Structure, Annotation and Tools in the Basque ZT Corpus

N. Areta (1), A. Gurrutxaga (1), I. Leturia (1), Z. Polin (1), R. Saiz (1), I. Alegria (2), X. Artola (2), A. Diaz de Ilarraza (2), N. Ezeiza (2), A. Sologaistoa (2), A. Soroa (2), A. Valverde (2)

(1) Elhuyar R&D
Zelai Haundi kalea, 3. Osinalde Industrialdea, 20170 Usurbil. Basque Country
agurrutxaga@elhuyar.com(1)
(2) Ixa Taldea. University of the Basque Country
649 Postakutxa. 20080 Donostia. Basque Country
i.alegria@ehu.es

Abstract

The ZT corpus (Basque Corpus of Science and Technology) is a tagged collection of specialized texts in Basque, which wants to be a main resource in research and development about written technical Basque: terminology, syntax and style. It will be the first written corpus in Basque which will be distributed by ELDA (at the end of 2006) and it wants to be a methodological and functional reference for new projects in the future (i.e. a national corpus for Basque). We also present the technology and the tools to build this Corpus. These tools, Corpusgile and Eulia, provide a flexible and extensible infrastructure for creating, visualizing and managing corpora and for consulting, visualizing and modifying annotations generated by linguistic tools.

1. Introduction

In the last years, corpora have become an essential tool in any domain of linguistics. Strictly speaking, any collection of texts can be called a corpus, but normally other conditions are required for a bunch of texts to be considered a corpus: it must be a 'big' collection of 'real' language samples, collected following some 'criteria' and 'linguistically' tagged (Bach *et al.* 1997:4).

Although Basque language has not a very long tradition regarding Science and Technology (it must be taken into account that its standardization and normalization only began in 1968, that it was not taught at schools until the 70s and used in Universities till the 80s), nowadays there are quite a lot of texts in Basque on Science and Technology, some dating back to 30 years ago. Even so, it is one of the areas with least 'de jure' normalization, and therefore the need of a Basque Science and Technology Corpus.

Corpora in Basque have so far been 'general'. There are no sources to study the Science and Technology branch of language. That is why we started the project of a 'specialized' (Sinclair 1996: 10) corpus, called *Zientzia eta Teknologiaren Corpusa* (henceforth *ZT Corpus*). It is a tagged collection of specialized texts in Basque, which wants to be a main resource in research and development about written technical Basque terminology, syntax and style. It will be the first written corpus in Basque which will be distributed by ELDA (at the end of 2006) and it wants to be a methodological and functional reference for new projects in the future (i.e. a national corpus for Basque).

The process of building the ZT Corpus has been done following a certain methodology. The guidelines followed involved the four steps of building the corpus: corpus design, raw corpus collecting, corpus tagging and corpus analysis and browsing. To help the process of building the corpus, some tools have been developed, which can be reused in the future to build new corpora.

2. Design of the Corpus

2.1. Features of the Corpus

The corpus intends to cover the texts about Science and Technology written in Basque in the years from 1990 to 2002 inclusive.

The corpus is divided in two main parts:

- a balanced corpus, tagged automatically and revised by hand
- an unbalanced corpus, as big as possible, tagged automatically

The aim is to collect 5 million words in the balanced section (currently more than 1.5 million words have been tagged) and more than 20 million words in the open section (at the moment more than 8 million words have been stored).

In order to balance the corpus, an inventory of all the articles and books about science and technology written in Basque between 1990 and 2002 was compiled as a previous step. The references were classified by topic and genre, and these factors were considered in the random selection of the samples (stratified sampling).

The topics we chose were exact sciences, matter and energy sciences, earth sciences, life sciences, technology, general and others. As to the genres, we chose schoolbooks and textbooks, high-level books (specialists' texts and University textbooks), popular science books, specialized articles, popular science articles and civil service books.

The total number of words in the inventoried texts is estimated in more than 85 millions words. In order to make a 5 million word corpus, we had to take a sample of the inventoried texts, in a 5/85 proportion (almost 6%). As the sampling was stratified, this proportion was to be taken in each of the topic/genre combinations.

The sampling of 6% can be done taking 6% of each and every item (book or article), which would be most representative but very costly (obtaining the books or articles has indeed proved to be the most difficult part of building the corpus!), or taking only a 6% of the items and them in full extent, which would be easier but not as representative as we would wish. Besides, this last solution could pose some problems regarding copyrights. So we took neither of these two ends, but a solution halfway of both: we took $\sqrt{5/85}$ of the items at random, and $\sqrt{5/85}$ of the words from each of them.

The sample that is taken from each of the items is not continuous. In order to get as much linguistic variety as possible, we were interested in taking different bits of the documents. So the sample to be taken is divided in 300 word chunks, spread out equally at random through the document.

The general scheme of the annotation process is shown in Fig. 1.

