Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing

Aduriz I., Aranzabe M.J., Arriola J.M., Atutxa A., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R.*

Department of Computer Languages and Systems
Computer Science Faculty
University of the Basque Country
P.O. box 649, E-20080 Donostia
jibaregj@si.ehu.es

* Department of Linguistics
Faculty of Philology
University of Barcelona.
E-08007

ABSTRACT

In this article, we will describe the different steps in the construction of EPEC (Reference Corpus for the Processing of Basque). EPEC is a corpus of standard written Basque that has been manually tagged at different levels (morphology, surface syntax, phrases) and is currently being hand tagged at deep syntax level following the Dependency Structure-based Scheme. It is aimed to be a "reference" corpus for the development and improvement of several NLP tools for Basque. This corpus has already been used for the construction of some tools such as a morphological analyser, a lemmatiser, or a shallow syntactic analyser.

1. Introduction

When specifying the strategic priorities for the development of language technology in minority languages, Sarasola (2000) stated:

> Language foundations and research are essential to create any tool or application; but in the same way tools and applications will be very helpful in research and improving language foundations. Therefore, these three levels [applications, tools, and language foundations] have to be incrementally developed in a parallel and coordinated way in order to get the best benefit possible.

Moreover, Sarasola (2000) proposes five phases as a general strategy to follow in the processing of a language: (1) laying foundations, (2) basic tools, (3) tools of medium complexity, (4) advanced tools and multilinguality, and (5) general applications. In all the phases proposed, corpora, first raw and then tagged, outstand as an essential language resource.

In this article, we will describe the different steps in the construction of EPEC (Reference Corpus for the Processing of Basque). EPEC is a corpus of standard written Basque that has been manually tagged at different levels (morphology, surface syntax, phrases) and is currently being hand tagged at deep syntax level. It is aimed to be a "reference" corpus for the development and improvement of several NLP tools for Basque.

In section 2, we explain how the raw corpus was compiled and we briefly describe the design of the tagset. In section 3, we account for the morphological disambiguation process carried out manually over the outcome of MORFEUS (the morphological analyser for Basque). The shallow syntactic tagging and phrase tagging are explained in sections 4 and 5 respectively. Finally, in section 6 we succinctly explain the tag system chosen for the dependency-based syntactic analysis and how the treebank is being manually tagged.

In figure 1, we can see a sketch of the different phases in the construction of EPEC, contrasting the manual tasks (right column) with the computer-based ones (left column) as well as the dependencies between them.

2. The tagged corpus

2.1 Compilation of the corpus

EPEC is a 50,000-word sample collection of written standard Basque. It is a strategic resource for the processing of Basque and it has already been used for the development and improvement of some tools. Half of this collection was obtained from the *Statistical Corpus of 20th Century Basque* (http://www.euskaracorpusa.net). The other half was extracted from *Euskaldunon Egunkaria* (http://www.egunero.info), the only daily newspaper written entirely in standard Basque.

The Statistical Corpus of 20th Century Basque is a reference corpus of Basque including 4,658,036 word-forms. It was created by UZEI (http://www.uzei.com), a non-profit organisation devoted to making Basque language suitable for any specialised field. The corpus constructed was based on an exhaustive inventory of Basque publications of the 20th century, from which a random sampling was extracted. This corpus has become an invaluable linguistic reference for written Basque of this period. It was classified taking into account the following criteria: the publications were divided into 4 **periods** (1900-1939, 1940-1968, 1969-1990, 1991-1999), 6 different **dialects** (Biscayan, Guipuzcoan, Souletin, Labourdin-Navarrese, Standard Basque, and non-classified), and 14 **genres** (literary prose, poetry, theatre, administration, newspapers...). Each book or article also contained information about the **author** (or authors) and its **title**. A subcorpus of about 25,000 word-forms was extracted from this corpus in order to build EPEC. Texts

