Procesamiento del lenguaje Natural, nº 41 (2008), pp. 147-154 recibido 7-05-2008; aceptado 16-06-2008

From Dependencies to Constituents in the Reference Corpus for the Processing of Basque (EPEC)

Arantza Diaz de Ilarraza Sánchez
Enrique Fernández Terrones
University of the Basque Country
LSI Department
Manuel Lardizabal pasealekua z/g
20018 Donostia (Gipuzkoa)
{a.diazdeilarraza}{sisfetek}@ehu.es

Izaskun Aldezabal Roteta
Maria Jesús Aranzabe Urruzola
University of the Basque Country
Basque Philology Department
Sarriena auzoa z/g
48940 Leioa (Bizkaia)
{izaskun.aldezabal}{maxux.aranzabe}@ehu.es

Resumen: En este artículo se expone el proceso adoptado para la transformación de un treebank anotado con dependencias a un treebank anotado con constituyentes. En este trabajo se toma en cuenta primeramente las características de ambos formalismos, para luego proponer las correspondientes equivalencias lingüísticas. Al final se explica brevemente el desarrollo, mediante refinamientos de las equivalencias lingüísticas, llevado a cabo. La evaluación del trabajo realizado es satisfactoria ya que el resultado es que en este momento es posible explotar y trabajar con corpus anotados en los dos formalismos normalmente usados en la tarea de etiquetado sintáctico. Si las equivalencias lingüísticas son iguales, la conversión es expansible a otros corpus; de lo contrario, habría que volver a definir nuevas equivalencias.
Palabras clave: treebank, formalismo de dependencias, formalismo de constituyentes, conversión de formalismos, equivalencias lingüísticas, conversor

Abstract: In this paper the process for turning a dependency-based corpus to a constituent-based one is explained. For this purpose, first both the Dependency and the Constituent formalism are analized and then the corresponding equivalences of linguistic phenomena are treated. This process has had different phases in which the linguistic equivalences have been improved. Finally, the evaluation process is briefly explained and, as a result, we get corpora annotated in the two different formalisms usually proposed for syntactic tagging. If the linguistic equivalences are the same, the conversion process could be expanded to other corpus; otherwise, new equivalences should be defined.
Keywords: treebank, dependecy-based, constituent-based, turning of formalism, linguistic equivalents, conversor

1 Introduction

In this paper we present the process followed to build CBT (Constituent based Basque Treebank). CBT is a new syntactically annotated resource built semiautomatically from the manually annotated Dependency-based Basque Treebank (DBT). It is a resource motivated by the CESS-ECE Project (HUM2004-21127; http://clic.ub.edu/cessece) in order to get compatible the resources developed for Spanish, Catalan and Basque.

As a result, we have the Corpus syntactically tagged following the two models generally used in the annotation task, so we get flexibility when interchanging information for the development of different parser types. This kind of works has been treated in Xia & Palmer, 2001 and Civit et al, 2006, between others. In this paper we discuss decisions taken during the automatic translation from the dependency-based to the constituent-based model.

A Treebank is a text corpus in which each sentence has been annotated with its syntactic structure. The construction of a Treebank although expensive, it is indispensable for the development of real applications in the field of Natural Language

ISSN: 1135-5948 © 2008 Sociedad Española para el procesamiento del Lenguaje Natural

Processing (NLP). At a purely linguistic level, the Treebank is an essential database for the study of a language given that it provides analyzed/annotated examples of real language. In Kakkonen (2005) and Abeillé (2003) we can find the state of the art of dependency-based Treebank.

2 EPEC Corpus

The Basque Dependency Treebank (BDT) is actually the Reference Corpus for the Processing of Basque (EPEC) annotated following the dependency model. The EPEC Corpus of Basque is a 300,000 words collection of written standard Basque that has been manually tagged at different levels (morphology, surface syntax, phrases). A small part of this collection has been obtained from the EEBS project (http://www.euskaracorpusa.net), and the other from Euskaldunon Egunkaria (not accesible at this moment), the only daily newspaper written entirely in standard Basque written in the second half of 1999 and in 2000. The articles were chosen so that they covered an assorted range of topics (economics, culture, international, local, opinion, politics, sports entertainment …). This corpus is being used for Natural Language Processing and, although its small size, it is a strategic resource for a minority language like Basque.

