ZT Corpus
Annotation and tools for Basque corpora

Areta N., Gurrutxaga A., Leturia I.
R&D
Elhuyar Fundazioa
agurrutxaga@elhuyar.com

Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Sologaistoa A.
IXA Taldea
University of the Basque Country
i.alegria@ehu.es

Abstract

The ZT Corpus (Basque Corpus of Science and Technology) is a tagged collection of specialised texts in Basque, which aims to be a major resource in research and development with respect to written technical Basque: terminology, syntax and style. It was released in December 2006 and can be queried at http://www.ztcorpusa.net.

The ZT Corpus stands out among other Basque corpora for many reasons: it is the first specialised corpus in Basque, it has been designed to be a methodological and functional reference for new projects in the future (i.e. a national corpus for Basque), it is the first corpus in Basque annotated using a TEI-P4 compliant XML format, it is the first written corpus in Basque to be distributed by ELDA and it has a friendly and sophisticated query interface. The corpus has two kinds of annotation, a structural annotation and a stand-off linguistic annotation. It is composed of two parts, a 1.6 million-word balanced part, whose annotation has been revised by hand, and another automatically tagged 6 million-word part. The project is not closed, and we have the intention to gradually enlarge the corpus, along with making improvements to it.

We also present the technology and the tools used to build this corpus. These tools, *Corpusgile* and *Eulia*, provide a flexible and extensible infrastructure for creating, visualising and managing corpora, and for consulting, visualising and modifying annotations generated by linguistic tools. And finally we will be introducing the web interface to query the ZT Corpus, which offers some interesting advanced features that are new in Basque corpora.

1. Introduction

In the last few 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 in accordance with certain 'criteria' and 'linguistically' tagged (Bach *et al.*, 1997: 4).

Although the Basque language does not have a very long tradition as far as science and technology is concerned (one needs to bear in mind that its standardization and normalisation only began in 1968, and it was not taught in schools until the 70s or used in Universities until the 80s), nowadays there are quite a lot of texts in Basque on science and technology, some dating back to thirty years ago. Even so, it is one of the areas with least *de jure* normalisation, and therefore the need for a Basque Science and Technology Corpus is clear.

Corpora in Basque have so far been 'general'. Previously there were another two annotated corpora for Basque (*Corpus of Basque in the 20th century*: http://www.euskaracorpusa.net and *Reference Prose Nowadays*: http://www.ehu.es/euskara-orria/euskara/ereduzkoa), both made up of generic texts (literature, press…). There are no sources for studying the science and technology branch of the Basque language. That is why we started the project of a 'specialised' (Sinclair, 1996: 10) corpus, called *Zientzia eta Teknologiaren Corpusa* (henceforth *ZT Corpus*). It is a tagged collection of specialised texts in Basque, which aims to be a major resource in research and development with respect to written technical Basque lexicology, terminology, syntax and style. It is the first written corpus in Basque to be distributed by ELDA, and it aims to be a methodological and functional reference for new projects in the future (i.e. a national corpus for Basque).

The corpus has two parts:

- 1.6 million words have been revised manually, since they are included in the balanced part
- 6 million words have been automatically tagged

These are the figures of the corpus so far, but they are not definitive. We intend to gradually add new texts to it, as well as other improvements. For example, during 2007 another 1.2 million words will be included in the corpus, 300,000 of which will be manually revised.

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

The ZT Corpus has been structurally and linguistically annotated. The tagging process has been carried out in two steps:

- Structural annotation: includes information about the document, text structure and typography. The work of the annotator is assisted by a tool which detects misspellings, split words, linguistic variations and phrases in other languages
- Linguistic annotation: the annotation scheme is stand-off, so the information for each document is stored in several files and can be seen as a composition of XML documents (an *annotation web*). The tool *Eulia* helps linguists with the revision of the balanced part

2. Design of the corpus

2.1. Features of the corpus

The corpus sets out to cover the texts on science and technology written in Basque from 1990 to 2002, inclusive.

The corpus is divided into 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 five million words in the balanced section (currently more than 1.6 million words have been tagged) and more than twenty million words in the open section (at the moment more than six million words have been stored). We have released a first version of the corpus with the amount of words we have at the moment, but the project is not finished and our intention is to continue adding new texts to the corpus in order to reach the desired size for the corpus.

In order to balance the corpus, an inventory of all the articles and books on science and technology written in Basque between 1990 and 2002 was compiled as a preliminary 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
- Others

As far as the genres were concerned, we chose:

- Schoolbooks and textbooks
- High-level books (specialists' texts and University textbooks)
- Popular science books
- Specialised articles
- Popular science articles
- Civil service books.

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

The sampling of 6 percent can be done by taking 6 percent of each and every item (book or article), which would be the most representative but very costly way (obtaining the books or articles has indeed proved to be the most difficult part in building the corpus!), or taking only 6 percent 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 options, but went for a halfway solution of each one: 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 sections

of the documents. So the sample to be taken is divided into 300-word chunks, spread out equally at random throughout the document.

An outline of the annotation process is shown in Figure 1.

