Construction of a Basque Dependency Treebank

Aduriz I.*, Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A.,
Garmendia A., Oronoz M.

IXA Group (http://ixa.si.ehu.es)
University of the Basque Country (UPV/EHU)
Faculty of Computer Science
649 p.k., 20080 Donostia (The Basque Country)

*Dept. of General Linguistics. University of Barcelona
E-08007 Barcelona
jiporanm@si.ehu.es

This paper presents the process followed to build the Basque Dependency Treebank. We think that it is a necessary resource for the linguistic research in general and for the development of real applications in the area of NLP. This work is part of a general project[1] which objective is to build annotated corpora with linguistic annotation at syntactic, semantic and pragmatic levels. We annotate syntactically the Eus3LB corpus following the dependency-based formalism as explained in Aduriz *et al,* (2002). This formalism is also used in the Prague Dependency Treebank for Czech (Hajic, 1998) and in Oflazer *et al*.,(1999) and, is the one that could best deal with the free word order (Skut *et al.* 1997) displayed by Basque syntax.

Eus3LB is a corpus of standard written Basque that contains 25.000 word-forms from EPEC (Aduriz *et al*., 2003) and 25.000 words coming from newspapers that can be considered equivalent to the corpora in the other languages in the project.

In order to define the syntactic tagging system, we adopted the framework presented in Carroll *et al.* (1998, 1999). However, there are certain differences: in our system, arguments that are not lexicalised may appear in grammatical relations (for example, the phonetically empty *pro* argument which appears in the so-called *pro-drop* languages). Figure 1 shows the coding-system based on hierarchies of grammatical relations.

The tag set employed describes the most important grammatical structures such as relative clauses, causative sentences, coordination, discontinuous elements, elliptic elements and so on.

The hierarchy distinguishes between several general levels, which are further specified in subsequent levels. Thus, for instance, in the general level we find structurally case-marked complements, modifiers, negation, linking-words, auxiliaries, others and semantic relations. Some of them, for example

[1] This work is part of a general project (http://www.dlsi.ua.es/projectes/3lb) which objective is to build three linguistically annotated corpora with linguistic annotation at syntactic, semantic and pragmatic levels: Cat3LB (for Catalan), Cast3LB (for Spanish) (Civit & Martí, 2002) and Eus3LB (for Basque). The Catalan and the Spanish corpora include 100.000 words each, and the Basque Corpus 50.000 words.

complements, in turn, are divided into noun phrases (nc)[2] and finite (ccomp)[3] and non finite (xcomp)[4] clauses . Each continuous gradation achieves further specification by taking into account their grammatical function (e.g. ncsubj, ncobj, and nczobj[5]).

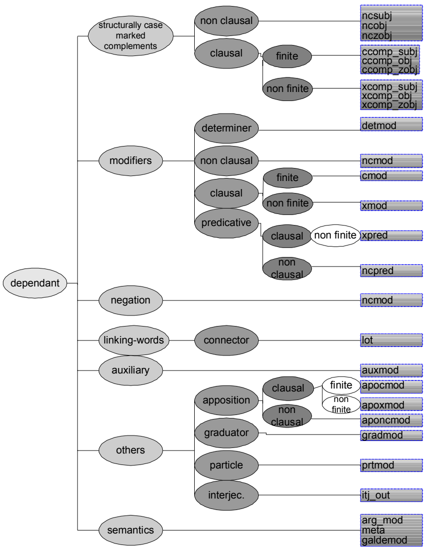


Figure 1: Hierarchy of grammatical relations

In the case of nominal phrases, it is necessary to point out that in Basque, the syntactic relation between verbs and the nominal head of the NP is realized by different categories (determiner, adjective, …). In order to represent this particularity and, considering the previous morphosyntactic analysis obtained by our tools, we add a new slot in the syntactic relation schema (in the example below, case-marked element within NP).