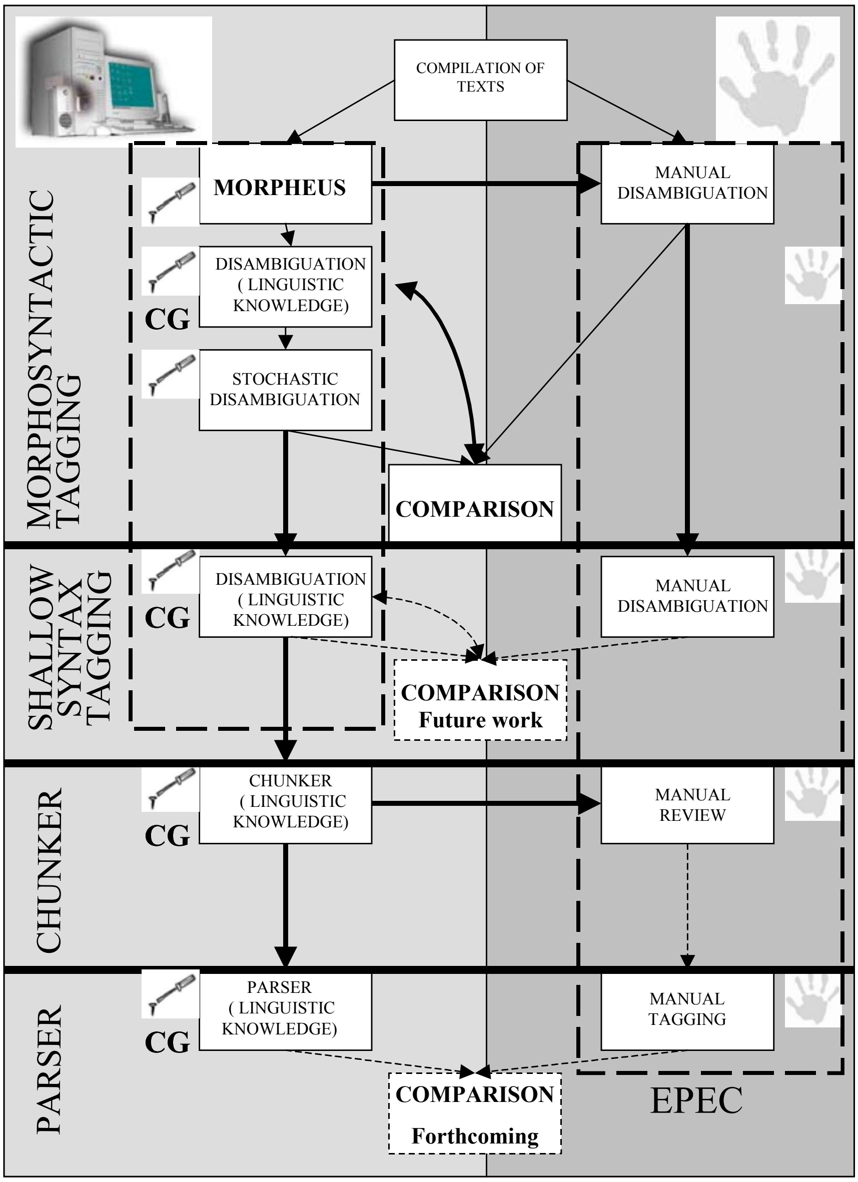


Figure 1. Sketch of the different steps in the completion of EPEC

written in standard Basque, corresponding to the last period (1991-1999) and belonging to both literary and non-literary prose, were chosen for this purpose. The second part of EPEC consists of several articles extracted from the Euskaldunon Egunkaria 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, entertainment, international, local, opinion, politics, sports...).

2.2 Design of the tagset

Choosing an appropriate tagset is a crucial task since the usefulness of further applications depend on it. The main problem we found while defining the tagset for Basque was the absence of an exhaustive one for automatic use. Moreover, Basque printed dictionaries also lacked systematisation of categories.

For the morphosyntactic treatment of Basque texts, the tag system we developed is a four level system, ranging from the simplest part-of-speech tagging scheme up to the full morphosyntactic information. In the first level, 20 general categories are included for lexical items (noun, adjective, verb, pronoun, conjunction...).

In the second one, each category tag is further refined by subcategory tags. For instance, the category 'pronoun' has 6 subcategories: common, emphatic, interrogative, indefinite, reflexive and reciprocal.

The third level includes some basic morphosyntactic information such as declension case, number, etc. This morphological information is carried by the dependent morphemes attached to the stem.

The full output of the morphosyntactic analysis constitutes the fourth level of tagging. The only difference with the previous level is that, here, all the morphological information is considered along with the tags for syntactic functions. Morphology and syntax are closely related in Basque, so most syntactic functions are provided by the database, along with the inflexional morphemes. For instance, the ergative case in Basque marks the subject in a clause (with transitive verbs) the absolutive case may either indicate the subject or the predicative (with intransitive verbs), or the direct object. The specification at this level is very detailed and constitutes the input for the morphosyntactic disambiguation process as well as for syntactic and other types of language processing.

In addition to these four levels, further tags are added to mark verb chains, noun phrases, and postpositional phrases (see sections 4.3.1. and 4.3.2.)

