BIZKAIFON: A sound archive of dialectal varieties of spoken Basque.

I. Hernáez, E. Navas, J. Sánchez, I. Madariaga, I. Gaminde, X. Zalbide

University of the Basque Country
Dpto. de Electrónica y Telecomunicaciones, ESI Bilbao
Alda. Urquijo s/n, 48013 BILBAO, SPAIN
{inma, eva, ion, imanol, xabier}@bips.bi.ehu.es

Abstract

This paper presents the sound archive of dialectal varieties of spoken Basque called BIZKAIFON. This database contains sound archives with their associated information and it is accessible via a web interface. A prototype of BIZKAIFON is available at http://bizkaifon.ehu.es/.

1. Introduction

The Basque language, (euskara in Basque) despite of being one of the oldest languages in Europe and being subject in several studies is full of linguistic unknowns, including its place of origin. Nowadays there are more than 1,000,000 speakers of Basque in the Basque Country area (near the political border of France and Spain), and about 200,000 in many other areas around the world, most of them in America.

In spite of the small number of speakers, the Basque language presents a huge dialectal fragmentation: depending on the author and on the classification criteria there are up to 7 main dialects and more than 50 minor dialectal varieties.

The process of standardization of the written Basque started in 1968 and is under continuous revision. Thanks to this standardization process, there has been an unquestionable recovery and expansion of the use of the Basque language. The standard Basque is learned from primary education to graduate level for some curricula. There is one public TV channel and one newspaper fully in Basque, and a few radio channels, and weekly or monthly publications. But this process has also been accompanied by some negative effects, one of them being the fact that some of the smaller local varieties are disappearing. This is one of the main reasons to collect, structure and make available existing audio material as well as some newly recorded sound archives regarding every local variation.

Another point to be considered is the lack of pronunciation rules for the standard. The standardization process has never included this aspect of the language, except for a few recommendations. The use of recorded audio can be a very appropriate guide for new speakers, as well as an invaluable tool for researchers if a minimum set of pronunciation rules were to be defined.

The organization of this paper is as follows: next section describes the main purposes of BIZKAIFON. Section 3 presents the potential users profile. Section 4 describes the database architecture, and in Section 5 we explain the way it has been implemented, including the addition of new data, SGML labeling and developed tools. Finally, some conclusions are drawn in section 6.

2. Main purposes

The main purposes of this sound archive are:

- To collect, preserve and classify the existing sound material, preventing linguistically highly valuable recordings from loss.
- To implement an open structure where new sound documents recorded with different purposes can be integrated.
- To provide an important space for the Bizkaia's dialectal variations aiding not only the diffusion of the standard Basque language, but particularly the most minority variations of Bizkaia's Basque.
- To provide a high valuable tool for linguistic researchers who have interest on different local Basque language variations, allowing linguistic or even social research.
- To support the Basque Government (Eusko Jaurlaritza) literacy campaign, providing a highly spread user friendly telematic tool.
- To supply a meeting point for researchers, engineers and linguists from all around the world who can be interested on phonetic and acoustic aspects of the speech.
- As a web site, it offers the advantages of using Internet: quick document access, place and time independent accessibility and multimedia presentation.

3. Potential user profile

One of the purposes of the BIZKAIFON sound archive, maybe the most important one, is to keep recorded speech from destruction and make it available for a wide spectrum of users. Taking this goal into account, three main types of users can be found:

- Professional people: ethnologists, linguists, teachers and technical researchers.
- Students: BIZKAIFON contents can be used as a material basis for different studies. This way, students can hear people speaking about subjects of their interest, to know details about folklore and popular culture, to know the way certain technical terminology is used in a local variation. Specially, Basque language students have a support, being able to make listening practices, speaking practices and paralinguistic matters study.
- Other people: everybody can use BIZKAIFON driven by knowledge desire. Many people can look for their local variety or even their neighbors' speech in order to compare.

4. Database architecture

4.1. Material types

Currently there are four different types of material in the database:

- Popular literature: archives that keep record of the popular knowledge, songs and folklore.
- Texts: any other speech that people tells.
- Sentences: grammatically complete sentences.
- Isolated words: words that have been pronounced separately, usually to make a complete lexicon of a certain place.

4.2. Database structure

The database is structured as follows: sound files placed in different directories, and, for every sound file, a text TEI SGML file which stores data according to the TEI Consortium P4 recommendation (Text Encoding Initiative, 1994), and a binary file containing the same data as the SGML, stored in a format that grants faster access to the data.

4.2.1. Audio files

The voice recordings are stored in Windows PCM audio files (wav). Every sound file in the database uses 16 bit samples, and various sample rates can be accepted. All the recordings are single channel (mono).