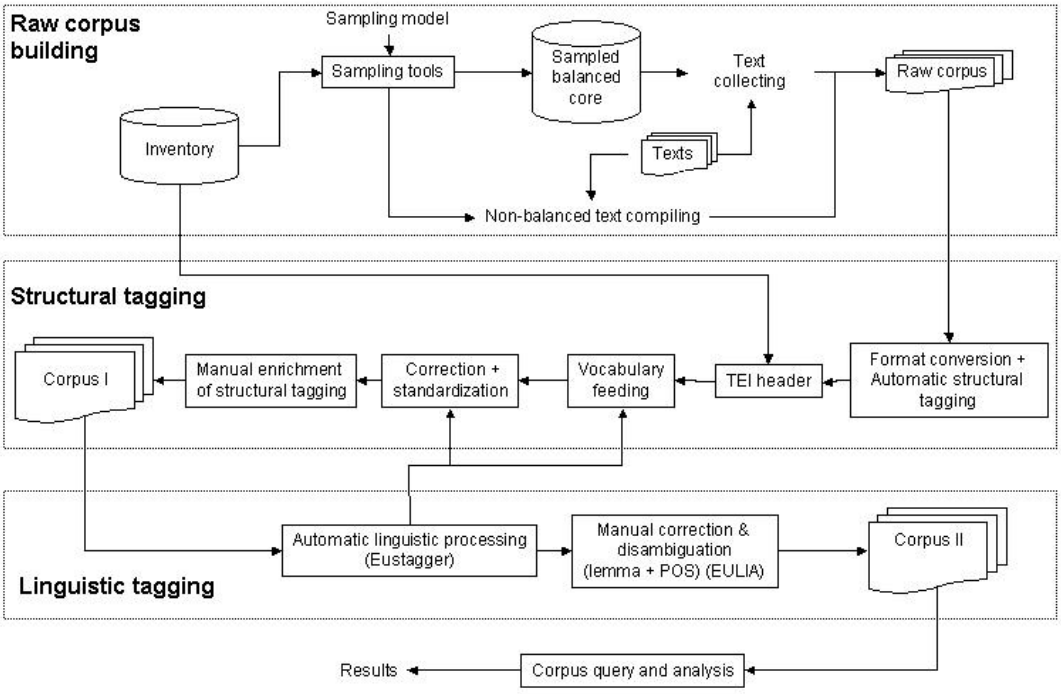

Figure 1.- General outline

2.2. Raw corpus

For obtaining the raw corpus, we got into 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 wished, they also sent the ones that did not get chosen, in electronic format wherever possible. 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, 2004). To convert the documents from their original formats to TEI, we developed an 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 applied to all documents during the conversion, and a second manual deeper one, which is applied 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>, , <head>, <p>, <table>, <row>, <cell>, <list>, <item>, <note>, <lb> and <seg>. 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 can be seen as a composition of XML documents (*annotation web*).

2.4.1. Linguistic annotation process

The steps which are carried out in the linguistic annotation process are the following (Figure 2):

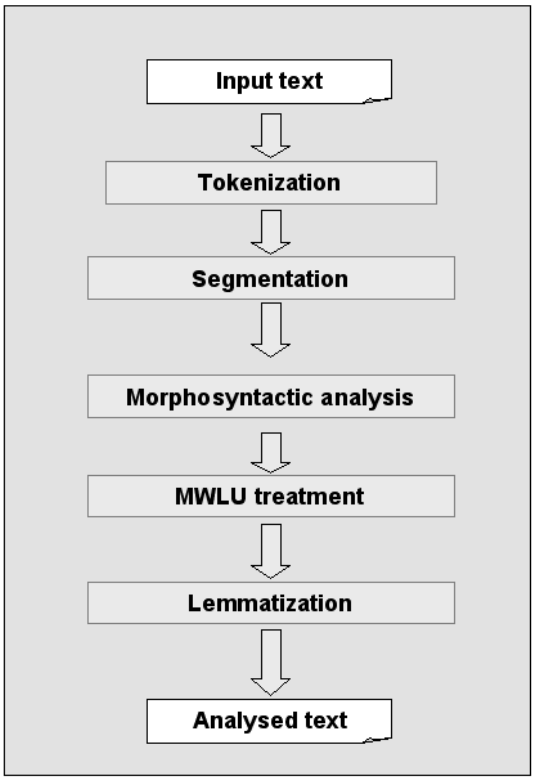

Figure 2.- Steps in linguistic tagging

- tokenizing to identify tokens and sentences
- morphological segmentising, which splits up a word into its constituent morphemes
- morphosyntactic analysis whose goal is to group the morphosyntactic information associated with a word
- the treatment of multiword lexical units (MWLU), like dates, numbers, named entities...
- disambiguation and lemmatisation: 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

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*.

2.4.2. Lexical resources and lemmatisation criteria

The lexical information used throughout the whole process comes from the EDBL lexical database (Aldezabal *et al*., 2001). EDBL is a permanent lexical storage facility, separate from the ZT Corpus. Its aim is to reflect the general lexicon of unified Basque, and the language units it collects are classified according to three different specialisation criteria: a) independent language units (dictionary entries) or non-independent morphemes; b) single-word language units or multi-word language units; and c) standard units or non-standard units; among the latter, they also note whether the two units are variants of each other.

With regard to the lemmatisation of the corpus, the treatment of standard and non-standard variants is very important. Needless to say, words that are not variants of each other have different lemmas, even if one is the standard or preferred form of the other. For example, even if the Unified Dictionary of Euskaltzaindia –the Royal Academy of the Basque language–, and therefore EDBL too, say that *oroimen* and *oroitzapen* are the standard or preferred forms for *memoria* ("memory"), it is clear that the lemma for the occurrences of the word *memoria* is *memoria*, irrespective of whether it is standard or not (aside from the fact that the decision of the Unified Dictionary regarding the word does not take into account its meaning in computer sciences). But *jarduera* / *iharduera* ("activity"), *elkarzut* / *elkartzut* ("perpendicular"), *immunitate* / *inmunitate* ("immunity") and many others are variants of each other. The EDBL has information about variants, and *Eustagger* uses this information when giving the lemma for a non-standard form, assigning it its standard variant. Thus, if we ask for the lemmas *jarduera*, *elkarzut* or *immune* in the web query interface of the ZT Corpus, occurrences of *iharduera*, *elkartzut* and *inmune* will also be displayed.

Besides, *Eustagger* can assign a single lemma to some systematic variant cases that are not present in the EDBL: for example, single phonological variants, like *–o*/*-u* endings, *tz*/*tx*/*ts* variants, *etc.* For instance, the EDBL has the lemma *kartutxo* ("cartridge") but not *kartutxu*; nevertheless, *Eustagger* proposes *kartutxo* as the lemma for the occurrences of *kartutxu* and its inflections, because you can arrive at the other by applying a single phonological rule.

As has been stated above, the EDBL's objective is to collect the general lexicon of Basque, so the need to enrich it with specific vocabulary is clear, if we want to use it in a specialised corpus. So in order to increase the precision of the linguistic annotation, we have added a complementary lexicon. The vocabulary consists of various scientific-technical terms that are not normally used in general language, and

therefore are not included in the EDBL. Thus, when lemmatising / annotating occurrences of those words or terms, the system will lemmatise them directly, before trying other options.

The complementary lexicon has been made up using the following sources:

- The Elhuyar dictionary database(ElhDB): We compared the database of the Elhuyar dictionary with that of the EDBL and added the lemmas that were not in the EDBL to the complementary lexicon; in some cases, when the new terms were quite general, they were added to the EDBL directly.
- The ZT Corpus itself: The corpus has been preprocessed to detect the words that were not recognized by the EDBL+ElhDB, and these words were sorted according to the frequency of the lemmas proposed by *Eustagger*; we checked the most frequent ones and, when appropriate, they were included in the complementary lexicon.

These two tasks are carried out prior to the linguistic annotation, during the structural annotation. For the second task, we developed a custom program and interface, which were included as a module in *Corpusgile*. The module is run before the stage of correcting the detected non-standards.