Next, we present an example showing the description of the grammatical relations specified in the hierarchy:

ncsubj *(Case, Head, Head of NP/dependent, Case-marked element within NP/dependent, subj)*

This is an example of structurally case-marked complement when complements are nc (non-clausal, Noun Phrases, henceforth NP). This description determines the number and type of tags needed for each relation (number of slots, the characteristics of each one, etc.).

Let us show an example by means of the sentence: *Kontu hori jakin genuenean, biziki poztu ginen*[6].

ncobj (abs, jakin, kontu, hori, obj)
detmod (- , kontu, hori)
arg_mod (- , jakin, kontu, obj)
ncsubj (erg, jakin, pro_1, pro_1, subj)

[2] nc: "*non-clausal complement*"; namely, noun and postpositional phrases.
[3] ccomp: "*clausal complement*"; namely, finite clauses.
[4] xcomp:" *clausal complement*"; namely, non finite clauses.
[5] nczobj would be equivalent to the English nciobj (non-clausal indirect object).
[6] When we learned 'Jakin genuenean' that story 'kontu hori', we were really happy 'biziki poztu ginen'

arg_mod (- , jakin, guk, subj)
auxmod (- , jakin, genuenean)
cmod (denb, poztu, jakin, genuenean)
ncmod (- , poztu, biziki, biziki)
ncsubj (abs, poztu, pro$_1$, pro$_1$, subj)
arg_mod (- , poztu, gu, subj)
auxmod (- , poztu, ginen)

The relation 'ncobj (abs, jakin, kontu, hori, obj)' shows the phenomena previously mentioned (case mark attached to the last word of the NP whatever the category of this word is). *Kontu*[7] is the NP head, but in this example it is the determiner *hori*[7] that wears the case marking. The case abs (absolutive) makes the dependency relation between the head of the NP (kontu) and the verb that is defined as head of the clause (jakin[7]). In the last slot, the syntactic function is marked.

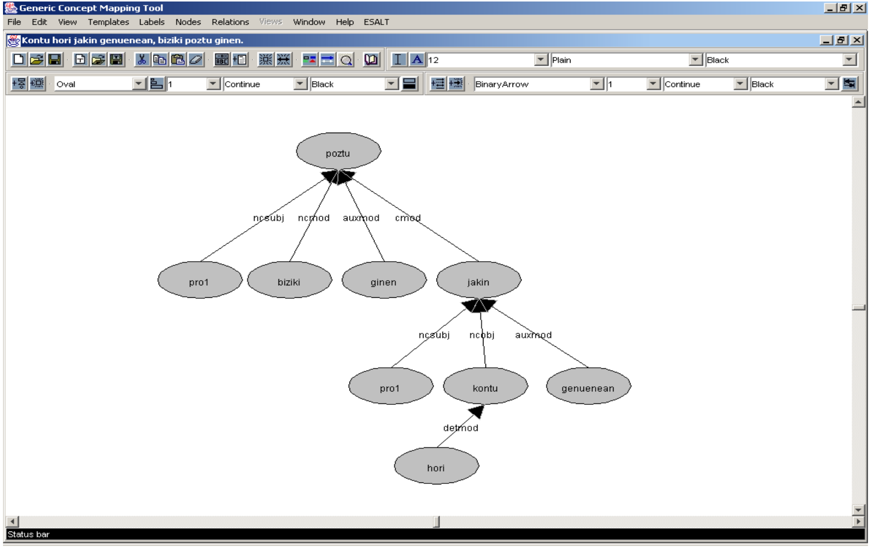
For the moment we are involved in the annotation process. This process was carried out following a methodology: i) three linguists annotated 20 sentences with the aim of adapting the tagging schema. As a result of this process the basic principles of the annotation were established[8]; ii) other two linguists annotated the same 150 sentences following the instructions of the report obtained in the first phase. These sentences cover the more representative phenomena of Basque. This task finished with a thorough description of the tagging system[8]; iii) then, other two linguists (different from the previous ones) analysed the corpus described in point ii) to become familiarised with the annotation schema and to check the proposal; iv) finally, three linguists tagged the 25.000 word-forms from EPEC contained in Eus3LB and are now working on the rest. As the tagging process goes on, and new solutions are found to arising problems, the defined tag set gets gradually improved in accuracy and robustness.