Nowadays, we are involved in the syntactic tagging of the corpus, following the Dependency Structure-based Scheme (see section 5). About 31 syntactic tags are being used for this purpose.

3. Morphosyntactic tagging of the corpus

A morphological analyser of words is an indispensable basic tool when defining a general framework for the automatic processing of agglutinative languages like Basque (Aduriz *et al.*, 1998). However, previous to the completion of the morphological analyser MORFEUS, the design of the tagset had to accomplished (see section 2.2) and a lexical database developed.

3.1 EDBL, a lexical database for Basque

EDBL (Aldezabal *et al.*, 2001) is a general-purpose lexical database used in Basque text-processing tools. This large repository of lexical knowledge is the basis in many different NLP tasks, and provides lexical information for several language tools including, obviously, the morphological analyser. At present, it consists of nearly 80,000 entries divided into (i) dictionary entries (the same you can find in any conventional dictionary), (ii) inflected verb forms, and (iii) dependent morphemes, all of them with their respective morphological information.

3.2 MORFEUS, automatic morphological analyser

MORFEUS is a robust morphological analyser for Basque. It is a basic tool for current and future work on NLP. The analyser is based on the two-level formalism proposed by Koskenniemi (1983), which has had widespread acceptance due mostly to its general applicability, declarativeness of rules and clear separation between linguistic knowledge and program.

The architecture of the analyser was defined using three main modules:

1 The standard analyser that uses a general lexicon and a user's lexicon. This module is able to analyse and generate standard language word-forms. In our applications for Basque, we defined more than 130 patterns of morphotactics and two rule systems in cascade, the first one for long-distance dependencies among morphemes and the second one for morphophonological changes. These elements are compiled together in the standard transducer.

2 The analysis and normalization of linguistic variants (dialectal uses and competence errors). Due to non-standard or dialectal uses of the language and competence errors, the standard morphology is not enough to offer good results when analysing real text corpora. This problem becomes critical in languages like Basque in which standardisation is in process and dialectal forms are still of widespread use. For this process the standard transducer is extended with new lexical entries and phonological rules producing the enhanced transducer.

3 The guesser or analyser of words without lemmas in the lexicons. In this case, the standard transducer is simplified removing the lexical entries and allowing the analysis of any string. Therefore, the standard transducer is substituted by a general transducer to describe any combination of characters.

The morphological analyser gives as a result *all* the possible analyses of each *token* in the text.

3.3 Manual disambiguation of the corpus

The manual disambiguation of the corpus was performed on the output of MORFEUS. Thus, the whole corpus was morphosyntactically analysed giving to each word-form every possible analysis, without taking into account the context in which it appeared. Once each word-form in the corpus was morphosintactically analysed, we carried out the manual disambiguation process. Two linguists marked independently the correct syntactic tag to each word in the corpus, applying the "double blind" method described in Voutilainen & Järvinen (1995). In case no right tag had been automatically assigned, they typed it themselves. Both linguists' answers were compared and, when differences occurred, they agreed a single tag.

This manually disambiguated corpus was used both to improve a Constraint Grammar disambiguator and to develop a stochastic tagger. After the corpus was manually disambiguated, we started to make up a grammar of constraint rules that would automatically select the correct syntactic tags in *any* real corpus. For this purpose, we chose the Constraint Grammar (CG) formalism (Karlsson *et al.*, 1995; Tapanainen & Voutilainen, 1994), which was designed with the aim of being a language-independent and robust tool to disambiguate and analyse unrestricted texts. The CG grammar statements are close to real text sentences and directly address some crucial parsing problems, especially ambiguity. The role of the CG system is to apply a set of linguistic constraints that discard as many alternatives as possible, leaving at the end as fully disambiguated sentences as possible.

Each rule produced for this grammar was checked on the manually disambiguated corpus so as to test its goodness and improve it iteratively whenever necessary. Moreover, in the cases in which the analyser didn't assign any correct analysis to a word-form in the corpus, the linguists contributed greatly to the improvement of the lexical database and the analyser itself.

Besides, we also developed a stochastic tagger. Statistical methods need little effort and obtain very good results (Church, 1998; Cutting *et al.*, 1992), at least when applied to English. In our case, we selected the TATOO tagger based on Hidden Markov Models (Armstrong *et al.*, 1995). TATOO was designed to be applied to the output of a morphological analyser and the tagset can be easily switched without changing the input text.

