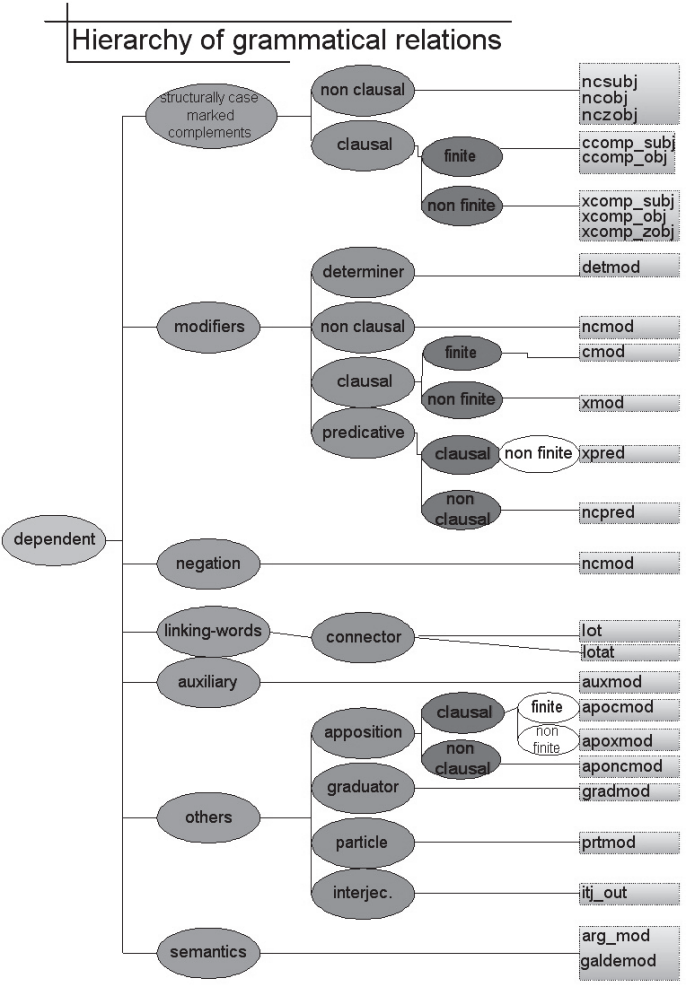


Figure 1. *Hierarchy of grammatical relations*

It is therefore necessary to define some auxiliary tags in order to analyze these kinds of multiwords as a syntactic unit. We distinguish four auxiliary tags for the four multiword expressions we are interested in:

`Haos:` components of multiword lexical units (e.g. *Ildo honetatik*/'thus': a linking word).

`Menos:` components of multiword that express subordinating conjunctions. (e.g. *hori egin bitartean* /'while doing that': a verb with a complex conjunction).

`Entios:` component of a multiword entity. (e.g. *Peru Badiola Salazar*: a proper name with its first and second surname).
`Postos:` component of a multiword postposition. (e.g. *liburuari dagokionez*/ 'regarding the book': a common name with a complex postposition).

In all of them, the annotator tags the component from left to the right. The last component of the multiword should have the corresponding common tag (e.g. *bitartean* is tagged as `xmod`, a non-finite subordinate clause). The process is the same with discontinuous components (e.g. *baita ni ere*/'me too', where *ere* is tagged as `lot`, a connector, and *baita* is tagged as `haos`).

4.2. Noun heads and their dependents inside clauses

Dependency grammatical relations corresponding to non-clauses can be described from two perspectives: i) as head of the relations (`ncsubj`, `ncobj`, `nczobj`, `ncmod`, `ncpred` and `itj_out`), and ii) as modifiers of heads (`detmod`, `ncmod`, `aponcmod` and `gradmod`).

When dealing with the structure of the non-clauses, we should say that we are not concerned with understanding the internal structure of noun phrases. We attempt to treat the phrases that are not clauses in a homogenous way.

Our approach is intended to provide consistent argument labelling that will facilitate the automatic extraction of relational data, without attempting to justify any theory.

4.2.1. Head of the non-clause and the tagging representation

Basque is what is known as a head-final language, since heads tend to be placed at the right-hand end of phrases. If we look at the structure of phrases in Basque, we can see that the morphological marker is placed in the last component of the phrase that carries it, regardless of the POS. Thus, the case marker can be attached to the head[4] of a noun-clause as in (4) (e.g. *zalantza-k*) or to a modifier of the head as in (5) (e.g. the adjective *altu-k*) and sometimes to the determiner as in (6) (e.g. *hori-ek*):

(4) *Zenbait zalantza-k ezusteko bide-tik lortu*
Some doubt-ERG unexpected way-SG-ABL solve
zuten argi-a
AUX-PST-3SG-3PL light-SG-ABS
'Some doubts were solved in an unexpected way.'

(5) *Edozein mutil altu-k egiten du*
Any box tall-ERG do-IPFV AUX-PRS-3SG-3SG
'Any tall boy does it.'

(6) *Zalantza hori-ek ezusteko bide-tik lortu*
Doubt those-ERG unexpected way-SG-ABL solve
zuten argi-a
AUX-PST-3PL-3SG light-SG-ABS
'Those doubts were solved in an unexpected way.'

In order to maintain coherence in each relation when the element carrying the declension-case/determiner and the noun head are not coincident, we decide to include both elements[5] together explicitly in the description of the relation. We consequently use a list of tuples to represent head/modifier relations in the dependency tree. For example, a structurally case-marked complement in which the complement is `nc` (non-clausal) has the following format:

- Case: the case marker by means of which the relation is established between the head and the head of the phrase.
- Head: the governor of phrase.
- Head dependent.
- Case marker: the component of the phrase that carries the case.
- Syntactic function: the syntactic label assigned to the relationship.

The analyses of the phrases included in the following sentences exemplify this formalization. In the phrase *zenbait zalantzak* in (4), *zalantzak* is the element that carries the case marker and, at the same time, it constitutes the head of phrase, so, the subject relation looks like the `ncsubj` dependency shown below.