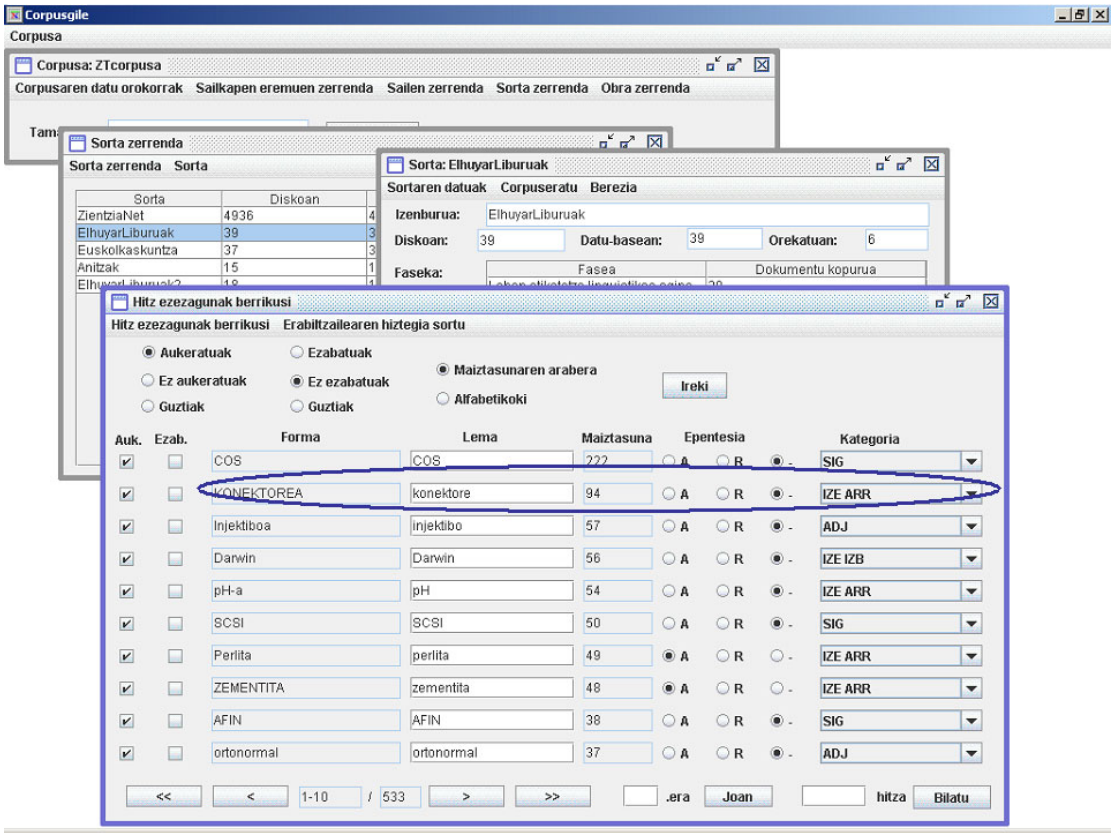

Figure 3.- Interface for enriching the complementary lexicon

Variants and non-standards that were not in the EDBL or which cannot be linked to any entry in the EDBL through a single phonological rule, have not been assigned a standard lemma, neither in the automatic processing nor in the manual revision. For example, neither *protisto* nor *protista* ("protist") appear in the EDBL. The occurrences of *protisto* and *protista* in the corpus have been assigned their own lemma, without a decision being taken as to the correct lemma. As a matter of fact,

these kinds of decisions should be taken by the members of the Language Academy on the basis of the occurrence frequencies of each in the corpora.

Additionally, to ease the 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.

2.4.3. Linguistic information

At the end of the linguistic annotation process, every word in the corpus will have some linguistic information attached, such as:

- Lemma and POS (100 percent correct in the manually revised part, and automatically assigned otherwise)
- Case and syntactic function (automatically assigned)
- In the case of multi-word expressions, their structure will also be represented

N-N (noun-noun) compounds joined by a hyphen have been annotated as a single lemma: *mahai-inguru* ("panel discussion"), *haize-energia* ("wind power")…

In any case, with respect to multi-word terms or NN compounds, the linguistic information of each component has also been kept, and the user of the query interface has the option of looking into the components as well. This option is very interesting, as NN compounds in Basque can be written with or without hyphen. For instance, to say "wind power" both *haize energia* and *haize-energia* are possible, so if we only kept the compound lemma of unions with hyphen, the hyphen unions would not appear when looking for *haize* ("wind"). If we activate the option of looking into components, looking for *haize* will bring results of both *haize energia* and *haize-energia*.

2.4.4. Stand-off annotation model

The linguistic annotation of the ZT Corpus has been done using a stand-off annotation model. 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
- one does not have 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

Figure 4 provides a graphical display of the links between documents.

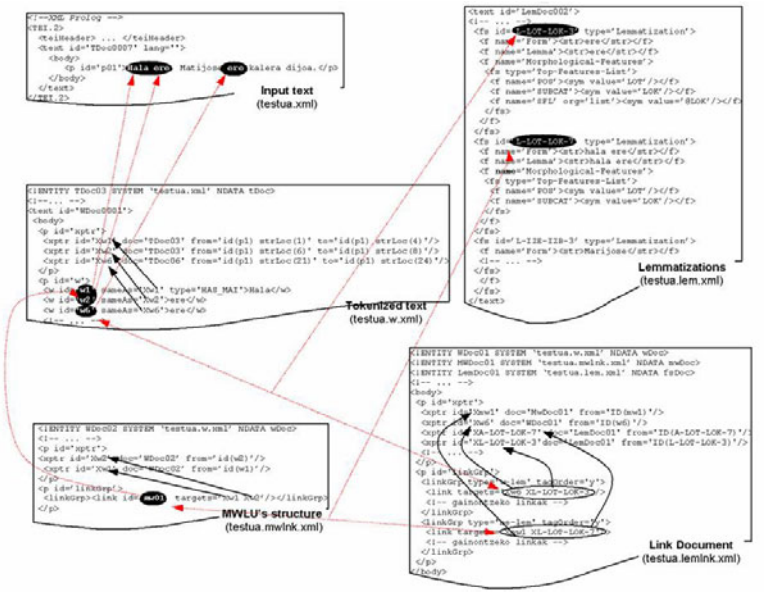

Figure 4.- Stand-off representation: anchors, linguistic information and links

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 three main modules in the application:

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

These modules have been used during the corpus compilation and construction. Apart from *Corpusgile*, a query interface to consult the corpus via the Internet has also been developed.