However, being Basque an agglutinative and free-order language, the stochastic tagger turned out to be much less accurate than for English when trained directly on the output of the morphological analyser. So, we performed a supervised training on the output of the CG grammar. Since the CG disambiguator leaves a relatively low ambiguity rate, the results of TATOO were much better. Currently, we apply a combination of the CG disambiguator with the stochastic tagger and we get good results (Ezeiza 2003). The CG disambiguator is first applied and then the remaining ambiguities are solved using the results of TATOO.

4. Shallow syntax tagging

After disambiguating the morphological tags in the corpus, the next step was to assign the corresponding syntactic tag to each word-form. Syntactic function tags follow the philosophy of the Constraint Grammar (CG) formalism in the sense that they are based on a functionally labelled dependency syntax[1]. By adopting the CG formalism, we express the syntactic functions of words and the interdependencies that exist among them rather than deep structural relations. So, the syntactic tags at this level refer to shallow syntactic functions, i.e. they may provide information about the surface structure of verb chains, noun phrases, or postpositional phrases. Therefore, this results in a shallow parsing of the corpus.

As we mentioned before, most syntactic functions are added to the word-forms together with inflectional morphemes. Morphological suffixes and syntactic functions are closely related in Basque and both are included in the database. Thus, the output of the morphological analyser displays most of these shallow syntactic tags.

However, some other syntactic tags that are not inherited from the database are added to the analysis through CG mapping rules. These functions are mostly attached to parts of speech, and they are generally assigned to word-forms provided that they comply with some given contextual conditions.

Mainly, the syntactic function tags are divided into three groups: *main functions* (subject, object, indirect object…), *modifiers* (indicating the direction relative to their head), and *verb functions* (used to detect verb chains). This distinction of the syntactic functions is essential for the tagging of the different kind of phrases (see section 5).

The ambiguity rate related to the shallow syntactic tagging is over 22%.[2], that is, for each 100 word-forms 22 are assigned more than one syntactic tag.

4.1 Manual disambiguation and applications

Once each word-form in the corpus was given at least one syntactic tag, we carried out the manual disambiguation process again. The method was similar to the one used for the morphological disambiguation in the previous step. Two linguists marked independently the correct syntactic tag to each word in the corpus or, in case no right tag had been automatically assigned, they typed it themselves. Then, both linguists agreed a single tag when differences occurred.

After the corpus was manually disambiguated, we started to make up a grammar of constraint rules that would automatically select the correct syntactic tags in *any* real corpus. Each rule produced was checked on the manually disambiguated corpus so as to test its goodness and improve it if necessary.

[1] The concept of dependency-syntax has a long tradition in grammatical analyses since the Greco-Roman era. More recently, within the application of formalisms to syntactic theory, among others we find Tésniere (1959), Hays (1964) and Mel'cuk (1988), the ones who have recovered dependency-syntax in theoretical terms.

[2] This ambiguity was estimated taking into account the syntactic functions of a subset of 200 common words.

5. Tagging phrases

At this stage we have the corpus manually tagged with surface syntactic tags following the CG syntax. No phrase units are marked yet, although based on this representation, the identification of various kinds of phrase units, such as verb chains, noun phrases, and postpositional phrases is reasonably straightforward.

5.1 Tags for verb chains

In order to detect verb chains, we use the verb function tags (@+FAUXVERB, @-FAUXVERB, @+FMAINVERB, @-FMAINVERB3…) and some particles (the negative particle, modal particles…). Based on these elements we are able to detect not only continuous verb chains but also dispersed ones.

So as to mark up continuous verb chains, the following tags are attached using again CG mapping rules:

- %VCH: this tag is attached to a verb chain composed of a single element.
- %VCHI: this is attached to the initial element of a complex verb chain.
- %VCHF: this is attached to the final element of a complex verb chain.

The tags used to mark-up the dispersed verb chains are:

- %NCVCHI: this tag is attached to the initial element of a non-continuous verb chain.
- %NCVCHC: this tag is attached to the second element of a non-continuous verb chain.
- %NCVCHF: this tag is attached to the final element of a non-continuous verb chain.

5.2 Tags for noun phrases and postpositional phrases

Our assumption is that any word having a modifier function tag is linked to some word with a main syntactic function tag. Moreover, a word with a main syntactic function tag can, by itself, constitute a phrase unit. Taking into account this assumption, we establish three tags to mark up this kind of phrase units (noun phrases or postpositional phrases):

- %PHR: this tag is attached to words with main syntactic function tags that constitute a phrase unit by themselves.
- %PHRI: this tag is attached to the initial element of a phrase unit.
- %PHRF: this tag is attached to the final element of a phrase unit.