The corpus has been morphosyntactically analyzed by means of MORFEUS (Alegria et al, 1996). Thus, each word-form of the whole corpus was assigned their every possible segmentation, without taking into account the context in which it appeared. After that, we carried out the manual disambiguation process (Aldezabal et al., 2007a) by selecting the correct segmentation and analysis.

This manually disambiguated corpus was used both to improve a Constraint Grammar disambiguator and to develop a stochastic tagger. We chose the Constraint Grammar (CG) formalism (Karlsson et al., 1995; Tapanainen & Voutilainen, 1997).

These two automatic taggers helped us in the task of manually disambiguate at lemmatization level.

The corpus manually disambiguated at lemmatization level is then processed sequentially by means of the two tools we'll briefly explain below: EIHERA and IXATI.

- EIHERA identifies entities corresponding to the categories: Institution, Person and Location (Alegria et al, 2006).
- IXATI Chunker (Aduriz et al., 2006). IXATI chunker identifies, besides verb chains and noun phrase units, complex postpositions. As far as the manually tagging process is concerned, only the detection of the latest, complex postpositions, is useful.

The dependency tagging process starts with the outcome of these tools. The linguistic information obtained in all the processes have been represented following a general stand-off schema that uses TEI-conformant feature structures (FS) coded in XML (Artola *et al*., 2005).

3 Two models: dependency-based and constituent-based

Phrase-structure theory and dependency theory are two different methods of conceptualizing the linguistic structure of sentences. Focusing on the dependency theory, we should stress that in grammars constructed following dependencies (e.g., 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. In addition, the head also controls the characteristics and placement of the other word (the dependent). The syntactic structure of a sentence as a whole is built up from such dependency relations between individual pairs of words.

On the one hand, based on a number of tests set out in Skut et al. (1997), Tapanainen & Järvinen (1998) and Oflazer et al. (1999) to deal with the free word-order languages, we decided to follow the dependency-based procedure rather than phrase-structure. On the other hand, requirements for integrating the Catalan, Spanish and Basque Treebank imposed in the framework of CESS-ECE project lead us to perform the translation to constituent-based model.

It should be noted that the formalization of the syntactic tagging that follows the

Dependency Model was the first approach done for Basque. The syntactic description of Basque has been mainly developed within the generative framework by Goenaga (1991), Eguzkitza (1993), Laka (1993), Artiagoitia (2002), Trask (2003), and other attempts have been made in general and applied linguistics (Odriozola & Zabala, 1993; Zabala, I., 2003).

3.1.1 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 most widely-used English corpus, namely the Penn Treebank, (Marcus et al., 1993) employs this sort of tagging.

This method has two outstanding properties:

1. It is based on linear word order; that is to say, the order of syntactic components reflects the order in which they appear in the sentence.
2. Hierarchical information is made explicit.

3.1.2 Dependency-based formalism

Unlike the constituency-based approach, dependency-based formalism (Järvinen & 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[1] (Böhomovà *et al*., 2003), among others.

The properties of this method are:

1. The relevance of word order is minimized.
2. It is a method strongly based on hierarchical relations.
3. The functional information is extremely important.

4 Equivalences of linguistic phenomena

In this section we will explain the two steps followed in the conversion process. First of all, we established the equivalences between constituent and dependency tags. It is known that the tags used are different depending on the criteria adopted. The constituent-based system we have based on is the one developed for Catalan and Spanish in the CESS-ECE project. We will explain these equivalences in subsection 4.1.

Secondly, the tree format has to be changed to the constituent format. This process will be briefly mentioned in the subsection 3.2.

4.1 Table of equivalences

Being our start point the dependency based annotation of the Treebank, we have split up our study of equivalences in three groups. In the first one, we deal with the tags related those elements that are classified as non-clauses; in the second one, those related to subordinated clauses; and finally, we focus on coordination. Added to that, we will mention some other equivalence needed for elements that are not considered as belonging to phrase-level.

Before going on giving details about the equivalences established in each group, let us show an example annotated following both formalisms.

(1) Dima Arratiako bailaran dago.