3.1. The corpus builder

It is based on a relational database and 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.

Figure 5 shows the main interface for the Corpus Builder.

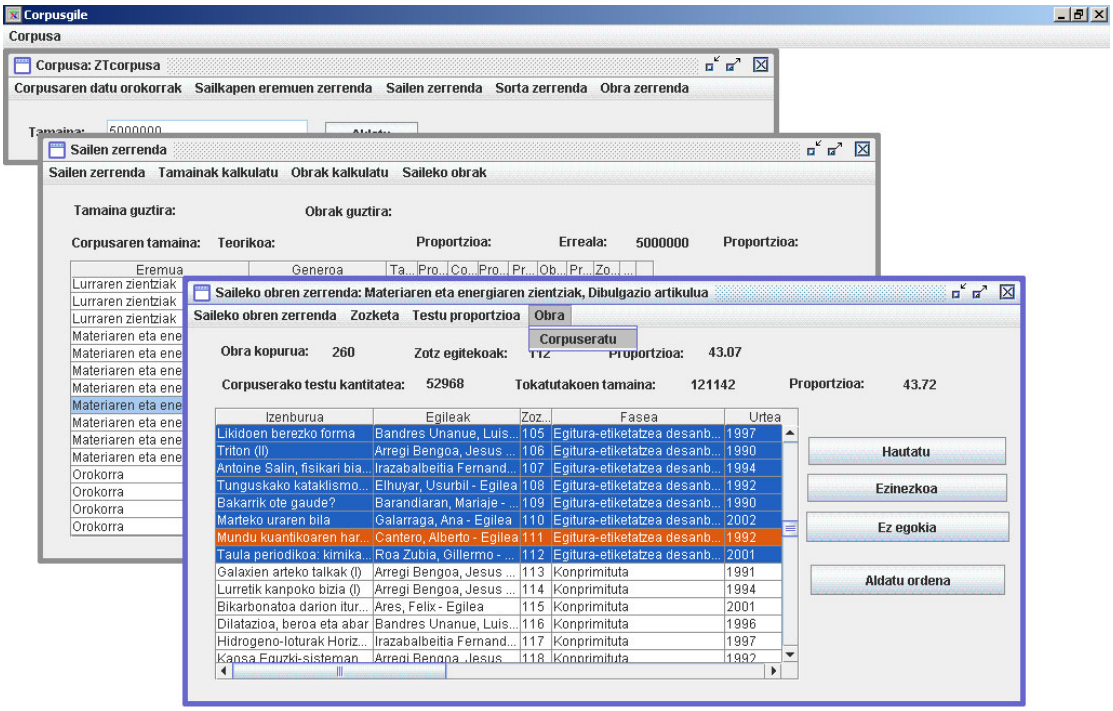

Figure 5.- Main interface for the Corpus Builder

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 specialised or technical words to the corpus-specific lexicon in order to improve future linguistic tagging; to achieve this, an NLP tool called *Eustagger*, a lemmatising/disambiguating tool based on the former *Euslem* (Aduriz *et al*., 1996), is used to detect non-correct words, which are then ordered according to the 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>`, `<head>` and `<note>`) when appropriate
- interface for correcting, improving and disambiguating the structural tagging of the balanced part
- verification and validation of XML/TEI structures

Figure 6 shows the interface when a non-standard use is tagged and linked to the standard one.

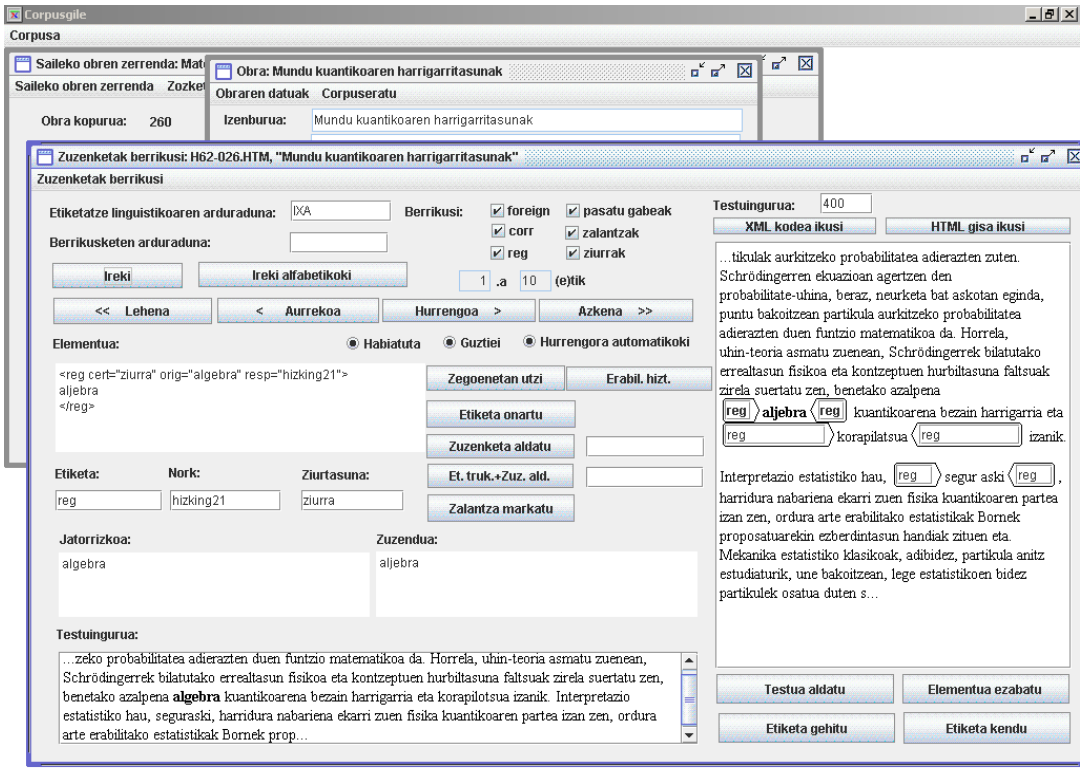

Figure 6.- Interface for manual revision of <reg>

3.3. The linguistic tagger

Linguistic annotation is carried out using *Eulia*, a framework for creating, browsing and editing linguistic annotations. It is based on a class library named LibiXaML, and 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, lemmatisation/disambiguation, shallow syntax and dependency-based analysis.