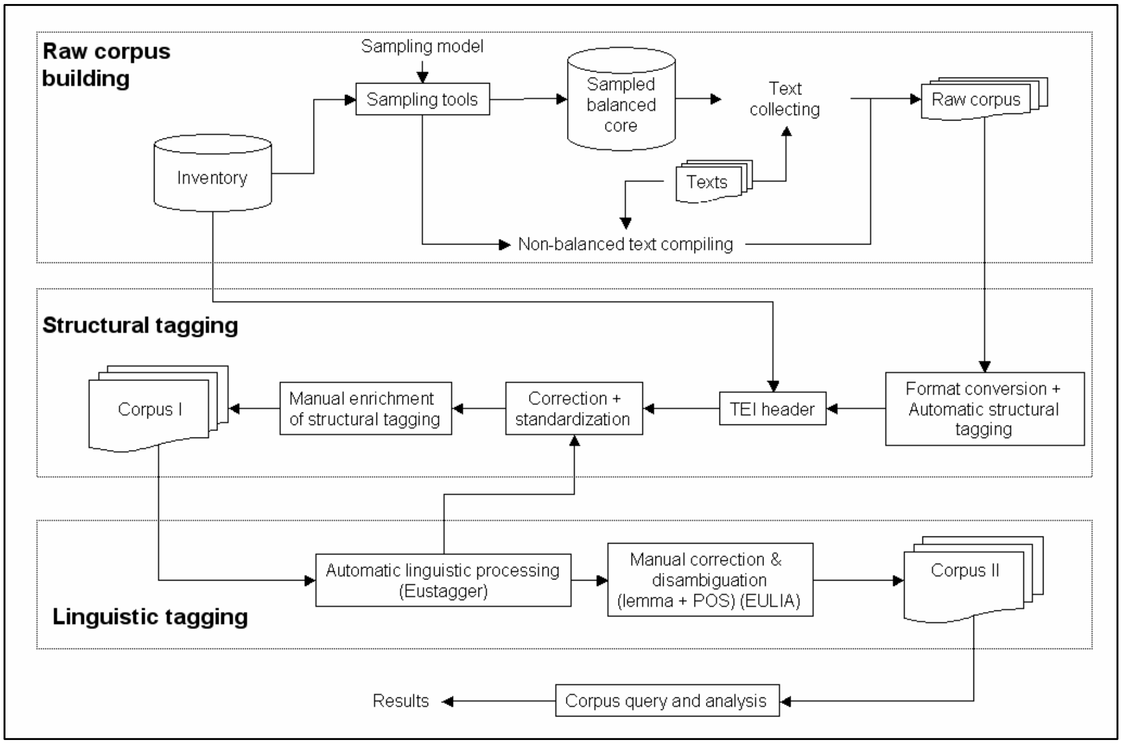

Raw corpus building
Sampling model
Sampling tools
Sampled balanced core
Text collecting
Raw corpus
Inventory
Texts
Non-balanced text compiling
Structural tagging
Corpus I
Manual enrichment of structural tagging
Correction + standardization
TEI header
Format conversion + Automatic structural tagging
Automatic linguistic processing (Eustagger)
Manual correction & disambiguation (lemma + POS) (EULIA)
Corpus II
Linguistic tagging
Results
Corpus query and analysis

Fig 1.- General scheme

2.2. Raw Corpus

For obtaining the raw corpus, we got in contact with Basque publishers. We told them about the corpus and signed an agreement with each of them. So the publishers sent us the texts selected for the balanced part and, if they wanted to, the ones that did not get chosen too, preferably in electronic format. The texts for the balanced part of the corpus that could not be obtained in electronic format were scanned, OCRed and reviewed. For the unbalanced part, only texts in electronic format were accepted.

For the annotation of the ZT Corpus, we chose TEI P4 (Ide et al., 2004) (TEI, 2005). To convert the documents from their original formats to TEI, we developed a HTML-TEI converter and a Doc-TEI converter. Conversion from other formats (Quark, PDF...) is done via external programs that convert from these to HTML first.

When we say balanced corpus and unbalanced corpus, we are not talking about two different corpora. There is only one collection of documents, and the paragraphs that are sampled for the balanced part are marked with an *'orekatua'* (for *balanced*) attribute.

2.3. Structural Annotation

The structural annotation is done in two steps: a first automatic one, which is done to all documents during the conversion, and a second manual deeper one, which is done only to the documents in the balanced part.

The automatic structural mark-up includes information about the document, information about text structure and typography. The information about the document is put under the `<teiHeader>` section. Text structure (titles, sections, subsections, paragraphs, lists, tables, footnotes...) is marked using the following tags: `<body>`, `<div>`, `<head>`, `<p>`, `<table>`, `<row>`, `<cell>`, `<list>`, `<item>` and `<note>`. Typography is marked using the tag `<hi>` combined with the attribute 'rend'.

In the balanced part deeper structural information is annotated. The typographical information is converted manually to more detailed tags: `<foreign>`, `<emph>`, `<distinct>`, `<q>`, `<soCalled>`, `<term>`, `<gloss>`, `<mentioned>`, `<name>`, `<head>` and `<note>`. The *lang* attribute is used for chunks in other languages.

Additionally, to ease the subsequent linguistic annotation process, NLP tools are used to detect chunks in other languages, typographical errors and non-standard uses, which are then manually reviewed for correctness and annotated using the `<foreign>`, `<corr>` and `<reg>`

tags. Statistics of these manual revisions are kept and afterwards used to improve the aforementioned NLP tools.