`detmod` (- , *zalantzak, zenbait*)
`ncsubj` (erg, *lortu*, *zalantzak*, *zalantzak*, subj)

In (6), the phrase *zalantza horiek*, *zalantza* is the head of the phrase, and so we would add the component that carries the case marker, namely *horiek*. Some of the relations associated to the NP follow:

`ncsubj` (*erg, lortu, zalantza, horiek, subj*)
`detmod` *(* - , *zalantza, horiek*)

4.2.2. The dependency tags used

Regarding non-clause heads, we will distinguish two perspectives used to tag phrases: i) the relations established between the noun and the verb: `ncsubj`, `ncobj`, `nczobj`, `ncmod`, `ncpred` and `itj_out` and ii) the modifiers of noun heads: `detmod`, `ncmod`, `aponcmod` and `gradmod`.

Below we present the analysis and dependency-tree of the examples given using the aforementioned dependency tags. The description of each of the grammatical relations is extremely important, since it determines the number

and type of arguments needed for each relation (number of slots, the characteristics of each one, etc.). This work will be very useful for future treatments, for example in getting all this information into XML[6] format.

The description of all tags is presented in the appendix. We will give some of them here to provide a better understanding of the tagging format used in the dependencies.

Let us begin by giving the description of non-clausal tags. Some of them have 5 slots (`ncsubj`, `ncobj`, `nczobj`, `ncmod`[7], `ncpred`), others 4 (`ncmod`[8], `aponcmod`, `itj_out`), and others 3 (`detmod`, `gradmod`). Below we present some examples of the representation of these relations:

`ncsubj` (Case, VerbHead, Head of NP, Case-marked element within NP, Role)
`ncmod` (Case, VerbHead, Head of NP, Case-marked element within NP)
`ncpred` (-,VerbHead, Head of NP, Case-marked element within NP)
`ncmod` (-, Noun Head, Case-marked element within NP)
`detmod` (-, Noun Head, Determiner)

In the examples bellow, the above-mentioned tags are used when tagging the EPEC corpus. In this way it will better understand our annotation. For each example, we will present the sentence in Basque and the translation to English in two ways: English words in the same order as they are in the Basque source text and the correct translation. Furthermore, the complete relation set is added together with a graphical representation of the analysis tree.

A characteristic in (7) is that the elements of the phrase linked to the verb contain the case marker.

(7) *Zu-k* *galdu* *zenion* *beldurr-a* *itsaso-ari*
You-ERG lost AUX-PST-2SG-3SG-3SG fear-SG-ABS sea-DAT
txiki-txiki-tatik
childhood-ABL
'You have lost your fear of the sea since your childhood.'

Below we present the list of relations used for tagging the sentence:

`ncsubj` (erg, *galdu, zuk, zuk,* subj)
`auxmod` *(* - *, galdu, zenion*)
`ncobj` (abs*, galdu, beldurra, beldurra,* obj)
`nczobj` (dat*, galdu, itsasoari, itsasoari,* zobj)
`ncmod` (abl, *galdu, txiki-txikitatik, txiki-txikitatik*, *adlg*[9])

The dependency tree of this example can be seen in Figure 2.

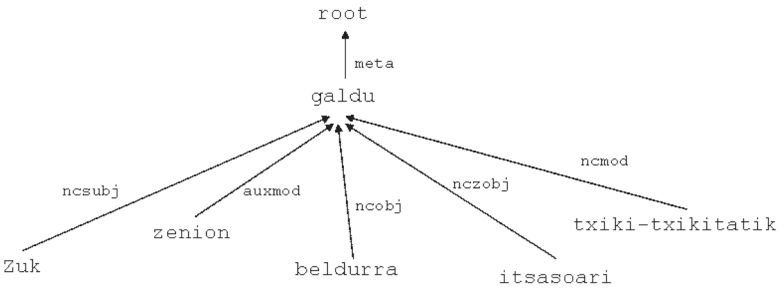


Figure 2. *Dependency tree for* Zuk galdu zenion beldurra itsasoari txiki-txikitatik

In (8) (see dependency tree in Figure 3), the noun *iritzi* is linked to the verb by means of a `ncsubj` dependency relation although the case marker is included in the determiner *hau*/'this' that modifies the noun. In this approach we make no distinction between the predicative noun and verb; this is why in this example, the noun *fruitu* is linked to the verb rather than to the noun *iritzi*.

(8) *Iritzi hau natura-ren behaketa zuzen-aren*
Opinion this-ABS nature-SG-GEN observation direct-GEN
fruitu zen
fruit-ABS be-PST-3SG
'This opinion was fruit of a direct observation of nature.'

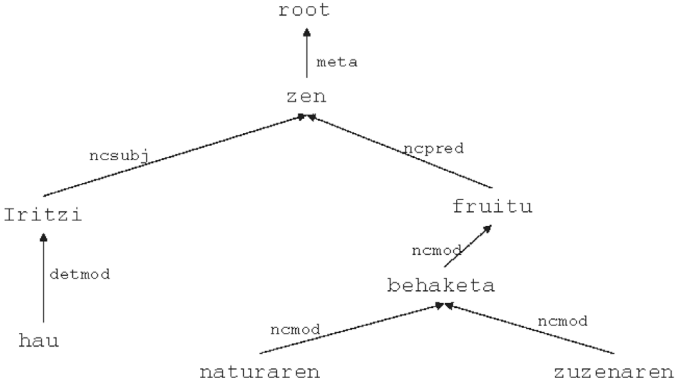


Figure 3. *Dependency tree for* Iritzi hau naturaren behaketa zuzenaren fruitu zen

In (9) (Figure 4) the `itj_out` relation is illustrated. This relation differs from the others insofar as it does not represent a common function in the sentence structure, because it is a vocative or exclamation related to the direct style, but it has been included in this group because it relates to a noun, *Valentine,* and a to a verb, *bustitzen*.