After the automatic processing, which generates the XML documents, a module for manual linguistic annotation can be used. This module integrates the results of the automatic processes and provides the linguists with a friendly interface for the annotation, hiding the complexity of the multiple files that have to be managed. The main interface is shown in Figure 7.

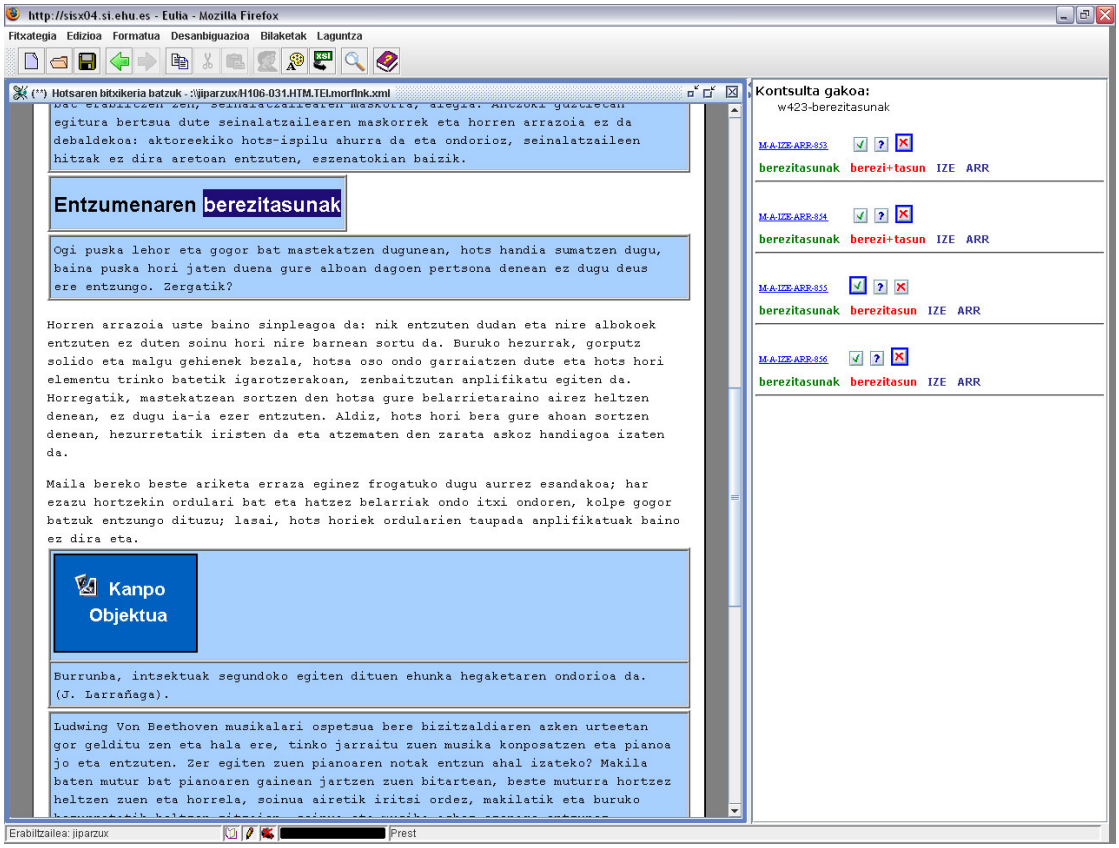

Figure 7.- Main interface in *Eulia*

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

In the text window the linguist can click a token and receive an offer for a set of actions to be performed.

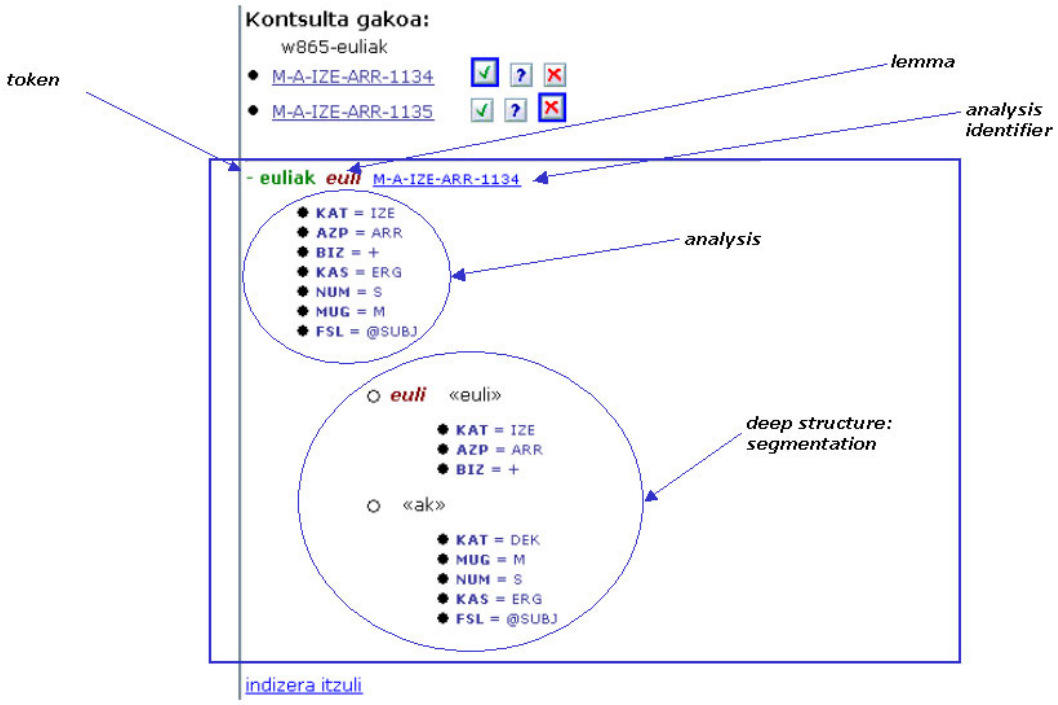

Figure 8.- Analysis window

The main action is to show, in the analysis window, the different possible analyses in order to disambiguate them. In any case, 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.

In Figure 8 we can see the top of this window by way of an example. Information for the whole word, *euliak* ("flies"), and for the two morphemes, *euli* ("fly") and *ak* (ergative singular morpheme), are given.

3.4. The query interface

The ZT Corpus has been put online for its querying through a web interface. This interface is user-friendly and easy to use in its *normal* mode, yet it also offers some very interesting more sophisticated options in the *advanced* mode. Many of the query options of the ZT Corpus are new in Basque corpora.