2.4. Linguistic Annotation

The linguistic annotation is based on TEI-P4 conformant typed features which are managed using EULIA (Artola et al., 2004), a web interface for creating, browsing and editing these structures. The annotation scheme is stand-off, so the information for each document will be divided in several files and it can be seem as a composition of XML documents (*annotation web*).

The steps which are carried out are the following (Fig. 2):

- a tokenizer that identifies tokens and sentences
- a morphological segmentizer which splits up a word into its constituents morphemes
- a morphosyntactic analyzer whose goal is to group the morphosyntactic information associated to a word.
- the treatment of multiword lexical units (MWLU) as dates, numbers, named entities, ...
- disambiguation and lemmatization: Based on the previous steps a combined tagger obtains an unique analysis for each lexical unit; so, lemma, part of speech and other morphosyntactic features are assigned.

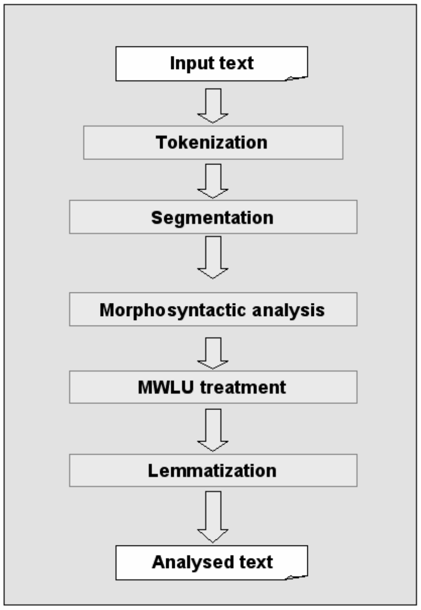


Fig 2.- Steps in linguistic tagging

This automatic process includes some errors. In the balanced corpus the results corresponding to lemma and part of speech are examined by linguists using EULIA.

Additional information about syntactic functions and semantic role will be included in a small part of the corpus. This information is obtained using NLP tools and revised and corrected in the balanced part.

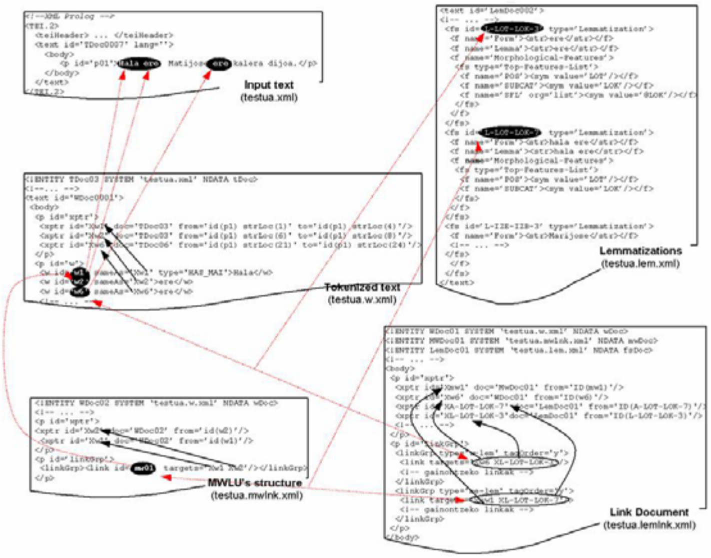
The use of a stand-off linguistic annotation is very interesting because:

- partial results and ambiguities can be easily represented
- information can be organized at different levels
- the representation of MWLUs is clear
- the level of disambiguation (automatic/manual) can be expressed
- there are not different mechanisms to indicate the same type of information

In this architecture three elements are distinguished in different documents:

- text anchors: text elements found in the input
- linguistic information: feature structures obtained from the analyses
- links between anchors and their corresponding analyses