(9) *Euri-ak* *ez* *zaitu* *bustitzen* *Valentine*
Rain-SG-ERG not AUX-PRS-2SG-3SG wet-IPFV Valentine-VOC
'The rain is not wetting you, Valentine.'

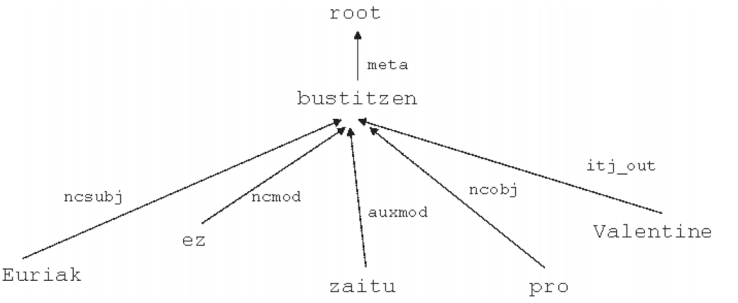


Figure 4. *Dependency tree for* Euriak ez zaitu bustitzen, Valentine

In (10) (see Figure 5) the internal relations of NP are shown; that is, the dependents of the non-clausal head. Some types of NP structures have been included in order to show their internal dependency relations. *Arrasateko* is a noun modifier and the demonstrative *hau* appears to the right of the noun while the quantifier *zenbait* and the ordinal *bigarren* precede the noun. They are both linked to the noun.

(10) *Arrasate-ko* *zenbait* *familia-k* *bigarren* *tarifa* *hau*
Arrasate-GENLOC some famili-ERG second rate this-ABS
kontratatu *zuen*
hired AUX-3SG-3SG
'Some families from Arrasate hired this second rate.'

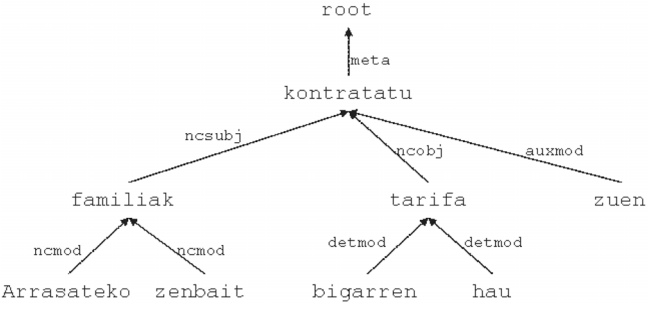


Figure 5. *The dependency tree for the sentence* Arrasateko zenbait familiak bigarren tarifa hau kontratatu zuen

In (11) (see Figure 6) we have the apposition structure classified in the hierarchy in Figure 1 as others. It represents the relation between a noun and the head of the preceding NP. In that case it is the relation between the heads of two phrases. In the modifier relation expressed by `aponcmod` the modifier is *idazle* and the head *Axularrek*.

(11) *Axularr-ek, gure idazle handi-ak idatzi zuen*
Axular-ERG our writer great-SG-ERG write AUX-PST-3SG-3SG
liburu hori
book that-ABS
'Axular, our great writer, wrote that book.'

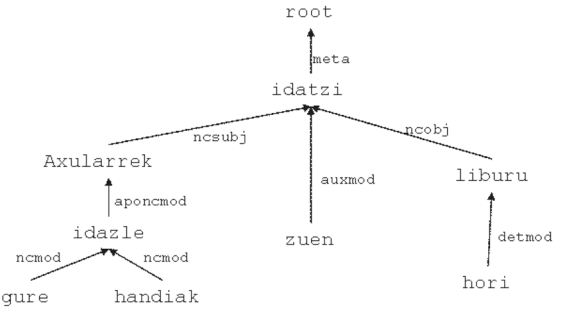


Figure 6. *The dependency tree for the sentence* Axularrek, gure idazle handiak, idatzi zuen liburu hori

Once we have revised, by means of examples, the tagset used for noun heads and their dependents inside clauses, we will explain in a similar way the relations defined for subordinate clauses.

4.3. Subordinate clauses

Subordinate clauses are divided into complement and modifier. Sentence 12 (Figure 7) exemplifies the case in which the verb of the subordinate clause (i.e. *dituela*) is finite. In this case, the verb of the subordinate clause is tagged as `ccomp` and, depending on the function it performs with respect to the main verb we will use `ccomp_subj` or `ccomp_obj` (see structure of the relations in appendices). From here to the end of the paper, we will show the relations for each example by means of their tree representation.

(12) *Gero, diote Euskal Herria-k zazpi probintzi*
Then say-PRS-3SG-3PL Euskal Herria-ERG seven province-ABS
ditu-ela
have-PRS-3SG-3PL-COMPL
'Then they say that the Basque Country has seven provinces.'

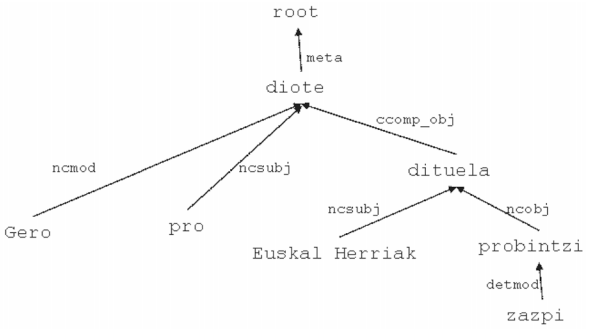


Figure 7. *The dependency tree for the sentence* Gero diote Euskal Herriak zazpi probintzi dituela

In Figure 8 we present a case of a sentence in which the verb of the subordinate clause (i.e. *edateari* in (13)) is not finite (xcomp). Non-finite is represented by the `x` in `xcomp` and, depending on the function it performs with respect to the main verb we will use `xcomp_subj`, `xcomp_obj`, or `xcomp_zobj` (see structure of the relations in appendices).

(13) *Edateari* *eman* *nion*
drink-NMLZ-SG-DAT give AUX-PST-1SG-1SG
'I started drinking.'