In order to attach one of these tags to each word-form, we have simultaneously developed two subgrammars containing CG mapping rules. The first subgrammar is aimed at delimiting verb chains whereas the second one marks noun and postpositional phrases.

[3] Finite auxiliary verb, non-finite auxiliary, finite main verb, non-finite main verb…

5.3 Manual tagging and applications

At present, a linguist is checking the tags that the first set of mapping rules marked up in the corpus. Whenever necessary, she adds, removes, or changes the tags automatically assigned. Once this work is finished, the first set of mapping rules developed will be tested on the corpus and the results will be used to improve the rules iteratively as well as to develop new ones.

6. Treebank

The next logical stage in the completion of the corpus is deep syntax tagging, in order to build a treebank (Aduriz *et al*., 2002.) Although tagging manually a treebank is an expensive and time-consuming task, it is also an essential step for the development of syntactic tools and applications for Basque. A group of linguists in our research group is currently involved in this arduous job[4].

After considering a number of diverse choices −Skut *et al.* (1997), Oflazer (1999)– we decided to follow a dependency-based procedure, for it was, in our opinion, the one that could best deal with the free word order displayed by Basque syntax. The dependency-based analysis describes the relations existing between components (i.e., word-forms). This way, for each sentence in the corpus we explicitly determine the syntactic dependencies between the heads and the dependants.

In order to define the syntactic tagging system, we adopted the framework presented in Carroll *et al.* (1998, 1999). By following this line of work, we developed a coding-system based on hierarchies of grammatical relations (both for lexical and empty elements, such as *pro*) (see figure 2).

As it can be seen in figure 2, 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, thematic roles (arg_mod), modifiers, auxiliaries and conjuctions. In turn, structurally case-marked complements, for example, are divided into noun phrases and clauses. Each continuous gradation achieves further specification by taking into account their grammatical function (e.g. ncsubj, ncobj, and nczobj).

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

ncsubj (Case, Head, Head of NP, the Case-marked element within NP, subj)
ncobj (Case, Head, Head of NP, the Case-marked element within NP, obj)
nczobj[5] (Case, Head, Head of NP, the Case-marked element within NP, ind.obj)

[4] In this research line, our group is taking part in the project entitled "The IXA group, tools for an automatic treatment of Basque: creating a database composed of syntactic-semantic trees" (See 'acknowledgments')

[5] nczobj would be equivalent to the English nciobj (non-clausal indirect object).

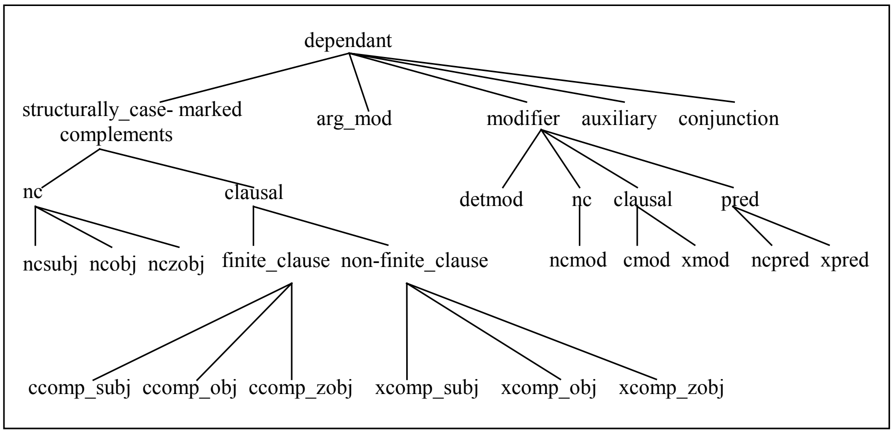


Figure 2: Hierarchy of grammatical relations.

These are examples of structurally case-marked complements when complements are nc (non-clausal, Noun Phrases, henceforth NP), as, for instance, in the sentence *Aitak haurrari sagarra eman dio* 'Father has given an apple to the child' (lit. 'Father to-child apple given has'):

ncsubj (erg, eman, **aitak**, aitak, subj)
nczobj (dat, eman, **haurrari**, haurrari, ind.obj)
ncobj (abs, eman, **sagarra**, sagarra, obj)

This description is extremely important, since it determines the number and type of tags needed for each relation (number of slots, the characteristics of each one, etc.). This formalisation will be very useful for future treatments, for example, to get all this information in XML format (see section 7).

Tagging the corpus manually has enabled us to find solution to problems that emerge in the analysing process, such as discontinuous constituents, coordination, or comparative clauses. Moreover, it is not unusual that similar phenomena are treated as distinct by the different linguists tagging the corpus. In these cases, the group of linguists tries to agree a single analysis that will be considered as correct thereafter.