In Fig. 3 we can see in a graphical mode the links between documents.



```
<!--XML Prolog -->
<TEI.2>
  <teiHeader> ... </teiHeader>
  <text id='TDoc0007' lang=''>
    <body>
      <p id='p01'>Hala ere Matijose ere kalera dijoa.</p>
    </body>
  </text>
</TEI.2>
```

```
<text id='LemDoc002'>
<!-- ... -->
 <fs id='L-LOT-LOK-3' type='Lemmatization'>
  <f name='Form'><str>ere</str></f>
  <f name='Lemma'><str>ere</str></f>
  <f name='Morphological-Features'>
   <fs type='Top-Features-List'>
    <f name='POS'><sym value='LOT'/></f>
    <f name='SUBCAT'><sym value='LOK'/></f>
    <f name='SFL' org='list'><sym value='@LOK'/></f>
   </fs>
  </f>
 </fs>
 <fs id='L-LOT-LOK-7' type='Lemmatization'>
  <f name='Form'><str>hala ere</str></f>
  <f name='Lemma'><str>hala ere</str></f>
  <f name='Morphological-Features'>
   <fs type='Top-Features-List'>
    <f name='POS'><sym value='LOT'/></f>
    <f name='SUBCAT'><sym value='LOK'/></f>
   </fs>
  </f>
 </fs>
 <fs id='L-IZE-IZB-3' type='Lemmatization'>
  <f name='Form'><str>Marijose</str></f>
  <!-- ... -->
 </fs>
  </f>
 </fs>
</text>
```

```
<!ENTITY TDoc03 SYSTEM 'testua.xml' NDATA tDoc>
<!--... -->
<text id="WDoc0001'>
 <body>
  <p id="xptr'>
   <xptr id='Xw1' doc='TDoc03' from='id(p1) strLoc(1)' to='id(p1) strLoc(4)'/>
   <xptr id='Xw2' doc='TDoc03' from='id(p1) strLoc(6)' to='id(p1) strLoc(8)'/>
   <xptr id='Xw6' doc='TDoc06' from='id(p1) strLoc(21)' to='id(p1) strLoc(24)'/>
  </p>
  <p id="w'>
   <w id='w1' sameAs='Xw1' type='HAS_MAI'>Hala</w>
   <w id='w2' sameAs='Xw2'>ere</w>
   <w id='w6' sameAs='Xw6'>ere</w>
   <!-- ... -->
```

```
<!ENTITY WDoc01 SYSTEM 'testua.w.xml' NDATA wDoc>
<!ENTITY MWDoc01 SYSTEM 'testua.mwlnk.xml' NDATA mwDoc>
<!ENTITY LemDoc01 SYSTEM 'testua.lem.xml' NDATA fsDoc>
<!-- ... -->
<body>
 <p id='mptr'>
  <xptr id='Xmw1' doc='MwDoc01' from='ID(mw1)'/>
  <xptr id='Xw6' doc='WDoc01' from='ID(w6)'/>
  <xptr id='XA-LOT-LOK-7' doc='LemDoc01' from='ID(A-LOT-LOK-7)'/>
  <xptr id='XL-LOT-LOK-3'doc='LemDoc01' from='ID(L-LOT-LOK-3)'/>
  <!-- ... -->
 </p>
 <p id='linkGrp'>
  <linkGrp type='w-lem' tagOrder='y'>
   <link targets='Xw6 XL-LOT-LOK-3'/>
   <!-- gainontzeko linkak -->
  </linkGrp>
  <linkGrp type='mw-lem' tagOrder='y'>
   <link targets='Xmw1 XL-LOT-LOK-7'/>
   <!-- gainontzeko linkak -->
  </linkGrp>
 </p>
</body>
```

```
<!ENTITY WDoc02 SYSTEM 'testua.w.xml' NDATA wDoc>
<!-- ... -->
<p id='mptr'>
<xptr id='Xw2' doc='WDoc02' from='id(w2)'/>
 <xptr id='Xw1' doc='WDoc02' from='id(w1)'/>
</p>
<p id='linkGrp'>
 <linkGrp><link id='mw01' targets='Xw1 Xw2'/></linkGrp>
</p>
```

Fig 3.- Stand-off representation: anchors, linguistic information and links

3. The Tools

An application named Corpusgile has been developed for this corpus and for developing new corpora in the future. Some previous NLP tools for Basque have been reused. There are 4 main modules in the application:

- The corpus builder
- The structural tagger
- The linguistic tagger
- The browser

The first 3 modules are being currently used and the browser is in design phase, with the aim of being finished in April. So by the time the LREC Conference will take place the browser will be finished and available online.

3.1. The Corpus Builder

It is based on a relational database and it includes all the main functions: inventorying, classification, stratified sampling of documents (random selection of documents for the balanced part), storage, format-conversion, sampling inside documents and search, all of them with a user-friendly interface.

In Fig. 4 we can see the main interface for the Corpus Builder.

3.2. The Structural Tagger

The following steps are controlled and carried out by this tool:

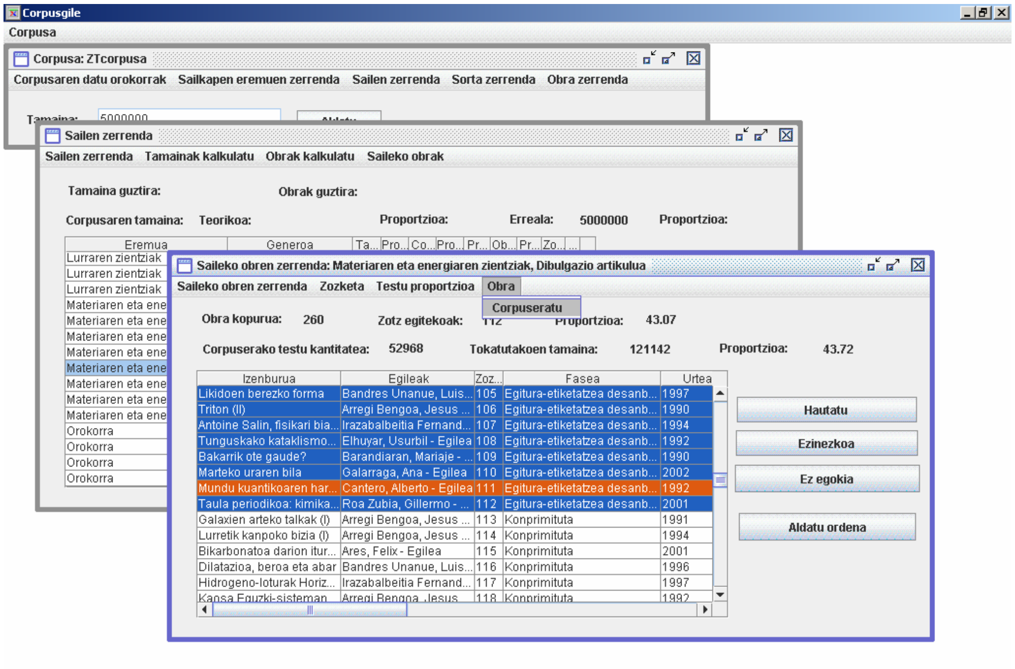
- tagging and parsing the TEI-XML format at structural level
- adding specialized or technical words to the corpus-specific lexicon in order to improve the future linguistic tagging; to achieve this, an NLP tool called EusTagger, a lemmatizing/disambiguating tool, based on the former Euslem (Aduriz *et al*., 1996) is used to detect non-correct words, which are then ordered by frequency of the lemmas proposed by Eustagger and presented to the user for acceptance and assignment of lemma and POS
- NLP process for recognition of misspellings, non-standard uses and presence of chunks in other languages, marked via `<corr>`, `<reg>` and `<foreign>` tags
- manual revision of `<corr>`, `<reg>` and `<foreign>` tags in the balanced part
- interface for scanning typographical changes, highlighting and quotation (mainly `<hi>` tags) and assigning them a sense (`<emph>`, `<distinct>`, `<q>`, `<soCalled>`, `<term>`, `<gloss>`, `<mentioned>`, `<name>`...) when appropriate
- interface for correcting, improving and disambiguating the structural tagging of the balanced part
- verification of XML structures