3.4.1. Query options

The user can query for up to three words, which can be at a distance of up to four words –either forwards or backwards– from each other. For each of the words, the user can choose to query for the lemma or a specific word form, and he or she can ask for the complete word, the beginning, or the ending of it. Optionally, he or she can restrict the query of the word to a particular POS –when combined in a multi-word query, it is possible to ask for the POS alone.

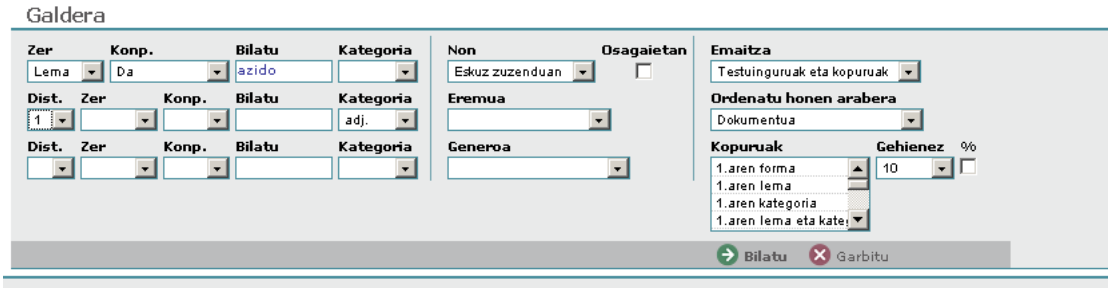

Figure 9.- Advanced query options

All these choices make it possible to conduct a wide combination of queries, from very simple to very complex ones, such as:

- Words whose lemma is *ekuazio* ("equation")
- Words whose lemma begins with *programa* ("program"), thus obtaining occurrences of *programa* ("program"), *programatu* ("to program"), *programatzaile* ("programmer"), *programazio* ("programming"), *programazio-lengoaia* ("programming language"), *etc.*
- Words whose lemma is *azido* ("acid") followed by an adjective
- Words whose lemma is *energia* ("energy") that have an adjective in a neighbourhood of four words

We can also choose to run our query either in the manually disambiguated part alone or in the whole corpus. Furthermore, queries can be restricted to a single specific topic, genre, or both.

Apart from individual words, many common expressions, terms, entities, phrasal verbs, word combinations, *etc.* have been indexed, so it is possible to look for these too, even non-contiguous occurrences of multi-word terms.

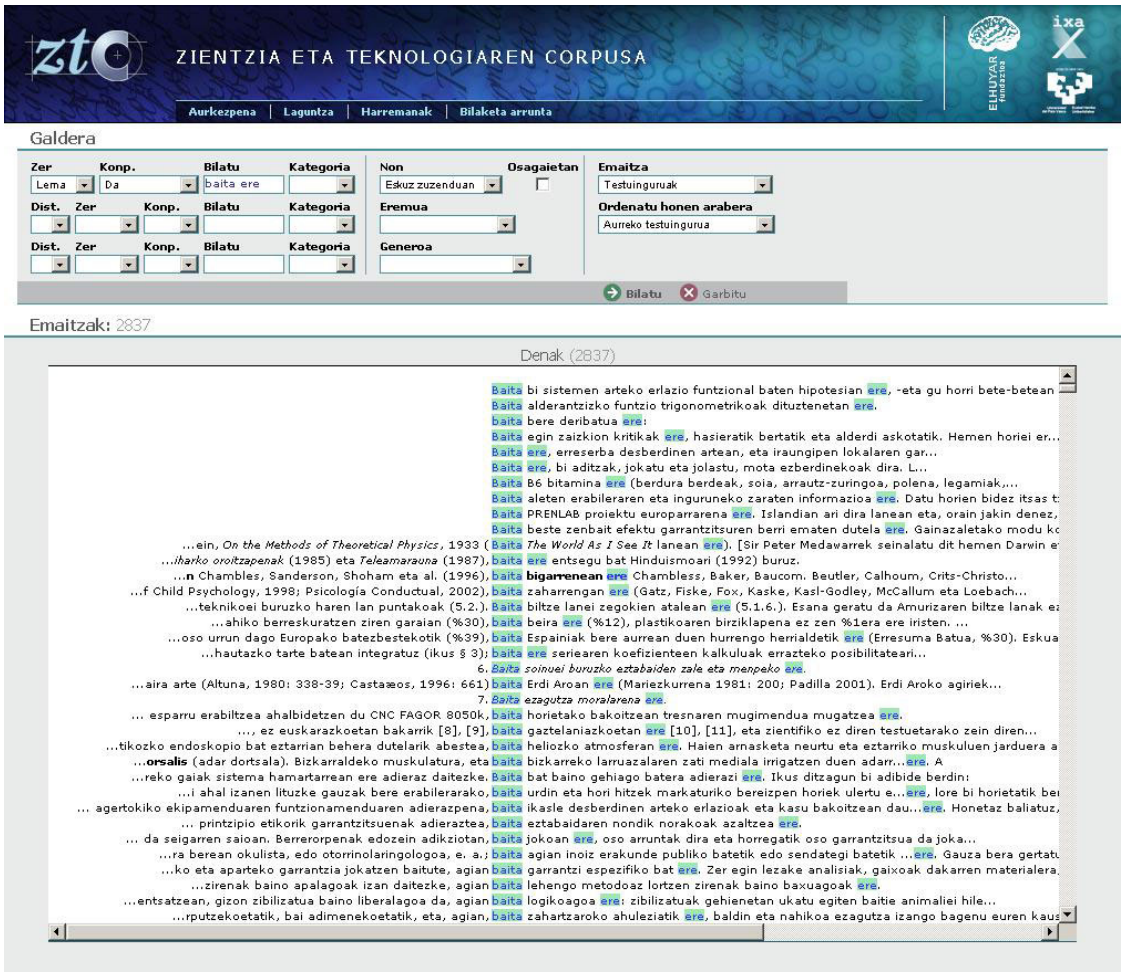

Figure 10.- Multiword units: results of lemma=*baita ere* query
(non-contiguous occurrences are recovered)

The user can choose to execute the query on the components of those multiword units activating the option *Osagaietan* ("look also into the components") –if this option is not activated, the whole multi-word unit is taken as the lemma–. This is very useful to analyse a kind of Basque NN compounds, which can be written with or without an inner hyphen, as *energia-iturri* or *energia iturri* ("energy source"). The NN compounds with hyphen, even those not included in dictionary, are processed as a single token. Nevertheless, the linguistic information about each component is also tagged and stored; this is necessary if we are interested in analysing the NN terms of a given noun. For instance, if the *Osagaietan* option is not activated, the query for the lemma *energia* would not retrieve the occurrences of *energia-iturri*, as this is tagged as one lemma. Using the option *Osagaietan*, we can retrieve all the occurrences of the nouns before or after *energia* with their frequencies, no matter the type of hyphen usage. This functionality is very useful for terminological work.