('Dima is in the valley of Arratia')

Dependency-style

```
ncmod (gel, bailaran, Arratiako, Arratiako)
ncmod (ine, dago, bailaran, bailaran)
ncsubj (abs, dago, Dima, Dima, subj)
```

Constituent-style

```
(S
  (sn =func:SUJ=
    (grup.nom
      (w62 Dima Dima)))
  (sp =func:CC=
    (sp =func:CC=
      (grup.nom
        (w63 Arratiako Arratia)))
    (grup.nom
      (w64 bailaran bailara)))
  (gv
    (w65 dago egon))))
```

The set of tags used in the dependency model is based on the proposal made in (Carroll *et al*., 1998) and thoroughly explained in (Aldezabal et al, 2007b; Aranzabe *et al*., 2003). About the constituent-style, some references about the tags used and the syntactic functions defined can be found in (Civit et al, 2006). Here we will only mention the syntactic functions employed in the dependency-style system. That is: subject, associated to SUJ, direct object,

[1] http://ufal.mff.cuni.cz/pcedt/doc/PCEDT_main.html

associated to CD, indirect object, associated to CI, predicative and attribute, associated to CPRED and ATR, and circumstantial complements associated to CC. Other functions such as CAG, C.REG, CCT and CCL, used only in the constituents, are not treated in this step.

4.1.1 Non-clausal phrases

In the dependency-style tags used for Basque, we have not distinguished among phrases headed by noun, adjective or adverb, neither if there is preposition or not in the phrase. We make a generalization and we consider all them as non-clausal phrases (nc). We have to mention that these non-clausal phrases have indicated their respective declension case. On the other hand, the constituent-style distinguishes the phrases having a preposition (sp) from those that have not (sn, sa, sadv).

In non-clausal phrases quite a range of categories can be the head: noun (IZE), determine or adjective -when the noun is omitted- (DET, ADJ), pronoun (IOR), adverb (ADB), and ellipsis just after the verb (ADI_IZEELI, ADT_IZEELI). Therefore, all of them have to be taken into account.

Other information in the dependency tag is the function (subject, object, indirect object, predicative and modifier). This information is given apart in the constituent-based Treebank, so all the combinations have to be defined. I.e.: ncsubj -> sn-SUJ / sa-SUJ.

There is one dependency-tag (gradmod) that not being "nc" has the same equivalence as a non-clausal modifier ("ncmod") headed by an adverb.

1[2]	2	3	4
ncsubj	IZE ADI_IZEELI ADT_IZEELI DET ADJ IOR	sn	SUJ
ncsubj	-	sn	SUJ
ncsubj	ADJ	sa	SUJ
ncobj	IZE ADI_IZEELI ADT_IZEELI	sn	CD
ncobj	DET ADJ IOR	sn	CD
ncobj	ADJ	sa	CD
nczobj	-	sp	CI
ncmod	ADJ	sa	CC
ncmod	ADB	sadv	CC
ncmod	-	sp	CC
ncpred		sn	ATR
gradmod		sadv	

Table 1: equivalences for non-clausal phrases

For instance, in the previous example (1) the second "ncmod" in the dependencies ("bailaran" 'in the valley') is equivalent to the most prominent "sp-CC" in the constituents; this "ncmod" has, at the same time, another "ncmod" inside ("Arratiako", 'of Arratia') that has been decided to map also as a "sp-CC". On the other hand, the "ncsubj" ("Dima" 'Dima') is equivalent to the "sn-SUJ" of the constituents.

4.1.2 Subordinated clauses

Regarding subordinated clauses, in the dependency-tags we distinguish between finite (c) and non-finite (x) clauses, and then the function is added (i.e. xcomp_obj for non-finite subordinated clauses that have object function). In the constituent tags there is no finiteness distinction and only the "S" tag is used. The function is added apart, and then, all the combinations have to be taken into account again.

1	2	3	4
cmod		S	CC
xmod		S	CC
xcomp_obj		S	CD
ccomp_obj		S	CD
xcomp_subj		S	SUJ
ccomp_subj		S	SUJ
xcomp_zobj		S	CI
xpred		S	ATR

Table 2: equivalences for subordinated clauses

For instance, in the example (2) the subordinated clause "ekitaldi guztiak eguraldiaren beldurrik gabe egin ahal izateko" ('so that all the events could be held without any problem') is tagged as "xmod", and it is equivalent to the "S-CC" tag of the constituents; this "xmod" has, at the same time,

[2] The meaning of the numbers is the following: 1-The dependency tag. 2- The category of the head, and sometimes also the case of the phrase. 3- The constituent tag. 4-The function assigned to the constituent.

a "xcomp_obj" inside ("ekitaldi guztiak eguraldiaren beldurrik gabe egin", 'be held without any problem') that is equivalent to the "S-CD" of the constituents.