In Fig. 5 we can see the interface when a non-standard use is tagged linked to the standard one.

3.3. The Linguistic Tagger

It is carried out using EULIA (Artola et al., 2005) a framework for creating, browsing and editing linguistic annotations. It is based on a class library named LibiXaML and the the huge amount of generated information is stored in a XML database.

It is an extensible, user-oriented and component-based software-architecture. At the moment several NLP processors for Basque are integrated: tokenization, morphological segmentation, multiword recognition, lemmatization/disambiguation, shallow syntax and dependency-based analysis.



Tamaina: 5000000

Tamaina guztira: Obrak guztira:

Corpusaren tamaina: Teorikoa: Proportzioa: Erreala: 5000000 Proportzioa:

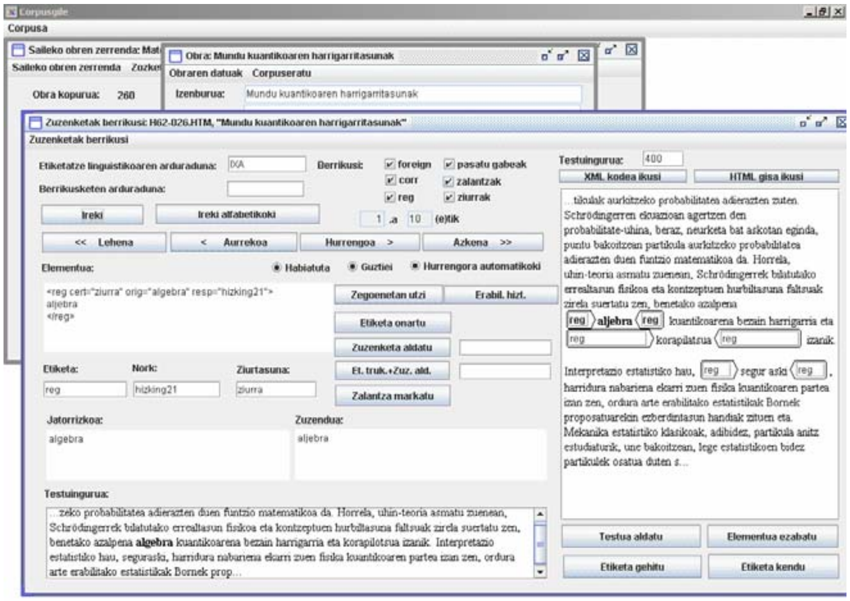
Eremua	Generoa	Ta...	Pro...	Co...	Pro...	Pr...	Ob...	Pr...	Zo...	...
Lurraren zientziak										
Lurraren zientziak										
Lurraren zientziak										
Materiaren eta ene										
Materiaren eta ene										
Materiaren eta ene										
Materiaren eta ene										
Materiaren eta ene										
Materiaren eta ene										
Materiaren eta ene										
Materiaren eta ene										
Orokorra										
Orokorra										
Orokorra										
Orokorra										

Obra kopurua: 260 Zotz egitekoak: 112 Proportzioa: 43.07

Corpuserako testu kantitatea: 52968 Tokatutakoen tamaina: 121142 Proportzioa: 43.72