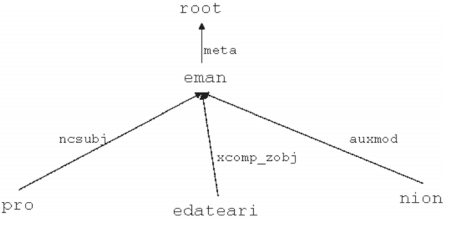


Figure 8. *The dependency tree for* Edateari eman nion

Below we show subordinate clauses that have the modifier function. Although modifier subordinates can be of different types (time, cause, etc.), we use two different variations of dependency tag, depending on the finiteness of the verb of the subordinate clause; we have therefore associated the `cmod` tag with finite verbs, (see (14)[10] in Figure 9) and `xmod` with non-finite verbs (see (15)[11] in Figure 10). We have defined the following slots for the `cmod` relation: i) clause type, ii) head of the main clause, iii) head of the subordinate clause, and iv) auxiliary carrying the relational suffix. The `xmod` relation takes the slots: i) clause type, ii) head of the main clause, iii) head of the subordinate clause, and iv) word carrying the relational suffix.

In (14), we can also see a relative clause in which the cmod relation is established between the verb *inguratzen* in the subordinate clause and the noun *likidoari* of the main clause (antecedent).

(14) *Ezaugarri hau oso erraz froge daiteke*
property this-ABS very easily prove AUX-can-3SG
koazerbatu-ak inguratzen ditu-en
coacervate-PL-ABS surround-IPFV AUX-PRS-3SG-3SG-REL
likido-ari koloratzaile desberdin-ak eransten
liquid-SG-DAT colourant different-ABS-PL add-IPFV
ba-dizkiogu
COND-AUX-PRS-1PL-3SG-3PL
'This property can be proved very easily if we add different colourants to the liquid surrounding the coacervate.'

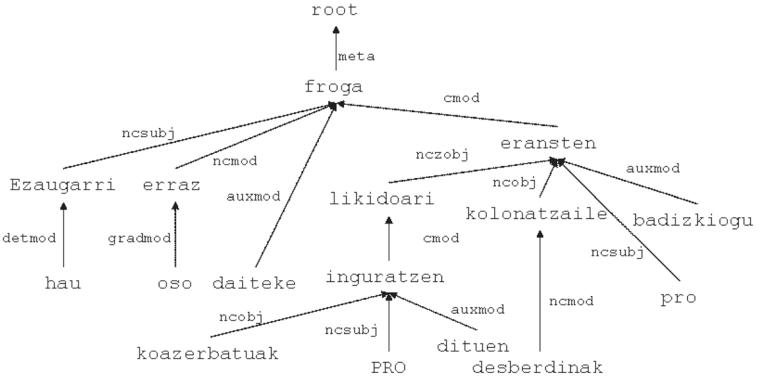


Figure 9. *The dependency tree for* Ezaugarri hau oso erraz froga daiteke koazerbatuak inguratzen dituen likidoari koloratzaile desberdinak eransten badizkiogu

In the rest of the cases, see example (15), the relation is given between two verbs (main and subordinate clause).

(15) *Gertatu-tako-az jabetzen has-tean gaizki sentitu*
Happen-REL-INS realisze-IPFV begin-TEMP terrible prove
nintzen
AUX-PST-1SG
'When I geban to realise what had happened, I felt terrible.'

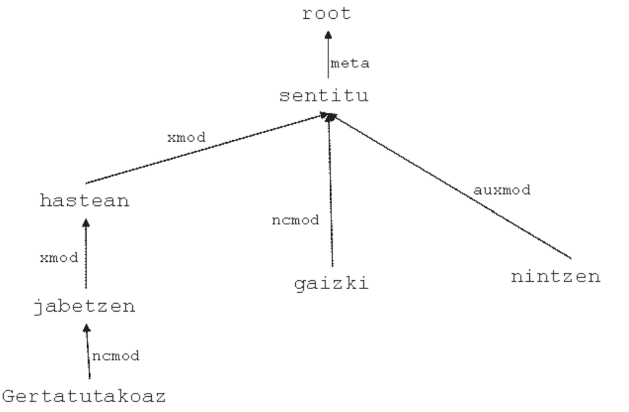


Figure 10. *The dependency tree for* Gertatutakoaz jabetzen hastean gaizki sentitu nintzen

4.4. Apposition and predicative clauses

In (16) (Figure 11) an apposition clause is shown. The apposition (represented by the `apo` abbreviation, as in the non-clause tag) is an explanation or specification of an element (either a complement or a modifier) that is the head of the apposition. In the apposition clauses there are also two different variations of `apo` dependency tag, depending on the finiteness of the verb of the subordinate clause; we have therefore associated the `apocmod` tag with finite verbs and `apoxmod` with non-finite verbs.

(16) *Jokin jokalari atzerriratu-a etorri da*
Jokin-ABS player abroad-ABS turn up AUX-PRS-3SG
'Jokin, the player who went abroad, has turned up.'

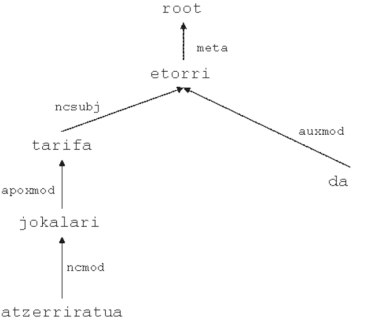


Figure 11. *The dependency tree for* Jokin, jokalari atzerriratua, etorri da

4.5. Analysis of coordination

Coordination is as problematic for Dependency Grammar formalism as for other traditional theories. In order to capture the idea that the constituents that are coordinated are at the same level, we have considered two options extensively explained in the literature (Böhmová et al. 2003; Järvinen and Tapanainen 1997): i) to presume one of the elements coordinated depends on the other and ii) to add a new imaginary node, maintaining the coordinated elements at the same level.

In our case, for computational reasons, we opt for the second one, which is expressed by considering the coordinator element as a head of the coordinate phrase; (17) (Figure 12) shows a case of noun phrase coordination that illustrates our choice.