(2) Lau estalpe ezarri dituzte Zumeltzako zelaietan, ekitaldi guztiak eguraldiaren beldurrik gabe egin ahal izateko.

('Four shelters have been put in the field of Zumeltza, so that all the events could be held without any problem')

Dependency-style

```
auxmod (-, ezarri, dituzte)
detmod (-, ekitaldi, guztiak)
detmod (-, estalpe, Lau)
ncmod (gel, zelaietan, Zumeltzako, Zumeltzako)
ncmod (gen, beldurrik_gabe, eguraldiaren, eguraldiaren)
ncmod (ine, ezarri, zelaietan, zelaietan)
ncmod (par_post_zero, egin, beldurrik_gabe, beldurrik_gabe)
ncobj (abs, egin, ekitaldi, guztiak, obj)
ncobj (abs, ezarri, estalpe, estalpe, obj)
xcomp_obj (konpl, ahal_izateko, egin, egin)
xmod (helb, ezarri, ahal_izateko, ahal_izateko)
```

Constituent-style

```
(S
  (sn =func:CD=
    (espec
      (w93 Lau lau))
    (grup.nom
      (w94 estalpe estalpe)))
  (gv
    (w95 ezarri ezarri)
    (w96 dituzte *edun))
  (sp =func:CC=
    (sp =func:CC=
      (grup.nom
        (w97 Zumeltzako Zumeltza)))
    (grup.nom
      (w98 zelaietan zelai)))
  (S =func:CC=
    (S =func:CD=
      (sn =func:CD=
        (grup.nom
          (w100 ekitaldi ekitaldi))
        (espec
          (w101 guztiak guzti)))
      (sp =func:CC=
        (sp =func:CC=
          (grup.nom
            (w102 eguraldiaren eguraldi)))
        (pos3 beldurrik_gabe gabe))
      (gv
        (w105 egin egin)))
    (gv
      (mw1 ahal_izateko ahal_izan)))))
```

4.1.3 Coordination

The coordinated elements are marked as "lot" in the dependencies, and the conjunction is the head of them, taking the corresponding function. In the constituents, the conjunction marks the coordination and the coordinated elements have their corresponding phrasal category and the function added.

Due to almost all the main category elements can be coordinated, all the specifications must be done. I.e. A "lot" element will be "sp" if the head of the phrase is a noun (IZE) and the case is neither absolutive (ABS) nor ergative (ERG); or the other way round: a "lot" element will be "sn" if the head of the phrase is a noun (IZE) and the cases are either ABS or ERG. Then, the function is specified (subj-SUJ, obj-CD…)

1	2	3
lot	ADI ADT	S
lot	IZE	-
lot	IZE, neither ABS nor ERG	sp
lot	ADB	sadv

Table 3: equivalences for coordination

For instance, in the example (3) "trenak" 'trains' and "autobusak" 'buses' are coordinated objects. Thus, in the dependencies the tag in both cases is "lot" and the conjunction "eta" is tagged as "ncobj". In the constituents, there is the conjunction "coord" coordinating the two nominal groups ("group_nom"), which are tagged as a "sn-CD".

(3) Eusko Trenbideak trenak eta autobusak jarriko ditu Donostiako geltokian.

('Eusko Trenbideak is going to put trains and buses in the Donostia station')

Dependency-style

```
auxmod (-, jarriko, ditu)
lot (emen, eta, autobusak)
lot (emen, eta, trenak)
ncmod (gel, geltokian, Donostiako, Donostiako)
ncmod (ine, jarriko, geltokian, geltokian)
ncobj (abs, jarriko, eta, autobusak, obj)
ncsubj (erg, jarriko, Eusko_Trenbideak, Eusko_Trenbideak, subj)
```

Constituent-style

```
(S
  (sn =func:SUJ=
    (grup.nom
      (ent8 Eusko_Trenbideak eusko_trenbide)))
  (sn =func:CD=
    (sn
      (grup.nom
        (w70 trenak tren)))
    (coord
      (w71 eta eta))
    (sn
      (grup.nom
        (w72 autobusak autobus))))
  (gv
```

```
    (w73 jarriko jarri)
    (w74 ditu *edun))
  (sp =func:CC=
    (sp =func:CC=
      (grup.nom
        (w75 Donostiako Donostia)))
    (grup.nom
      (w76 geltokian geltoki))))
```

4.1.4 Not phrase-level elements

Sometimes, not phrasal level elements have to be tagged, and they need to be mapped element by element. Some of them, such as "grup.nom" and "gv", can be coordinated. Therefore, they have to be grouped.