Izenburua	Egileak	Zoz...	Fasea	Urtea
Likidoen berezko forma	Bandres Unanue, Luis...	105	Egitura-etiketatzea desanb...	1997
Triton (II)	Arregi Bengoa, Jesus ...	106	Egitura-etiketatzea desanb...	1990
Antoine Salin, fisikari bia...	Irazabalbeitia Fernand...	107	Egitura-etiketatzea desanb...	1994
Tunguskako kataklismo...	Elhuyar, Usurbil - Egilea	108	Egitura-etiketatzea desanb...	1992
Bakarrik ote gaude?	Barandiaran, Mariaje - ...	109	Egitura-etiketatzea desanb...	1990
Marteko uraren bila	Galarraga, Ana - Egilea	110	Egitura-etiketatzea desanb...	2002
Mundu kuantikoaren har...	Cantero, Alberto - Egilea	111	Egitura-etiketatzea desanb...	1992
Taula periodikoa: kimika...	Roa Zubia, Gillermo - ...	112	Egitura-etiketatzea desanb...	2001
Galaxien arteko talkak (I)	Arregi Bengoa, Jesus ...	113	Konprimituta	1991
Lurretik kanpoko bizia (I)	Arregi Bengoa, Jesus ...	114	Konprimituta	1994
Bikarbonatoa darion itur...	Ares, Felix - Egilea	115	Konprimituta	2001
Dilatazioa, beroa eta abar	Bandres Unanue, Luis...	116	Konprimituta	1996
Hidrogeno-loturak Horiz...	Irazabalbeitia Fernand...	117	Konprimituta	1997
Kapsa Eguzki-sisteman	Arregi Bengoa, Jesus	118	Konprimituta	1992

Fig 4.- Main interface for the Corpus Builder


Corpusgile
Corpusa
Saileko obren zerrenda: Mat
Saileko obren zerrenda Zozke
Obra kopurua: 260
Obra: Mundu kuantikoaren harrigarritasunak
Obraren datuak Corpuseratu
Izenburua: Mundu kuantikoaren harrigarritasunak
Zuzenketak berrikusi: H62-026.HTM, "Mundu kuantikoaren harrigarritasunak"
Zuzenketak berrikusi
Etiketatze linguistikoaren arduraduna: IXA
Berrikusketen arduraduna:
Berrikusi: ☑ foreign ☑ pasatu gabeak ☑ corr ☑ zalantzak ☑ reg ☑ ziurrak
Ireki
Ireki alfabetikoki
1 .a 10 (e)tik
<< Lehena
< Aurrekoa
Hurrengoa >
Azkena >>
Elementua: ● Habiatuta ● Guztiei ● Hurrengora automatikoki
Zegoenetan utzi
Erabil. hizt.
Etiketa onartu
Zuzenketa aldatu
Et. truk.+Zuz. ald.
Zalantza markatu
Etiketa: reg
Nork: hizking21
Ziurtasuna: ziurra
Jatorrizkoa: algebra
Zuzendua: aljebra
Testuingurua:
Testuingurua: 400
XML kodea ikusi
HTML gisa ikusi
Testua aldatu
Elementua ezabatu
Etiketa gehitu
Etiketa kendu

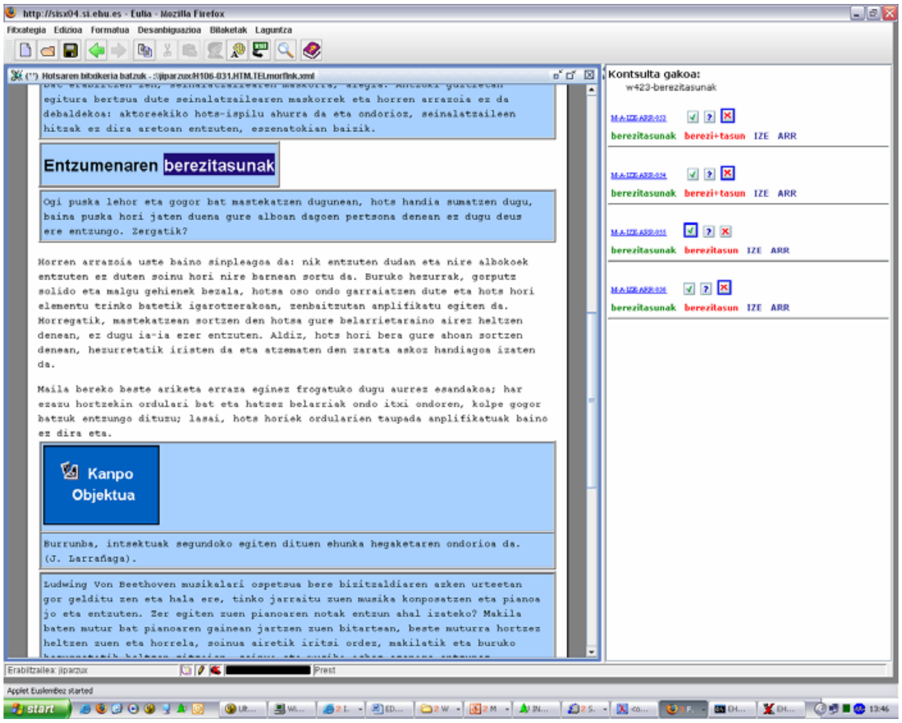
```
<reg cert="ziurra" orig="algebra" resp="hizking21">
aljebra
</reg>
```

…zeko probabilitatea adierazten duen funtzio matematikoa da. Horrela, uhin-teoria asmatu zuenean, Schrödingerrek bilatutako errealtasun fisikoa eta kontzeptuen hurbiltasuna faltsuak zirela suertatu zen, benetako azalpena **algebra** kuantikoarena bezain harrigarria eta korapilotsua izanik. Interpretazio estatistiko hau, seguraski, harridura nabariena ekarri zuen fisika kuantikoaren partea izan zen, ordura arte erabilitako estatistikak Bornek prop…

…tikulak aurkitzeko probabilitatea adierazten zuten. Schrödingerren ekuazioan agertzen den probabilitate-uhina, beraz, neurketa bat askotan eginda, puntu bakoitzean partikula aurkitzeko probabilitatea adierazten duen funtzio matematikoa da. Horrela, uhin-teoria asmatu zuenean, Schrödingerrek bilatutako errealtasun fisikoa eta kontzeptuen hurbiltasuna faltsuak zirela suertatu zen, benetako azalpena reg **aljebra** reg kuantikoarena bezain harrigarria eta reg korapilatsua reg izanik.