(17) *Horixe* *zen* *magoak* *eta* *nik*
that-ABS be-PST-1SG illusionist-SG-ERG and I-ERG
genuen *sekretua*
have-PST-1PL-3SG-REL secret-SG-ABS
'That was the secret the illusionist and I had.'

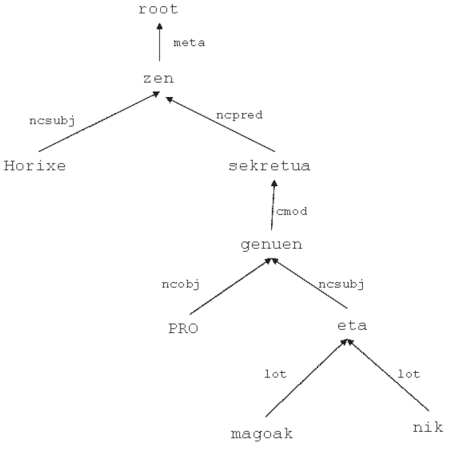


Figure 12. *Example of the dependency tree of NP coordination*

In the example, the coordinated elements *magoak* and *nik* are represented at the same level and they have as their governor the connective *eta*, which takes the dependency relation with respect to the verb, in this case `ncsubj`.

We use *emen* for copulative coordination, *aurk* for adversative, *haut* for disjunctive, *espl* for explicative and so on.

The explanation given above could be extended to the coordination of more than two elements.

4.6. The empty category pro

Basque displays a rich inflectional morphology. Indeed, it provides information about the case (Absolutive, Ergative or Dative) on either synthetic or auxiliary predicates. Interestingly, it is possible for the argument phrase corresponding to one or several case markings not to appear in the sentence (the so-called *pro*). However, precisely because the auxiliary displays case agreement with this argument (which is a possibility with the so-called *pro-drop* languages) we have assumed that this *pro* should be taken into account in the sense that it belongs to the predicate when analyzing sentences. A subset of 50,000 words of EPEC has been manually annotated taking into account the empty category as shown in (18) (Figure 13).

(18) *Begietatik igarri nionan ez*
Eye-SG-ABL figure out AUX-PST-1SG-3SG not
zela bizi
AZX.OST.3SG-COMPL live
'From his eyes I could see that he was dead.'

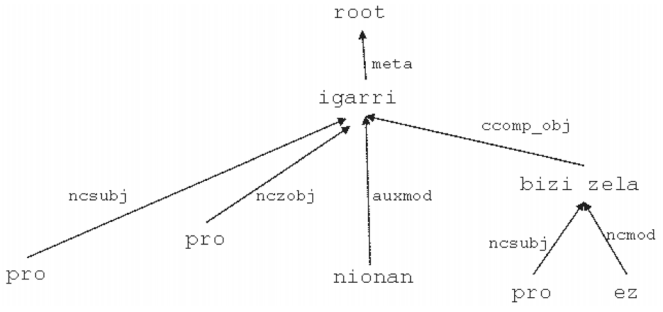


Figure 13. *The dependency tree for* Begietatik igarri nionan ez zela bizi

Once we have presented the details of the syntactic annotation we will briefly explain the annotation tool used in the manual annotation process. We have designed and implemented Abar-Hitz (Díaz de Ilarraza et al. 2004), a general and friendly tool especially designed to help in the definition of dependencies among the words of a sentence. It is important to emphasize that the design of Abar-Hitz follows the general annotation schema we established for representing linguistic information. It is part of a general environment we have developed thus far, in which general processors and resources have been integrated.

Abar-Hitz communicates with the user by means of a user-friendly interface providing the following facilities:

1) It displays the morphosyntactic information obtained thus far, which has previously been manually disambiguated in our specific tagging process. The linguist is hardly requested to take this information into account when beginning the syntactic tagging of a sentence. The tool is able to simultaneously use outputs from several tools (a morphological parser, a POS tagger and a syntactic parser) to guide the annotator's decisions.
2) It graphically presents the dependency tree for each sentence. In addition, the tree drawn can be graphically manipulated in such a way that the user can change the tags and their fields, roll up subtrees, remove/add nodes, remove/add connectors (dependencies) and so on. The changes in the tree will be automatically verified when it is explicitly required or when the window is closed.
3) It provides an environment for syntactic checking while tagging. We have to take into account that mistakes can be made while tagging, both in the number and type of slots and the name of the tag itself. Abar-Hitz avoids these mistakes by showing specific pop-up menus where the only thing the linguist can do is to select the appropriate tag.
4) It keeps track of unfinished sentences making it clear when these appear on the screen.

Finally, the table below gives some figures for the occurrences and percentages of the main dependency tags identified in EPEC:

Table 1. *Occurrences and percentages of dependency tags*

Dependency tag	Number	Percentage
ncmod	47817	34.17%
lot	18769	13.46%
auxmod	15172	10.61%
ncsubj	15287	10.73%
ncobj	11633	6.18%
detmod	7842	5.65%
xmod	5728	3.96%
xmod	4101	2.80%
ncpred	3548	2.50%
ccomp_obj	2029	1.42%
others	9361	8.53%

5. Conclusions

This paper has described the first formalization for the annotation of Basque syntax using the Dependency Grammar Theory. We have started by setting out the reasons for creating the BDT Treebank; i.e., a syntactically tagged

corpus. After considering and analyzing the principal possibilities that exist, we decided to follow the formalism based on dependency relations, basically for two reasons: first, because it is known to be more suitable for languages with a free word order, like Basque; and second, because, apart from being intuitive and easy to use, its flexibility allows new types of tags to be introduced, such as those corresponding to thematic roles. This will be an important aspect for any research we carry out in the future.

We have taken the step of analyzing the syntactic structures by explicitly expressing the relation between the head and the dependent.

Additionally, we have found solutions to problems that have emerged when describing some syntactic phenomena such us coordination, discontinuous constituents, and so on. To date, 300,000 words have been annotated. The Abar-Hitz annotation tool has been used in the annotation process. It was created taking into account the characteristics of our XML linguistic annotation.