We are currently developing a computational tool, ESALT, aimed to make the manual tagging easier and faster and ensuring the syntactic correctness of the written tags. This tool provides facilities for establishing the dependencies and visualizing the resulting tree for each sentence. In figure 2, we show an example of the tree visualizer, which is based on a concept map editor (Arruarte *et al.,* 2001). Once the tree is drawn, the user can change the tags and their fields, roll up subtrees, remove/add nodes, remove/add connectors (dependencies) and so on. When the user decides that the tree is correct, ESALT can write the correct tree with the changes in an XML-tagged document.

To conclude, in this paper, we have presented: a) the definition of the tag set, b) the application of it to the corpus and c) the implementation of an annotation tool. So far, the %50 of the corpus has been deeply analysed and, in a few months, we plan to finish the tagging with the help of ESALT. In the near future, 100.000 words of a new corpus that is being compiled (http://www.hizking21.org) will be syntactically tagged.

[7] kontu (story), hori (that), jakin (learn)
[8] working reports in http://www.dlsi.ua.es/projectes/3lb

Finally, we would like to stress the urging necessity of a syntactically tagged corpus, which would serve to evaluate and improve the *parser* for Basque that we are developing in the group.



Generic Concept Mapping Tool

Kontu hori jakin genuenean, biziki poztu ginen.

Figure 2: ESALT

Acknowledgements

This research is supported by the University of the Basque Country (9/UPV00141.226-14601/2002), the Ministry of Industry of the Basque Government (project XUXENG, OD02UN52), the Interministerial Commission for Science and Technology of the Spanish Government (FIT-150500-2002-244).

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. (2003). 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. Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands. Forthcoming.

Aduriz I., Aldezabal I., Aranzabe M.J., Arrieta B., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Gojenola K., Oronoz M., Sarasola K. (2002). Construcción de un corpus etiquetado sintácticamente para el euskera. *Actas del XVIII Congreso de la SEPLN*. Universidad de Valladolid, septiembre de 2002.

Arruarte A., Elorriaga J.A., Rueda U. (2001). "A Template-Based Concept Mapping Tool for Computer-Aided Learning". T. Okamoto, R. Hartley, Kinshuk, J.P. Klus (Eds.), IEEE *International Conference on Advanced* Learning *Technologies*. IEEE Computer Society, pp. 309-312.

Carroll J., Briscoe E., Sanfilippo A. (1998). Parser evaluation: a survey and a new proposal. *Proceedings of the 1st International Conference on Language Resources and Evaluation*, 447-454. Granada, Spain.

Carroll J., Minnen G., Briscoe T. (1999). Corpus Annotation for Parser Evaluation. *Proceedings of* Workshop *on Linguistically Interpretated Corpora*, EACL´99. Bergen.

Civit M. & Martí M. (2002). Design Principles for a Spanish Treebank. *Proceedings of The Treebank and Linguistic Theories* (TLT2002). Sozopol, Bulgaria.

Hajic J. (1998). Building a Syntactically Annotated Corpus: The Prague Dependency Treebank. In *Issues of Valency and Meaning,* 106-132. Karolinum, Praha.

Oflazer K., Zeynep D., Tür H., Tür G. (1999). Design for a Turkish treebank. *Proceedings of Workshop on Linguistically Interpretated Corpora*, EACL´99. Bergen.

Skut W., Krenn B., Brants T., Uszkoreit H. (1997). An Annotation Scheme for Free Word Order Languages. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*. Washington, DC, USA.