Consequently, as the tagging process goes on and we find new solutions to arising problems, it will get gradually improving in accuracy, robustness, and speed. Besides, we are currently developing a computational tool aimed to make the manual tagging easier and faster.

All this work is being carried out within a project that aims at constructing treebanks for Catalan, Spanish, and Basque (Civit & Martí, 2002).

6.1 Applications

When the manual tagging of the corpus is finished, we plan to develop a tool based on linguistic knowledge that will be able to parse real corpora automatically. Like in the previous steps of manual tagging, each rule produced for the parser will be tested on the manually tagged corpus in order to assess it goodness and improve it accordingly.

In the future, we also plan to apply machine learning methods to the corpus, in order to carry out an automatic tagging.

7. Representation of the Corpus using XML

During the last three years a great effort has been done in our research group (Artola *et al.*, 2002) to integrate the NLP tools for Basque described in previous sections. Due to the complexity of the information to be exchanged among the tools, Feature Structures (FSs) are used to represent it. Feature structures are coded following the TEI's DTD for FSs, and Feature Structure Definition descriptions (FSD) have been thoroughly defined. The documents used as input and output of the different tools, contain TEI-P4-conformant feature structures (FS) coded in XML. The use of XML for encoding the I/O streams flowing between programs forces us to describe the mark-up formally, and provides software to check that these mark-up hold invariantly in an annotated corpus.

We could deeply analyse the framework for linguistic knowledge representation and integration developed in our group, but as it is not the goal of this paper, we will only show the output of the tools of the analysis chain (figure 3).

Different representations of the sentence *Noizean behin itsaso aldetik Donostiako Ondarreta hondartzara enbata iristen da* (Once in a while, a storm arrives from high seas to the Donostia's beach of Ondarreta) coded in XML are shown in figure 3.

8. Future work

The lexical information gathered in the lexical database (EDBL), which is the basis for several NLP tools in our research group, is constantly being renewed. New entries from diverse sources are periodically added to the database. Moreover, new tools such as multiword units, named entities, or postposition recognisers have been developed. These changes must be reflected in the corpus, so we must review it regularly. Therefore, in the near future, we intend to update EPEC with all these information. This will be done semiautomatically, so that only the new information needs to be reviewed.

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), and the European Community (MEANING project, IST-2001-34460).

9. References

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. (1998) A Framework for the Automatic Processing of Basque. *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada.

Aduriz I., Aldezabal I., Aranzabe M., Arrieta B., Arriola J., 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*, Valladolid, Spain.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernández G., Lersundi M. (2001) EDBL: a General Lexical Basis for the Automatic Processing of Basque. *IRCS Workshop onLinguistic Databases*, Philadelphia (USA).

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. (2002) Robustness and customisation in an analyser/lemmatiser for Basque. *Proceedings of Workshop on "Customizing knowledge in NLP applications ". Third International Conference on Language Resources and Evaluation*, Las Palmas de Gran Canaria (Spain).

Armstrong S., Russell G., Petitpierre D., Robert G. (1995) An Open Architecture for Multilingual Text Processing. *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics,* Dublin, Ireland, pp 101-106.

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Hernández G., Soroa A. (2002) A Class Library for the Integration of NLP Tools: Definition and implementation of an Abstract Data Type Collection for the manipulation of SGML documents in a context of stand-off linguistic annotation. *Proceedings of the Third International Conference on Language Resources and Evaluation,* Las Palmas de Gran Canaria, Spain.

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

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

Church K. W. (1998) A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. *Proceedings of the Second Conference on Applied Natural Language Processing,*Québec, Canada , pp 136-143.

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

Cutting D., Kupiec J., Pederson J., Sibun P. (1992) A Practical Part-of-speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing,* Philadelphia, USA, pp 133-140.

Ezeiza, N. (2003) *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile sintaktiko sendo eta malgua*. PhD thesis, University of the Basque Country.

Hays D. C. (1964) Dependency theory: a formalism and some observations. *Language 40*, pp 511-525.

Karlsson F., Voutilainen A., Heikkila J., Anttila A. (1995) Constraint Grammar: Language-independent System for Parsing Unrestricted Text. *Mouton de Gruyter*, Berlin.

Koskenniemi K. (1983) *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. University of Helsinki, Department of General Linguistics. Publications 11.

Mel'cuk I. (1988) *A Dependency Syntax: Theory and Practice*. State University of New York Press.