To conclude, we would like to stress the urgent need for a syntactically tagged corpus, which would serve to evaluate and improve the parser for Basque that we are developing in the group. Furthermore, it will also be a key ingredient for syntactic studies from a theoretical point of view. The Treebank can be used to verify our linguistic intuitions.

Appendix

A) All the dependency tags (29) with their general representation, and the meaning of the abbreviations within the tags

aponcmod: (null, head, head of the apposition phrase, element with a declension case)
apocmod: (null, head, head of the apposition phrase, element with a subordinating conjunction)
apoxmod: (null, head, head of the apposition, element with a subordinating conjunction)
auxmod: (null, head, auxiliary)
ccomp_subj: (comp/indirect style, head, head of the dependent, element with a subordinating conjunction)
ccomp_obj: (comp/indirect style, head, head of the dependent, element with a subordinating conjunction)
cmod: (relation, head, head of the dependent, element with a subordinating conjunction)
detmod: (null, head of the phrase, determiner)
entios: (null, right-hand entity component, entity component)
galdemod: (null, head, reinforcing element)
gradmod: (null, head, graduator)
haos: (null, right-hand multiword component, multiword component)

itj_out: (null, head, head of the interjection, element with a declension case)
lot: (relation, conjunction, head)
lotat: (null, root, connector)
menos: (null, subordinating conjunction component, subordinating conjunction component)
ncmod: (declension case*, head *, dependent*, dependent*)
- If it is a noun: (declension case, head, head of the phrase, element with a declension case)
- If it is the negation particle: (neg, head, ez, ez)
- If it is a complex postposition phrase: (the case of the complex postposition, head, postposition, postposition)
- If it is an adverb: (null, head, adverb, adverb)
- If it is an adjective modifying a noun: (null, head of the phrase, adjective, adjective)
- If it is the left-hand component of a compound: (null, head, component of the compound, component of the compound)

ncpred: (abs/pro, head, head of the phrase, element with a declension case)
ncsubj: (erg/abs/par, head, head of the phrase, element with a declension case, subj)
ncobj: (abs/par, head, head of the phrase, element with a declension case, obj)
nczobj: (dat, head, head of the phrase, element with a declension case, zobj)
postos: (kasua, right-hand postposition component, postposition component)
prtmod: (null, head, particle)
xcomp_subj: (konp/zhg, head, element with a subordinating conjunction, element with a subordinating conjunction)
xcomp_obj: (konpl/zhg, head, element with a subordinating conjunction, element with a subordinating conjunction)
xcomp_zobj: (konpl, head, element with a subordinating conjunction, element with a subordinating conjunction)
xmod: (relation head, element with a subordinating conjunction, element with a subordinating conjunction)
xpred: (null, head, element with a subordinating conjunction, element with a subordinating conjunction)
arg_mod: (semantic role)

B) Meaning of the abbreviations within the tags

apo apposition
aux auxiliary
c finite clause
comp complement
enti entity
galde reinforcing element
grad graduator
ha multiword
itj interjection
lot conjunction

lotat connector
mod modifier
null empty
nc non clause
obj object
os component
out element out of the clause
post postposition
pred predicative
prt particle
subj subject
x non finite clause
zobj indirect object

Bionotes

Izaskun Aldezabal is Assistant Lecturer of Basque for Especial Purposes at the University of the Basque Country. She received her PhD in Basque Philology in 2004 at the University of the Basque Country. She is a researcher in the field of Natural Language Processing. Her research interests include knowledge bases, syntax and semantic of verbs. E-mail: izaskun.aldezabal@ehu.es

Maria Jesus Aranzabe is Assistant Lecturer of Basque for Especial Purposes at the University of the Basque Country. She received her PhD in Basque Philology in 2008 at the University of the Basque Country. She is a researcher in the field of Natural Language Processing. Her research interests include Treebank construction and shallow and deep syntax. E-mail: maxux.aranzabe@ehu.es

Jose Mari Arriola is Assistant Lecturer of Basque for Especial Purposes at the University of the Basque Country. He received his Master in Computational Linguistics from the University of Barcelona in 1991 and his PhD in Basque Philology in 2000 at the University of the Basque Country. He is a researcher in the field of Natural Language Processing. His research interests include shallow and deep syntax. E-mail: josemaria.arriola@ehu.es

Arantza Diaz de Ilarraza is Aggregate Professor in Computer Languages and Systems at the University of the Basque Country. She received her PhD in Computer Sciences in 1990 at the University of the Basque Country. She is a researcher in the field of Natural Language Processing. Her research interests include development of Natural Language Processing Resources, Machine Translation and Linguistic Annotations. E-mail: a.diazdeilarraza@ehu.es

Notes

* This work was partly supported by the Spanish Ministry of Education and Science (IMLT: General Model for Integration of Linguistic Tools and Resources: a proposal based on XML standards. TIN2007-63173) and the Local Government of the Basque Country (AnHITZ 2006: Language Technologies for Multilingual Interaction in Intelligent Environments, IE06-185).

1. *pro*: elided syntactic arguments that typically arise when the predicate displays agreement with the elided argument *pro* itself.

2. Example taken from Carroll et al. (1998).
3. *http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.html.*
4. Head is associated to any kind of analysis unit: multiwords or entities.
5. The decision, however, is not specific to Basque: more generally, it arises in the word-based Constraint Grammar analyzer (Karlsson et al. 1995). Our manual tagging seeks to be as compatible as possible with output obtained by the parser, for evaluation purposes. The easiest way to achieve this involved adapting the original tagset as proposed by Carroll et al. (1998), including, in some cases, an additional slot. Note that we do not change the initial dependency philosophy; we merely adapt it to our needs.
6. XML stands for *Extended Generalized Markup Language.* This is the standard and generalized language used for tagging texts, namely, a metalanguage used for specifying sets of tags as opposed to a single set of tags.
7. ncmod represents the relation between the verb and the head of the non clausal phrase.
8. ncmod represents the relation between the noun and the modifiers in the non clausal phrase.
9. Verb modifier
10. Only the analysis of the subordinate clause is provided.
11. Same as in Example 14.