Interpretazio estatistiko hau, reg segur asko reg, harridura nabariena ekarri zuen fisika kuantikoaren partea izan zen, ordura arte erabilitako estatistikak Bornek proposatuarekin ezberdintasun handiak zituen eta. Mekanika estatistiko klasikoak, adibidez, partikula anitz estudiaturik, une bakoitzean, lege estatistikoen bidez partikulek osatua duten s…

Fig 5.- Interface for manual revision of <reg>


http://sisx04.si.ehu.es - Eulia - Mozilla Firefox
Fitxategia Edizioa Formatua Desanbiguazioa Bilaketak Laguntza
(*) Hotsaren bitxikeria batzuk - :/jiparzux/H106-031.HTM.TELmorfink.xml
Kanpo Objektua
Erabiltzailea: jiparzux
Prest
Applet EuslemBez started
start
Kontsulta gakoa:
w423-berezitasunak
M-A-IZE-ARR-053
M-A-IZE-ARR-054
M-A-IZE-ARR-055
M-A-IZE-ARR-056

egitura bertsua dute seinalatzailearen maskorrek eta horren arrazoia ez da debaldekoa: aktoreekiko hots-ispilu ahurra da eta ondorioz, seinalatzaileen hitzak ez dira aretoan entzuten, eszenatokian baizik.

Entzumenaren berezitasunak

Ogi puska lehor eta gogor bat mastekatzen dugunean, hots handia sumatzen dugu, baina puska hori jaten duena gure alboan dagoen pertsona denean ez dugu deus ere entzungo. Zergatik?

Horren arrazoia uste baino sinpleagoa da: nik entzuten dudan eta nire albokoek entzuten ez duten soinu hori nire barnean sortu da. Buruko hezurrak, gorputz solido eta malgu gehienek bezala, hotsa oso ondo garraiatzen dute eta hots hori elementu trinko batetik igarotzerakoan, zenbaitzutan anplifikatu egiten da. Horregatik, mastekatzean sortzen den hotsa gure belarrietaraino airez heltzen denean, ez dugu ia-ia ezer entzuten. Aldiz, hots hori bera gure ahoan sortzen denean, hezurretatik iristen da eta atzematen den zarata askoz handiagoa izaten da.

Maila bereko beste ariketa erraza eginez frogatuko dugu aurrez esandakoa; har ezazu hortzekin ordulari bat eta hatzez belarriak ondo itxi ondoren, kolpe gogor batzuk entzungo dituzu; lasai, hots horiek ordularien taupada anplifikatuak baino ez dira eta.

Burrunba, intsektuak segundoko egiten dituen ehunka hegaketaren ondorioa da. (J. Larrañaga).

Ludwing Von Beethoven musikalari ospetsua bere bizitzaldiaren azken urteetan gor gelditu zen eta hala ere, tinko jarraitu zuen musika konposatzen eta pianoa jo eta entzuten. Zer egiten zuen pianoaren notak entzun ahal izateko? Makila baten mutur bat pianoaren gainean jartzen zuen bitartean, beste muturra hortzez heltzen zuen eta horrela, soinua airetik iritsi ordez, makilatik eta buruko

M-A-IZE-ARR-053 ✓ ? ☒	berezitasunak	berezi+tasun	IZE	ARR
M-A-IZE-ARR-054 ✓ ? ☒	berezitasunak	berezi+tasun	IZE	ARR
M-A-IZE-ARR-055 ☑ ? ✗	berezitasunak	berezitasun	IZE	ARR
M-A-IZE-ARR-056 ✓ ? ☒	berezitasunak	berezitasun	IZE	ARR