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

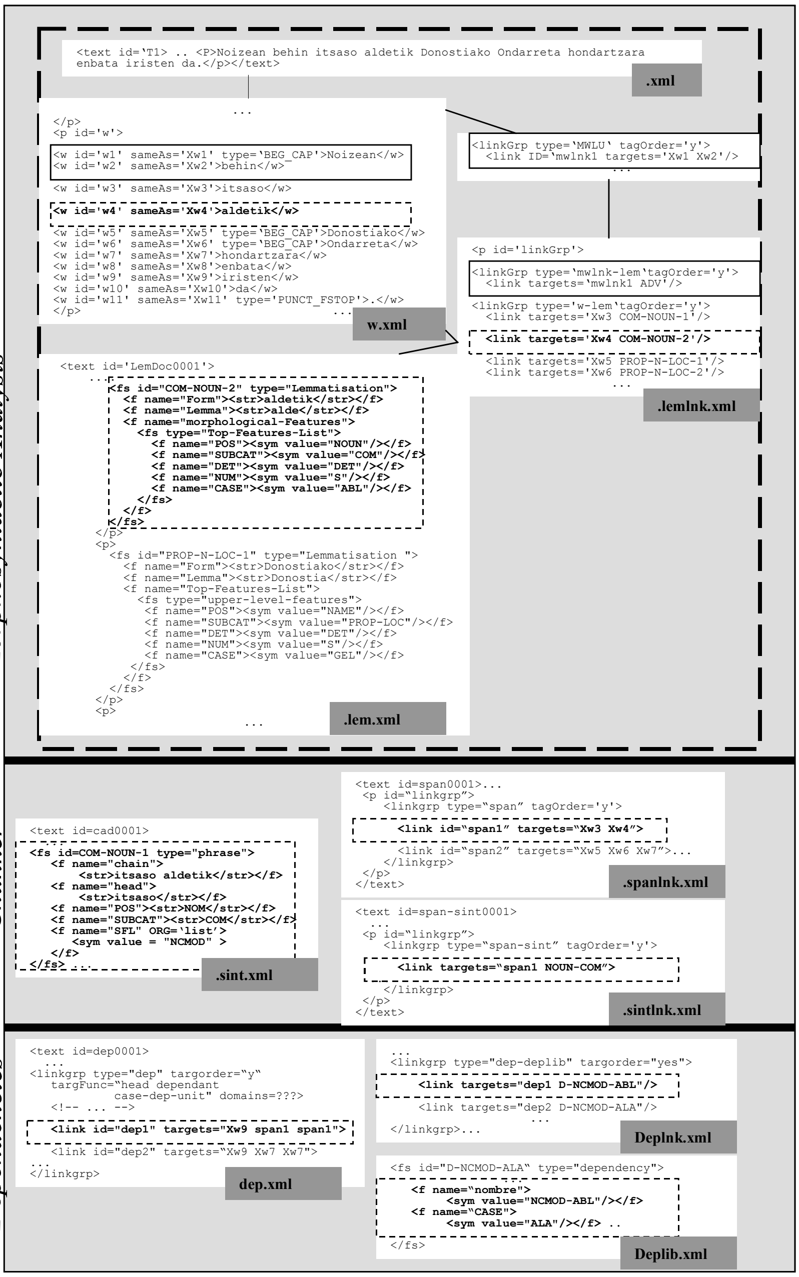
Sarasola K. (2000) Strategic priorities for the development of language technology in minority languages. *Proceedings of Workshop on "Developing language resources for minority languages: re-useability and strategic priorities". Second International Conference on Language Resources and Evaluation*, Athens, Greece.

Skut W., Krenn B., Brants T., Uszkoreit H. (1997) An Annotation Scheme for Free Word Order Languages. *Fifth Conference on Applied Natural Language Processing*, Washington, DC, USA, pp 88-95.

Tapanainen P., Voutilainen A. (1994) Tagging Accurately-Don´t guess if you know. *Proceedings of the 4th Conference on Applied Natural Language Processing*, Washington.

Tesnière L. (1959) *Elèments de Syntaxe Structurale*, (2nd ed.) Paris, Klincksieck.

