# The Lexicographic Work of Euskaltzaindia - The Basque Language Academy 1984-2009

**The roadmap**

Euskaltzaindia was founded in 1919, but until 1968 failed to start developing a standard variety of Basque Language. From 1968 to 1984, the lexicographic work of Euskaltzaindia was not very consistent, but rather limited to meet the most acute needs. In 1983 the Academy created a commission of lexicography, and in 1984 approved a long-term plan for the development of dictionaries. That plan provided a range of activities for the period 1984-1996, which were mainly the following:

1) The General Basque Dictionary, which should be a compilation of the lexicon used in the publications until 1970.

2) A lexicology project, whose aim was to study the formation of words in Basque.

3) A compilation of the lexicon used in current publications

All these activities have the common objective of laying the groundwork for developing a unified standard dictionary of Basque.

**The General Basque Dictionary**

In 1984 started The General Basque Dictionary. The aim of this dictionary is describing the basque lexical heritage of all time, taking into account all dialects. It spans from antiquity's inscriptions to medieval texts and everything published since the advent of printing until the mid eighteenth century. The literature is growing much from that time. For this reason, the corpus is not so comprehensive for modern times. However, one can say that this 6-million-words corpus is a very complete reference of what was published until 1970.

It is a computerized corpus, but has not been annotated nor lemmatized.

The first result of the project was a dictionary of 16 volumes in paper format, with a total of over 14,000 pages. These volumes were published between 1987 and 2005. The dictionary entries contain the following information:

The lemma, and the variants of the word as they appear in texts, information about the dialects corresponding to these variants, according to usage in the last hundred years, the meanings**,** the history and examples of use of the word**,** compounds, expressions and etymology**.**

Since October of 2009 this dictionary is available online at the address http://www.euskaltzaindia.net/oeh. Queries can be done in the field corresponding to the entry and response is a true reflection of the article on paper.

Users can also download the dictionary in PDF format at the address http://www.euskaltzaindia.net/oeh/jaisteko_gunea.

## The statistical corpus of twentieth century

In 1986 the committee of lexicography held sessions to determine the period of reference for the collection of the current lexicon. The period finally selected was the twentieth century. However, the volume of text published in the Basque Language throughout the whole century was too large for to opt for a comprehensive corpus. For that reason, it was decided that the body would comprise a representative sample selected through a statistical model.

Specifically, we used a stratified sampling, so that in the sample were suitably represented four time periods (1900-1939, 1940-1968, 1969-1990, 1991-1999), 14 types of text (literary prose, drama, essay, textbooks, etc.), three sizes of text, and the different dialects of Basque literature.

Previously we did an exhaustive list of all publications of the period and they were classified according to the above criteria.

We decided that the size of the corpus would be of 2,000,000 words, to make the project feasible with available resources.

Then was completed the corpus for the rest of the century, following the same statistical model.

Thus, the final size of the corpus was 4,658,036 words from 6,351 pieces of text.

The texts, mostly, were digitized by scanning, OCR and manual correction.

The entire corpus is lemmatized, using a semiautomatic procedure and corrected manually.

The result of this project is available online at the address

http://www.euskaracorpusa.net/XXmendea/Konts_arrunta_fr.html

The query can be about a lemma, a word or an initial or final fragment of a word. It also supports boolean combinations of these elements.

You can filter the results taking into account the period, dialect and text type.

The result of the query is a set of contexts that contain the element has been consulted, with bibliographic information about the texts to which it belongs.

## The Unified Dictionary of the Basque Language

In 1992 the Academy created a commission to prepare The Unified Dictionary of the Basque Language.

This dictionary should contain the words most needed in everyday life, setting its standard form.

The reference to establish the standard form of a word is the use of the same in the historical tradition and in modern times. The General Basque Dictionary and The Statistical Corpus of the Twentieth Century are precisely the information sources that provide insight into the use of words.

The commission's proposals are subject to review of certain qualified users (teachers, translators, writers, etc.) and finally to the approval by the plenary of the Academy.
In 2000, was published a list of standardized 20,000 words and in 2008 a second edition collected a total of 29,000 words.

By the end of 2011 the list should contain about 40,000 forms. Regardless of the publications on paper, the unified dictionary is available at http://www.euskaltzaindia.net/hiztegibatua and can be downloaded in PDF format from http://www.euskaltzaindia.net/eaeb along with the rest of the rules of the Academy.

The documentation generated in the process of developing the unified dictionary can be found at this address.

**The Dictionary of the Academy**

Currently a working group is drafting the articles of the Dictionary of the Academy, for the first 20,000 entries of the unified dictionary. This work will be completed in late 2010. Alongside a committee of academics is reviewing that work and the Academy is expected to approve it in 2011.

**The Lexicon Observatory**

This project aims to create a monitor corpus, that is to say a corpus designed to show the changes that are occurring in the use of Basque. It is intended as a first step towards a large reference corpus, balanced, lemmatized, annotated and corrected manually, but for now only is possible to be an opportunistic corpus for reasons of cost.

The corpus has the following characteristics:

- It contains texts published after 2000
- Basically there are texts from the media or informative nature
- Preference is given to texts that offer facilities to be processed automatically
- Will be fed continuously

The processing of texts consists of the following steps:

1) Collection of texts
2) Conversion of the format and structure labeling (TEI)
3) Automatic language processing: tokenization, morphological segmentation, analysis of units consisting of several words and automatic disambiguation
4) Manual disambiguation and correction

The corpus also has tools for querying, analysis and exploitation of data.

Automatic processing is based on tools developed over the last two decades and have been tested and improved in other projects: the lexical database **EDBL,** the MORFEUS morphological analyzer, the EUSLEM lemmatizer and the **EULIA** environment that provides a flexible and extensible environment for creating, consulting, visualizing, and modifying documents generated by existing linguistic tools. All of them have been developed by the IXA Group of Natural Language Processing, Computer Science Faculty, University of Basque Country.

In late 2009 the body had a volume of 4,000,000 words and by 2012 will reach 49 million words.

Euskaltzaindia collaborate on this project with the Basque Center for Terminology and Lexicography (UZEI), Elhuyar Fundazioa and the IXA group.

The Commission of the Unified Dictionary conceived the project, with the help of these institutions, and designed the strategy of exploitation of the data.

This strategy is intended mainly to establish whether the rules issued by the Academy are met in practice, and also detect new words and new uses of existing words.

The corpus also provides interesting data for the study of grammar, onomastics or stylistics.

Although not a reference corpus is prepared to correct the imbalances that inevitably has an opportunistic corpus. In this regard, cataloging provides the ranking of texts by various criteria.

In summary, one can say that in forty years the Academy has succeeded to improve many of the shortcomings that plagued the Basque language to mid-twentieth century as the lack of a standard variety and the lack of a standard dictionary.

Fortunately, the Academy is not alone in the challenges posed to the normalization of the Basque language; other institutions developed terminological dictionaries, bilingual dictionaries, encyclopedias and other reference works.