References

Abeillé, Anne (ed.)
2003 *Treebanks: Building and using parsed corpora*. Dordrecht, Boston & London: Kluwer Academic Publishers.

Aduriz, Itziar
2000 *EUSMG: morfologiatik sintaxira murriztapen gramatika erabiliz. Euskararen desanbiguazio morfologikoaren tratamendua eta azterketa sintaktikoaren lehen urratsak* [EUSMG: From morphology to syntax using Constraint Grammar. Morphological disambiguation in Basque and first steps in syntax]. Donostia-San Sebastián: University of Basque Country thesis.

Aduriz, Itziar, María Jesús Aranzabe, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Illarraza, Nerea Ezeiza, Koldo Gojenola, Maite Oronoz, Aitor Soroa & Ruben Urizar
2006 Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for automatic processing. In Andrew Wilson, Paul Rayson & Dawn Archer (eds.), *Corpus linguistics around the world*, 1–15. Netherlands: Rodopi.

Agirre, Eneko, Izaskun Aldezabal, Jone Etxeberria & Elixabete Pociello
2006 A preliminary study for building the Basque PropBank. Paper presented at the International Conference on Language Resources and Evaluations, Genoa (Italy), 22–28 May.

Aldezabal, Izaskun, Klara Ceberio, Itsaso Esparza, Ainhara Estarrona, Jone Etxeberria, Mikel Iruskieta, Eli Izagirre & Larraitz Uria
2007a *EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentaziomailan etiketatzeko eskuliburua* [A guide to tag EPEC (Reference Corpus for the Processing of Basque) at segmentation level]. Technical Report, No. TR-11, 1–45. University of Basque Country (Spain).

Aldezabal, Izaskun, María Jesús Aranzabe, Jose Mari Arriola, Arantza Díaz de Ilarraza, Ainhara Estarrona, Enrique Fernandez, Uria Larraitz & Mikel Iruskieta
2007b *EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua* [A guide to tag EPEC (Reference Corpus for the

Processing of Basque) with dependences]. Technical Report, No. TR-12, 1–113. University of Basque Country (Spain).

Alegria, Iñaki, Artola Xabier & Kepa Sarasola
1997 Improving a robust morphological analyser using lexical transducers. In Ruslan Mitkov and Nicolas Nicolov (eds.), *Recent advances in Natural Language Processing*, 97–110. Amsterdam & Philadelphia: John Benjamins.

Alegria, Iñaki, Olatz Arregi, Nerea Ezeiza & Izaskun Fernandez
2006 Lessons from the development of a named entity recognizer. *Procesamiento del Lenguaje Natural* 36. 25–37.

Aranzabe, María Jesús
2008 *Dependentzia-ereduan oinarritutako baliabide sintaktikoak: zuhaitz-bankua eta gramatika konputazionala* [Linguistic resources based on dependency formalism: A treebank and a computational grammar]. Donostia-San Sebastián: University of Basque Country thesis.

Artiagoitia, Xabier
2002 The functional structure of the Basque noun phrase. *Journal of Basque Linguistics and Philology* 44. 73–90.

Artola, Xabier, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Gorka Labaka, Aitor Sologaistoa & Aitor Soroa
2005 A framework for representing and managing linguistic annotations based on typed feature structures. Paper presented at the Recent Advances on NLP (RANLP), Borovets (Bulgaria), 18–24 September.

Bengoetxea, Kepa & Koldo Gojenola
2006 *Desarrollo de un analizador sintáctico estadístico basado en dependencias para el euskera* [Development of a statistical parser for Basque]. *Procesamiento del lenguaje natural* 39. 5–12.

Böhomová, Alena, Jan Hajic, Eva Hajicová & Barbora Hladká
2003 The PDT: A 3-level annotation scenario. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 103–127. Dordrecht, Boston & London: Kluwer Academic Publishers.

Brants, Thorsten, Skut Wojciech & Hans Uszkoreit
2003 Syntactic annotation of a German newspaper corpus. In Anne Abeillé (ed.), *Treebanks: building and using parsed corpor*a, 73–87. Dordrecht, Boston & London: Kluwer Academic Publishers.

Bunt, Harry, John Carroll & Giorgio Satta Giorgio
2004 New Developments in Parsing Technology. *Text, Speech, and Language Technology* 32(3). 439–442.

Carroll, John, Ted Briscoe & Antonio Sanfilippo
1998 Parser evaluation: A survey and a new proposal. Paper presented at the International Conference on Language Resources and Evaluations, University of Granada (Spain), May.

Charniak, Eugene
2000 A maximum-entropy-inspired parser. Paper presented at the first conference of the North American chapter of the Association for Computational Linguistics, Seattle (Washington), 29 April–4 May.

Chomsky, Noam
1970 Remarks on nominalizations. In Roman Jacobs & Peter Rosenbaum (eds.), *Readings in English Transformational Grammar*, 184–221. Massachusetts: Ginn Waltham.

Collins, Michael
1997 Three generative lexicalised models for statistical parsing. Paper presented at the Eighth Conference of the European Chapter of the Association for Computational Linguistics, Madrid (Spain), 7–12 July.

Collins, Michael
2000 Discriminative reranking for natural language parsing. Paper presented at the Seventeenth International Conference on Machine Learning, Stanford (California), 29 June–2 July.

Díaz de Ilarraza, Arantza, Garmendia Aitzpea & Maite Oronoz
2004 Abar-Hitz: An annotation tool for the Basque Dependency Treebank. Paper presented at the International Conference on Language Resources and Evaluation, Lisbon (Portugal), 24 May.

Eguzkitza, Andolin
1993 Adnominals in the grammar of Basque. In José Ignacio Hualde & Jon Ortiz de Urbina (eds.), *Generative studies in Basque linguistics*, 163–187. Amsterdam & Philadelphia: John Benjamins.