Dep. element	Const. element	Group yes/not
IZE	grup.nom	y
DET	espec	n
ITJ	interjeccio	n
LOT	coord	n
ADI ADT ADL	gv	y
PRT&lema =ez	neg	n

Tabla 4: not phrase-level elements

For instance, in the above example (3) "trenak" 'trains' and "autobusak" 'buses' are not tagged with phrase level tags (as seen in section 4.1.3) because they are in coordination; so they have to be identified by their category and then do the equivalence. In the example, "trenak" 'trains' and "autobusak" 'buses' are nouns (IZE), and their constituent equivalent are two nominal groups ("group_nom").

4.2 From tree to constituent format

Once the equivalences are well defined, a program starts analysing the dependency tree from the top to the branches. In this way, as dependency-tags are found, their corresponding constituent-based tags are being created opening brackets. Once the branch ends, the bracket is closed. Thus, we get the constituent hierarchy structure from the top level (sentence and phrase level) to word level. However, the hierarchy structure of some intermediate levels (such as group.nom) must be analyzed more deeply.

5 Evaluation

The process has been accomplished by refinements. In the first step, general equivalences were established and, accordingly to them, the conversion was done. After examining a sample of the resulting output, mistakes were solved and new refinements were faced. This sequence of steps was repeated until having satisfactory results.

As a first approach, we have manually evaluated 25 sentences of the corpus, and 5 of them failed when getting a successful constituent structure. We explain briefly the main reasons:

- Sentence connectors have not been treated in detail; then, a lot of them are not represented in the constituents.
- We have not studied in depth the correct representation of the multiword and discontinuous expressions; then, they appear separated and sometimes without any tag.
- Some phrases do not get any function since the equivalences do not cover all the possible contexts.
- Punctuation marks that make coordination are not treated as such; then they are just put as tokens in the hierarchy without any other information.

In any case, with an 80 % of correctness, we can say that the converser tool is quite robust at this stage. In the future we should improve the results solving the phenomena we have found and others that we probably have not detected yet.

6 Conclusions

In this paper the process for turning a dependency-based corpus to a constituent-based one has been explained. For this purpose, the corresponding equivalences of linguistic phenomena are treated. The process has had different phases in which the linguistic equivalences have been improved. The 300.000 words contained in EPEC have been converted. Treebank in both formats are freely available for research purposes.

Furthermore, the tool can be useful for other corpus if the linguistic equivalences are the same.

Bibliography

Abeillé, A. (2003). Treebanks: Building and Using Parsed Corpora, Kluwer Academic Publishers, Dordrecht, The Netherlands.

Aduriz, I.; Aranzabe, M. J.; Arriola, J. M.; Atutxa, A.; Díaz de Illarraza, 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 automatic processing. Corpus Linguistics Around the World. Book series: Language and Computers. Vol. 56 (pag 1- 15). ISBN 90-420-1836-4 Ed. Andrew Wilson, Paul Rayson, and Dawn Archer. Rodopi. Netherlands.

Aldezabal I., Ceberio K., Esparza I., Estarrona A., Etxeberria J., Iruskieta Quintian M., Izagirre E., Uria L. (2007a). EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) segmentazio-mailan etiketatzeko eskuliburua. UPV/EHU / LSI / TR 11-2007.

Aldezabal I., Aranzabe M., Arriola J., Díaz de Ilarraza A., Estarrona A., Fernandez K., Uria L., Quintian M. 2007b). EPEC (Euskararen Prozesamendurako Erreferentzia Corpusa) dependentziekin etiketatzeko eskuliburua. UPV/EHU / LSI / TR 12-2007.

Alegria I., Arregi O., Ezeiza N., Fernandez I. (2006). Lessons from the Development of a Named Entity Recognizer. Procesamiento del Lenguaje Natural, ISSN 1135-5948 Revista nº 36, pag. 25-37.

Alegria I., X. Artola & K. Sarasola. (1996). Automatic morphological analysis of Basque. Literary & Linguistic Computing Vol. 11, No. 4. Oxford: Oxford University Press. 193-203.