Voutilainen A., Järvinen T. (1995) Specifying a shallow grammatical representation for grammatical purposes. *Proceedings of the 7th Conference of European Association of Computacional Linguistics*, Dublin.



```
<text id='T1> .. <P>Noizean behin itsaso aldetik Donostiako Ondarreta hondartzara
enbata iristen da.</p></text>
```

```
                         ...
</p>
<p id='w'>
<w id='w1' sameAs='Xw1' type='BEG_CAP'>Noizean</w>
<w id='w2' sameAs='Xw2'>behin</w>
<w id='w3' sameAs='Xw3'>itsaso</w>
<w id='w4' sameAs='Xw4'>aldetik</w>
<w id='w5' sameAs='Xw5' type='BEG_CAP'>Donostiako</w>
<w id='w6' sameAs='Xw6' type='BEG_CAP'>Ondarreta</w>
<w id='w7' sameAs='Xw7'>hondartzara</w>
<w id='w8' sameAs='Xw8'>enbata</w>
<w id='w9' sameAs='Xw9'>iristen</w>
<w id='w10' sameAs='Xw10'>da</w>
<w id='w11' sameAs='Xw11' type='PUNCT_FSTOP'>.</w>
</p>                                            ...
```

```
<linkGrp type='MWLU' tagOrder='y'>
  <link ID='mwlnk1 targets='Xw1 Xw2'/>
                     ...
```

```
<p id='linkGrp'>
<linkGrp type='mwlnk-lem'tagOrder='y'>
  <link targets='mwlnk1 ADV'/>
<linkGrp type='w-lem'tagOrder='y'>
  <link targets='Xw3 COM-NOUN-1'/>
  <link targets='Xw4 COM-NOUN-2'/>
  <link targets='Xw5 PROP-N-LOC-1'/>
  <link targets='Xw6 PROP-N-LOC-2'/>
                     ...
```

```
<text id='LemDoc0001'>
    ...<fs id="COM-NOUN-2" type="Lemmatisation">
        <f name="Form"><str>aldetik</str></f>
        <f name="Lemma"><str>alde</str></f>
        <f name="morphological-Features">
          <fs type="Top-Features-List">
            <f name="POS"><sym value="NOUN"/></f>
            <f name="SUBCAT"><sym value="COM"/></f>
            <f name="DET"><sym value="DET"/></f>
            <f name="NUM"><sym value="S"/></f>
            <f name="CASE"><sym value="ABL"/></f>
          </fs>
        </f>
      </fs>
     </p>
     <p>
       <fs id="PROP-N-LOC-1" type="Lemmatisation ">
         <f name="Form"><str>Donostiako</str></f>
         <f name="Lemma"><str>Donostia</str></f>
         <f name="Top-Features-List">
           <fs type="upper-level-features">
            <f name="POS"><sym value="NAME"/></f>
            <f name="SUBCAT"><sym value="PROP-LOC"/></f>
            <f name="DET"><sym value="DET"/></f>
            <f name="NUM"><sym value="S"/></f>
            <f name="CASE"><sym value="GEL"/></f>
          </fs>
         </f>
       </fs>
     </p>
     <p>
                       ...
```

```
<text id=cad0001>
   ...
<fs id=COM-NOUN-1 type="phrase">
   <f name="chain">
       <str>itsaso aldetik</str></f>
   <f name="head">
       <str>itsaso</str></f>
   <f name="POS"><str>NOM</str></f>
   <f name="SUBCAT"><str>COM</str></f>
   <f name="SFL" ORG='list'>
      <sym value = "NCMOD" >
   </f>
</fs> ...
```

```
<text id=span0001>...
 <p id="linkgrp">
    <linkgrp type="span" tagOrder='y'>
      <link id="span1" targets="Xw3 Xw4">
      <link id="span2" targets="Xw5 Xw6 Xw7">...
    </linkgrp>
 </p>
</text>
```

```
<text id=span-sint0001>
   ...
 <p id="linkgrp">
    <linkgrp type="span-sint" tagOrder='y'>
      <link targets="span1 NOUN-COM">
    </linkgrp>
 </p>
</text>
```

```
<text id=dep0001>
  ...
<linkgrp type="dep" targorder="y"
   targFunc="head dependant
            case-dep-unit" domains=???>
   <!-- ... -->
   <link id="dep1" targets="Xw9 span1 span1">
   <link id="dep2" targets="Xw9 Xw7 Xw7">
...
</linkgrp>
```

```
...
<linkgrp type="dep-deplib" targorder="yes">
    <link targets="dep1 D-NCMOD-ABL"/>
    <link targets="dep2 D-NCMOD-ALA"/>
                        ...
</linkgrp>...
```

```
<fs id="D-NCMOD-ALA" type="dependency">
                   ...
   <f name="nombre">
        <sym value="NCMOD-ABL"/></f>
   <f name="CASE">
        <sym value="ALA"/></f> ..
</fs>
```

Figure3: Output of the different tools coded in XML.