Euskaltzaindia
1991 [1985] *Euskal Gramatika: Lehen Urratsak – I* [A Grammar of Basque: The first steps – I]. Bilbo: Euskaltzaindia.

Euskaltzaindia
1987 *Euskal Gramatika: Lehen Urratsak – II* [A Grammar of Basque: The first steps – II]. Bilbo: Euskaltzaindia.

Euskaltzaindia
1990 *Euskal Gramatika: Lehen Urratsak – III (Lokailuak)* [A Grammar of Basque: The first steps – III (the connectives)]. Bilbo: Euskaltzaindia.

Euskaltzaindia
1994 *Euskal Gramatika: Lehen Urratsak – IV (Juntagailuak)* [A Grammar of Basque: The first steps – IV (the conjunctions)]. Bilbo: Euskaltzaindia.

Euskaltzaindia
1999 *Euskal Gramatika: Lehen Urratsak – V (Mendeko perpausak – 1)* [A Grammar of Basque: the first steps – V (subordinated clauses – 1)]. Bilbo: Euskaltzaindia.

Ezeiza, Nerea, Itziar Aduriz, Iñaki Alegria, Jose Mari Arriola & Ruben Urizar
1998 Combining stochastic and rule-based methods for disambiguation in agglutinative languages. Paper presented at the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Montreal (Canada), 10–14 August.

Ezeiza, Nerea
2003 *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile morfosintaktiko sendo eta malgua* [Linguistic tools for exploiting corpora. The Basque Morphosyntactic robust and flexible tagger]. Donostia-San Sebastián: University of Basque Country thesis.

Goenaga, Patxi
1991 *Gramatika bideetan* [In the paths of grammar]. Donostia-San Sebastián: Erein.

Hajiè, Jan
1998 Building a syntactically annotated corpus: The Prague Dependency Treebank. In Eva Hajicová (ed.), *Studies in honour of Jarmila Panevová*, 106–132. Prague: Karolinum Charles University Press.

Hinrichs, Erhard & Kiril Simov
2005 Special issue on treebanks and linguistic theories. *Research on Language and Computation* 3(1).

Hudson, Richard
1990 *Word Grammar*. Oxford: Basil Blackwell Publishers.

Jackendoff, Ray
1997 *The architecture of the language faculty*. Cambridge, MA: The MIT Press.
Järvinen, Timo & Pasi Tapanainen
1997 *A dependency parser for English*. Technical Report, No. TR-1, 1–43. University of Helsinki (Finland).
Kakkonen, Tuomo
2005 Dependency treebanks: Methods, annotation schemes and tools. Paper presented at the 15th Nordic Conference of Computational Linguistics, Joensuu (Finland), 21–22 May.
Karlsson, Fred, Atro Voutilainen, Juha Heikkilä & Arto Anttila
1995 *Constraint grammar: A language-independent system for parsing unrestricted text*. Berlin & New York: Mouton de Gruyter.
Laka, Itziar
1993 Unergatives that assign ergative, unaccusatives that assign accusative. In Jonathan Bobaljik & Colin Phillips (eds.), *Case and agreement* (Linguistics 18), 149–172. Cambridge, MA: The MIT Press.
Lin, Dekang
1995 A dependency-based method for evaluating broad-coverage parsers. Paper presented at the International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal (Canada), 19–20 August.
Marcus, Mitchell, Beatrice Santorini & Mary Ann Marcinkiewicz
1993 Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics* 19(2). 313–330.
Mel'cuk, Igor Aleksandroviè
1988 *Dependency syntax: Theory and practice*. Albany: State University of New York Press.
Montemagni, Simonetta, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, Roberto Basili, Maria Teresa Pazienza, Dario Saracino, Fabio Zanzotto, Mana Nana, Fabio Pianesi & Rodolfo Delmonte
2003 Building the Italian Syntactic-Semantic Treebank. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 189–210. Dordrecht, Boston & London: Kluwer Academic Publishers.
Oflazer, Kemal, Zeynep Dilek & Gökhan Tür
1999 Design for a Turkish treebank. Paper presented at the Workshop on Linguistically Interpreted Corpora, Bergen (Norway).
Sampson, Geoffrey
2003 Thoughts on two decades of drawing trees. In Anne Abeillé (ed.), *Treebanks: Building and using parsed corpora*, 23–41. Dordrecht, Boston & London: Kluwer Academic Publishers.
Skut, Wojciech, Brigitte Krenn, Thorsten Brants & Hans Uszkoreit
1997 An annotation scheme for free word order languages. Paper presented at the Fifth Conference on Applied Natural Language Processing (ANLP'97), Washington (DC), 31 March–3 April.
Sleator, Daniel & Davy Temperley
1993 Parsing English with a link grammar. Paper presented at the Third International Workshop on Parsing Technologies, Tilburg, 13 August.
Tapanainen, Pasi & Atro Voutilainen
1994 Tagging accurately – don't guess if you know. In *Proceedings of Applied Natural Language Processing (ANLP'94)*, University of Stuttgart, 13–15 October.

Tapanainen, Pasi & Timo Järvinen
1997 A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington (DC), 31 March–3 April.
Tapanainen, Pasi & Timo Järvinen
1998 Dependency concordances. *International Journal of Lexicography* 11(3). 187–203.
Tesnière, Lucien
1959 *Eléments de Syntaxe Structurale*. Paris: Librairie Klincksieck.
Trask, Robert Lawrence
2003 The noun phrase: Nouns, determiners and modifiers; pronouns and names. In José Ignacio Hualde & Jon Ortiz de Urbina (eds.), *A Grammar of Basque*, 113–170. Berlin & New York: Mouton de Gruyter.
Zabala, Igone
2003 Nominal Predication. In José Ignacio Hualde & Jon Ortiz de Urbina (eds.), *A Grammar of Basque*, 426–446. Berlin & New York: Mouton de Gruyter.
Zwicky, Arnold M.
1985 Heads. *Journal of Linguistics* 21. 1–29.