3.4.2. Results

In the default behaviour, the query interface of the ZT Corpus shows a table and a chart of the form or lemma counts of the word that has been searched for –depending on whether the search was for a lemma or a form–, plus a list of all the occurrences of the word in a KWIC context.

These contexts are not plain text but formatted text, so they are as close to the original document as possible –we have already pointed out that the TEI structural annotation of the corpus keeps the typographical information of the documents–. We believe this is very important in a science and technology corpus, full of equations, chemical formulae, terminology, *etc.*, which would lose their sense and render the text incomprehensible if we removed the italics, subscripts, superscripts, *etc.*

The words searched for are shown in different colours, depending on the certainty of their linguistic analyses –manually corrected (clear green), unambiguous (dark green) or ambiguous (yellow or red)–. We have judged this as necessary on the ground that when the user makes a query on the whole corpus, it is worthy to make it clear whether the linguistic analysis has been manually surveyed or automatically processed, and, in the latter case, the level of ambiguity and certainty of the analysis. Moving the mouse over any of them will show the linguistic analyses –lemma and POS– of each one in a floating window, and clicking on any of them will open a new window which shows a bigger context of the word –about 300 words long– and reference information (authors, publisher, title, *etc.*) about the document it was in.

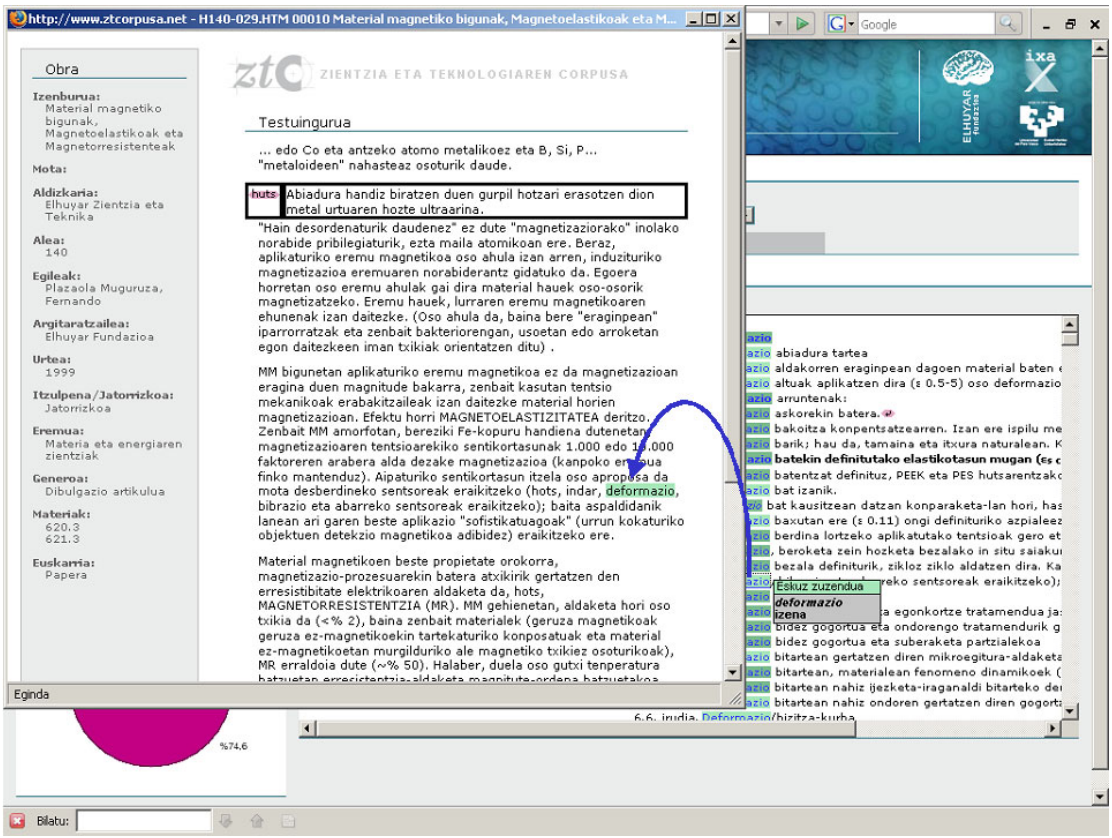

Figure 11.- Larger context display

The KWIC contexts can be ordered and grouped in accordance with different criteria: by document, lemma, POS, topic, genre, year, context after the word, context before the word, *etc.*

The occurrences of non-standard variants are retrieved and displayed according to the lemmatisation criteria mentioned above. For example, the occurrences of *inmunitate*, a non-standard variant of *immunitate*, are also shown when we enter *immunitat*e as the queried lemma.

In the *advanced* mode we can tell the system to show only the KWICs or the tables and graphs alone, and we can also choose which graphs we would like. There are many available: word forms, lemmas, POS, topic, genre, year, form or lemma of the word before, form or lemma of the word after, *etc.* In multi-word searches, the graphs can display the features of any of the searched for words –form, lemma, POS, form or lemma of the word before, form or lemma of the word after, *etc.*

These tables and charts are of great interest in order to show the collocational or combinational behaviour of words, as we can ask the system for a word like *ekuazio* ("equation") or *azido* ("acid") or *energia* ("energy") followed by an adjective, and show a chart with the counts of the lemmas of the second word, for example.

The 'tables and charts only' mode can be very useful to show the behaviour (counts, lemmas, combinations) of a word in a single view with various charts.