Aranzabe M., Arriola J., Atutxa A., Balza I., Uria L. (2003) Guía para la anotación sintáctica manual de Eus3LB (corpus del euskera anotado a nivel sintáctico, semántico y pragmático). *UPV/EHU/LSI/TR 13-2003*

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Labaka G., Sologaistoa A., Soroa A. A framework for representing and managing linguistic annotations based on typed feature structures. RANLP 2005. ISBN: 954-91743-3-6.

Artiagoitia, X. (2002). The functional structure of the Basque noun phrase. Erramu Boneta: Festschrift for Rudolf P.G. de Rijk. Supplements of International Journal of Basque Linguistics and Philology: 73-90.

Böhomovà, A.; Hajic, J.; Hajicova, E. and Hladka B. (2003). The PDT: a 3-level annotation scenario. In Abeillé, editor, Treebanks: Building and Using Parsed Corpora. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Brants, T.; Skut, W.; Krenn, B. and Uszkoreit, H. (2003). Syntactic annotation of a German newspaper corpus. In Abeillé, editor, Treebanks: Building and Using Parsed Corpora, Kluwer Academic Publisher, Dordrecht, The Netherlands.

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

Civit M., Martí A, Bufí N. Cat3lb and Cast3lb: from constituents to dependencies. Advances in Natural Language Processing (LNAI 4139), pp-141-153. Springer Berlag Berlin 2006.

Eguzkitza, A. (1993). Adnominals in the Grammar of Basque. In J. I. Hualde and J. Ortiz de Urbina, eds., Generative Studies in Basque Linguistics, 163-187. Amsterdam: John Benjamins.

Goenaga, P. (1991). Gramatika bideetan. Donostia: Erein.

Hudson, R. (1990). Word Grammar. Oxford, England: Basil Blackwell Publishers Limited.

Järvinen, T. and Tapanainen, P. (1997). A Dependency Parser for English. Technical Report, No. TR-1, Department of General Linguistics. University of Helsinki.

Kakkonen, T. (2005). Dependency Treebanks: Methods, Annotation Schemes and Tools. In Proceedings of the 15th Nordic Conference of Computational Linguistics. Finland.

Karlsson, F., Voutilainen, A., Heikkilä, J and Anttila, A. (1995). Constraint Grammar. Mouton Gruyter. Berlin.

Laka, I. (1993). Unergatives that Assign Ergative, Unaccusatives that Assign Accusative. In J. Bobaljik and C. Phillips, eds., Papers on Case & Agreement 1, 149-172. Cambridge: MIT Working Papers in Linguistics, Volume 18.

Marcus, M.; Santorini, B. and Marcinkiewicz, M.A. (1993). Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics, 19:313––330.

Mel'cuk, I.A. (1988). Dependency syntax: theory and practice. Albany: State University of New York Press.

Odriozola, J. C. and Zabala, I. (1993). Izen-sintagma. Idazkera teknikoa II. EHUko argitalpen zerbitzua, Bilbo.

Oflazer, K.; Zynep, D. and Tür, G. (1999). Design for a Turkish Treebank. Proceedings of Workshop on Linguistically Interpreted Corpora, at EACL'99, Bergen.

Skut, W.; Krenn, B.; Brants, T. and Uszkoreit, H. (1997). An Annotation Scheme for Free Word Order Languages, Fifth Conference on Applied Natural Language Processing (ANLP'97), Washington, DC, USA, 88-95.

Tapanainen, P. and Järvinen, T. (1997). A non-projective dependency parser. Proceedings of the 5th Conference on Applied Natural Language Processing (ANLP'97), 64-71

Tapanainen, P. and Järvinen, T. (1998). Dependency concordances. International Journal of Lexicography, 11 (3): 187-203. September.

Trask, R L. (2003). The Noun Phrase: nouns, determiners and modifiers; pronouns and names A Grammar of Basque. José Ignacio Hualde & Jon Ortiz de Urbina (eds.). Mouton de Gruyter. Berlin-New York.

Xia, F., Palmer, M. (2001). Converting dependency structures to phrase structures. In *Proc. Int. Conf. on Human Language Technology*, HLT-2001. San Diego, CA.

Zabala I. (2003). Nominal Predication J.I. Hualde y J. Ortiz de Urbina (eds.) A Grammar of Basque: 426-446. Mouton Gruyter. Berlin.