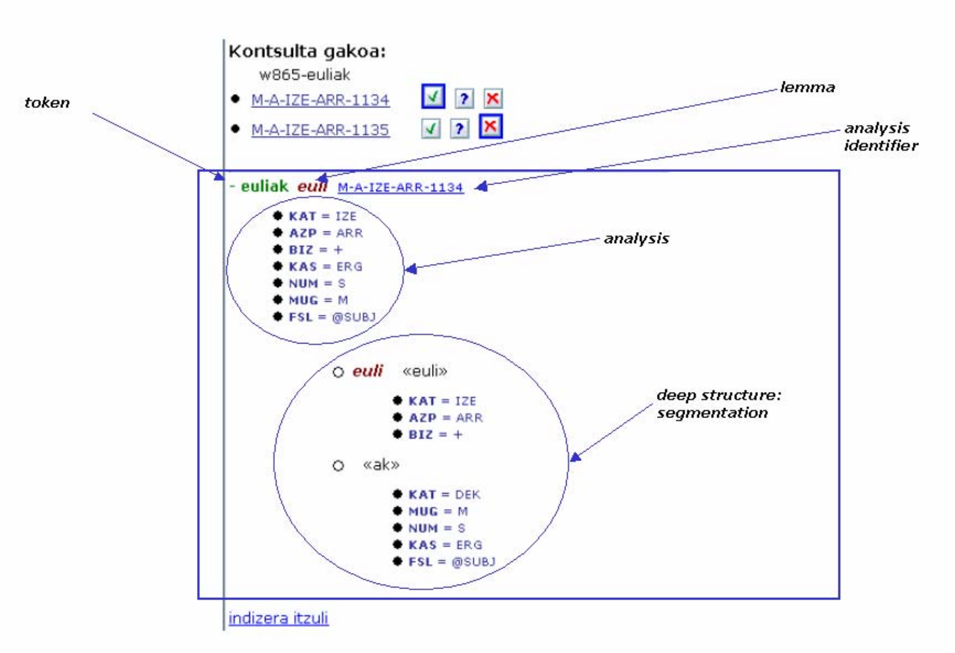
Fig 6.- Main interface in EULIA

After the automatic processing which generate the XML documents a module for manual linguistic annotation can be used. This module integrates the results of the automatic processes and gives to the linguistics a friendly interface for the annotation, hiding the complexity of the multiple files are being managed. The main interface is shown in Fig. 6.

As it can be observed there are two main windows: the *text window* on left and the *analysis window* on right.

In the text window the linguist can click a token and a set of actions are offered to be performed.

The main action is to show in the analysis window the different possible analyses in order to disambiguate them. Anyway different icons and display methods are used to indicate different features: hand-made disambiguation, multiple analyses and so on. In the analysis window details about the analyses are shown using style-sheets which hide the different files and tags.



```
Kontsulta gakoa:
    w865-euliak
● M-A-IZE-ARR-1134  ✓ ? ✗
● M-A-IZE-ARR-1135  ✓ ? ✗

- euliak euli M-A-IZE-ARR-1134
    ● KAT = IZE
    ● AZP = ARR
    ● BIZ = +
    ● KAS = ERG
    ● NUM = S
    ● MUG = M
    ● FSL = @SUBJ

    ○ euli  «euli»
        ● KAT = IZE
        ● AZP = ARR
        ● BIZ = +
    ○ «ak»
        ● KAT = DEK
        ● MUG = M
        ● NUM = S
        ● KAS = ERG
        ● FSL = @SUBJ

indizera itzuli
```

Fig 7.- Analysis window

In Fig. 7 we can see the top of this window for an example. Information for whole word *euliak* (flies) and for the two morphemes, *euli* (fly) and *ak* (nominative plural) are given.

4. Conclusions

Just as any other language, Basque needs corpora. Linguists, terminology specialists, language technology researchers, people that work in language standardization and normalization… Many people need corpora, nowadays an essential tool for the analysis of language. The Basque ZT corpus wants to be a useful and powerful tool for researching on specialized texts in Basque.

But Basque being a small language in terms of speakers and, therefore, resources dedicated to it, we not only need corpora. We also need the technology to build them easily; we need tools that will assist in the process of creating and managing corpora and that will reduce the usually expensive costs of building them. We have made such a tool, Corpusgile. This tool provides a flexible and extensible infrastructure for creating, visualizing and managing corpora and for consulting, visualizing and modifying annotations generated by linguistic tools. The interface has been designed to be informative, easy-to-use and intuitive. And due to its being based on the TEI standards, XML and stand-off annotation; it can be adapted by other builders of corpora using other tag sets, tools and languages.

Besides, in the making of Corpusgile we have defined and applied a methodology for building corpora more easily in the future.

These three things, a resource (the ZT Corpus), a methodology and a tool (Corpusgile) are the contributions we have done to this field we are so in need of, the field of corpora. And we are convinced that in the future they will prove to be very valuable contributions indeed.

5. Acknowledgements

This work was partially funded by the Basque Government (EJ-ETORTEK-2002/HIZKING21) and the University of the Basque Country (EHU/GIU05/52).

6. References

Aduriz I., Aldezabal I., Alegria I., Artola Zubillaga X., Ezeiza N., Urizar R. 1996. "EUSLEM: A Lemmatiser / Tagger for Basque." *Proc. EURALEX'96, Part 1, 17-26.* Göteborg.

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Sologaistoa A., Soroa A. 2004. "EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora." *LREC 2004. Workshop on Workshop on XML-based richly annotated corpora*. Lisbon.

Biber D., Conrad, S. eta Reppen, R. 2000. Corpus Linguistics. *Investigating Language Structure and Use.. Cambridge*: Cambridge University Press.

Bowkeron L, Pearson J. 2002. Working with Specialized Language. *A practical guide to using corpora*. New York: Routledge

Ide N., Romary L., Clergerie E. "International standard for a linguistic annotation framework." *Natural Language Engineering*, 2004

TEI. Text Encoding Initiative. The XML version of the TEI Guidelines. [on line] [date: 05-01-22] <http://www.tei-c.org/P4X/>