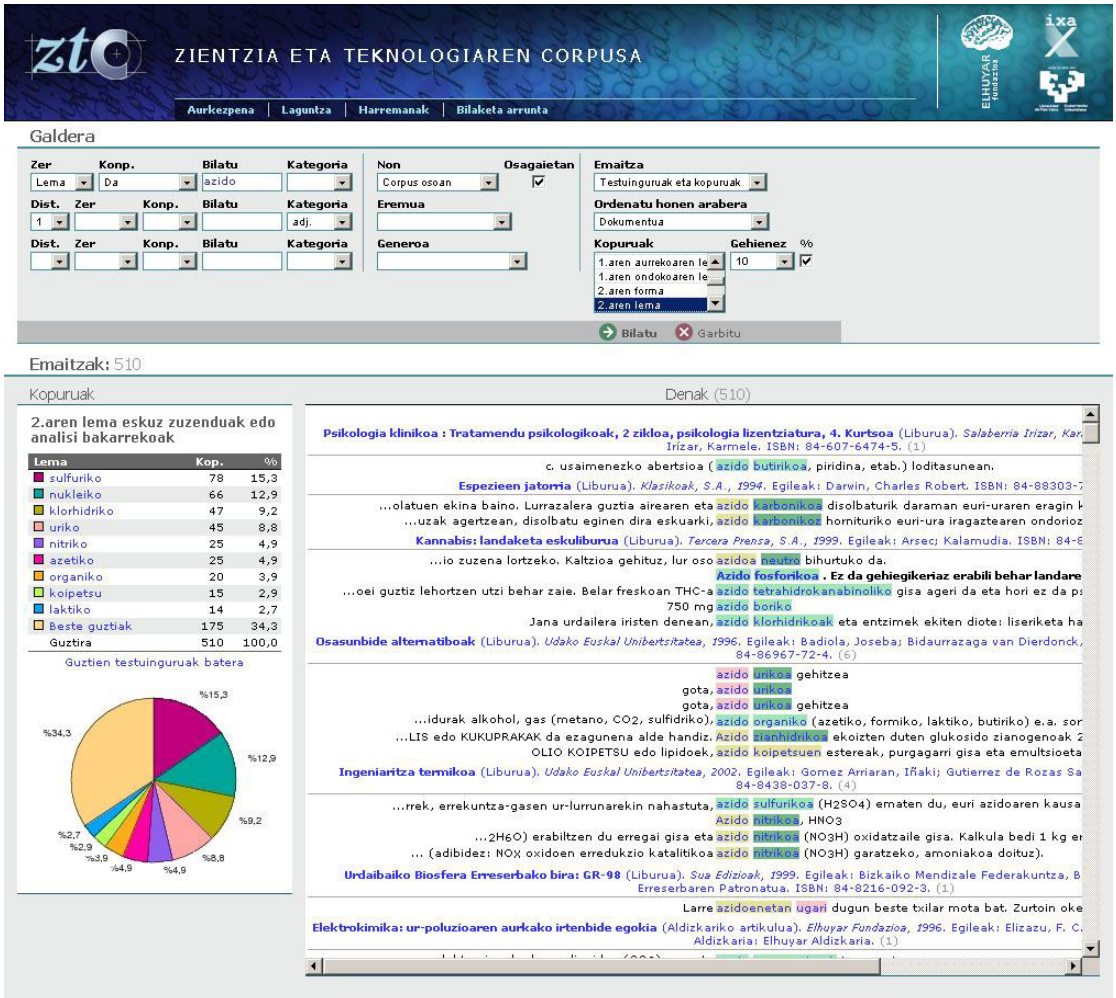

Figure 12.- Query interface with results: occurrences of adjectives following *azido* ("acid"), with KWIC ordered by document.

4. Conclusions

Just like any other language, Basque needs corpora. Linguists, terminology specialists, language technology researchers, people that work in language standardization and normalisation… Many people need corpora, as they constitute an essential tool nowadays for language analysis. The Basque ZT corpus aims to be a useful and powerful tool for conducting research on specialised texts in Basque.

However, Basque is a small language in terms of speakers and, consequently, in terms of the resources devoted to it, so we need something else in addition to 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 which will reduce the generally high costs involved in building them. We have made such a tool, *Corpusgile*. This tool provides a flexible and extensible infrastructure for creating, visualising and managing corpora, and for consulting, visualising and modifying annotations generated by linguistic tools. The interface has been designed to be informative, easy-to-use and intuitive. And due to the fact that it is based on 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 developed and applied a methodology that can be used 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 made to this field, which we are so very much in need of, which is the field of corpora. And we are convinced that in the future they will prove to be very valuable contributions indeed.

Additionally, the corpus is online, available to be queried through a user-friendly yet powerful interface, and is a very valuable tool for language researchers, dictionary makers, technical text writers, *etc.*

5. Acknowledgements

This work was partially funded by the Basque Government (EJ-ETORTEK-2002/HIZKING21, EJ-ETORTEK-2006/AnHitz, EJ-TEK-2005D0/0005 and EJ-IKT 2006Be/0001) and the University of the Basque Country (EHU/GIU05/52).

6. References

Aduriz I., I. Aldezabal, I. Alegria, X. Artola, N. Ezeiza and R. Urizar (1996) 'EUSLEM: A Lemmatiser / Tagger for Basque'. *Proc. EURALEX'96, Göteborg, Part 1*, 17-26. Available online from http://citeseer.ist.psu.edu/cache/papers/cs/8816/http:zSzzSzixa.si.ehu.eszSzdokumentzSzArtikuluzSz96EUSLEM.pdf/aduriz96euslem.pdf (accesed: 25 June 2007)

Aldezabal I., O. Ansa, B. Arrieta, X. Artola, A. Ezeiza, G. Hernández and M. Lersundi (2001) 'EDBL: a General Lexical Basis for the Automatic Processing of Basque'. *IRCS Workshop on Linguistic Databases, Pennsylvania*, 1-10. Available online from http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1011897592/publikoak/2001-IRCS.pdf (accessed: 25 June 2007)

Artola X., A. Díaz de Ilarraza, N. Ezeiza, K. Gojenola, A. Sologaistoa and A. Soroa (2004) 'EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora'. *LREC 2004, Lisbon, Workshop on XML-based richly annotated corpora*. Available online from http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1088448358/publikoak/04LREC_EULIA.pdf (accesed: 25 June 2007)

Bach C., R. Saurí, J. Vivaldi and M.T. Cabré (1997) El corpus de l'IULA: descripció. Bartzelona: Universitat Pompeu Fabra, Institut Universitari de Lingüística Aplicada. Available online from ftp://ftp.iula.upf.es/pub/publicacions/97inf017.pdf (accesed: 25 June 2007)

Ide N., L. Romary and E. Clergerie (2004) 'International standard for a linguistic annotation framework'. *Natural Language Engineering 10*, 211-225. Available online from http://acl.ldc.upenn.edu/W/W03/W03-0804.pdf (accessed: 25 June 2007)

Sinclair, J (1996) Preliminary Recommendations on Corpus Typology. EAGLES. Available online from http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html (accessed: 25 June 2007)

TEI - Text Encoding Initiative (2004) The XML version of the TEI Guidelines. Available online from http://www.tei-c.org/P4X/ (accessed: 25 June 2007)