4.2.2. SGML files

These text files store the following data in the TEI P4 recommendation format (Text Encoding Initiative, 1994; Real Academia Española, 1999):

- A technical description of the file, including the following data:
    - Signal file name or path.
    - TEI SGML text file actualization data.
    - Collection Title (when data from a previous compilation has been used)
    - Collection ID (according to the Collection Title)
    - File Title (in the case of popular literature)
    - Name of the person responsible for the labeling
    - Name of the publisher
    - Copyright Information
    - Name of the person responsible for the recording
    - Recording equipment
    - Recording date
    - Material Type
    - If the material type is either text or popular literature, sub-type.
    - Id codes of the speakers.
    - Sexof the speakers.
    - Name of the speakers (if known).
    - Recording region
    - Recording place
    - Some keywords to identify the subject
- The time marks for every marking level: multiple mark levels are accepted: text level, sentence level, word level, diphoneme level, phoneme level… New levels can be registered and then included in the SGML files.
- The transcription or transcriptions of the file (Listerri, 1997). A standardized Basque orthographic transcription is mandatory, except for the popular literature files, which are referenced by a standardized Basque title. Many other types of transcriptions are accepted: phonetic transcriptions, literacy transcriptions… When new recordings are added onto the database, new transcription types can be registered and then included in the SGML files.

4.2.3. Binary files

Though the use of standard TEI text files is very extended, parsing SGML text files needs extensive use of process capability. Thus, in order to search the database efficiently, the same data that is stored in that SGML text file is copied into binary files. The format of these files is explained in figure 1.

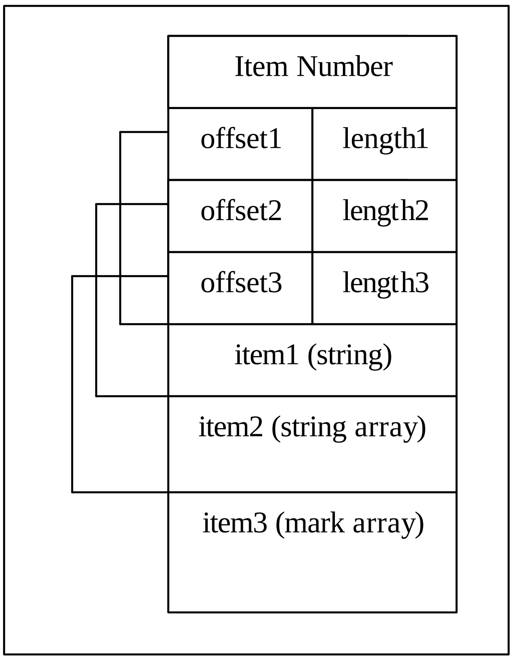


Figure 1: Binary file structure.

As seen in the figure, the binary file has a header which tells first the item number, and then, for every item, the starting position and the length of the item. Then there are the items themselves, leaving no gaps among them. Currently there are three types of items defined:

- Strings: when a single text data is stored for a single SGML field. I.e. the recording place.
- String arrays: when several text data is stored for a single SGML field. I.e. the keywords.
- Time marks: for every mark, the starting time, the duration and a label are stored.

Accessing these binary files instead of the SGML plain text files allows better performance. The translation from text to binary is made off-line.

5. System Implementation

In order to get a functional search engine, the database itself and the search and management software is placed into a Linux machine, making the search interface available on-line through an Apache web server.

5.1. Search Application

The web-based interface has three main tasks: sending and interpretating the search formularies; generating a full list of the data that fulfills the search criteria; and, for every recording, recovering all the regarding data from the data-base (including the sound itself), and sending it to the client browser when so is asked.

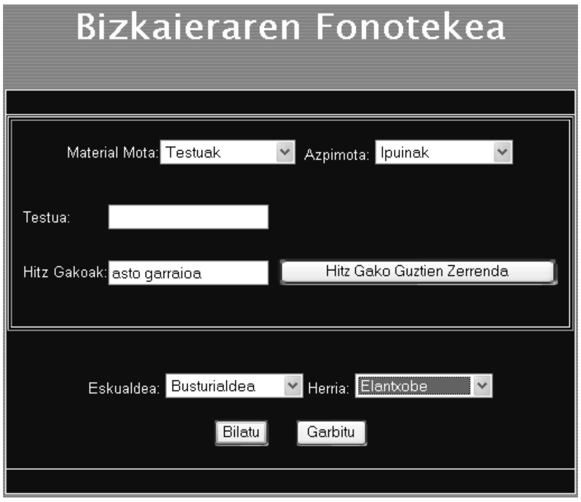
5.1.1. Search forms

Two different search forms have been developed: first, the standard search form, where the more efficient search is suggested according to the selected material type. In this case, different criteria can be imposed. First of all, the material type must be selected. Depending on the material type, the accepted choices are:

- Isolated Words and Sentences: searching for text in the standardized Basque transcription is allowed.
- Texts: Sub-type search and subject search (using keywords) are allowed, as well as the text search in the standardized Basque transcription.
- Popular literature: Sub-Type search and subject search are allowed, and the text search is replaced with a title search.

Together with these conditions, the search can be geographically restricted, by selecting a desired place or region.

For a search performed with the standard form, every result must accomplish all of the search conditions.



Bizkaieraren Fonotekea

Material Mota: Testuak Azpimota: Ipuinak

Testua:

Hitz Gakoak: asto garraioa

Eskualdea: Busturialdea Herria: Elantxobe

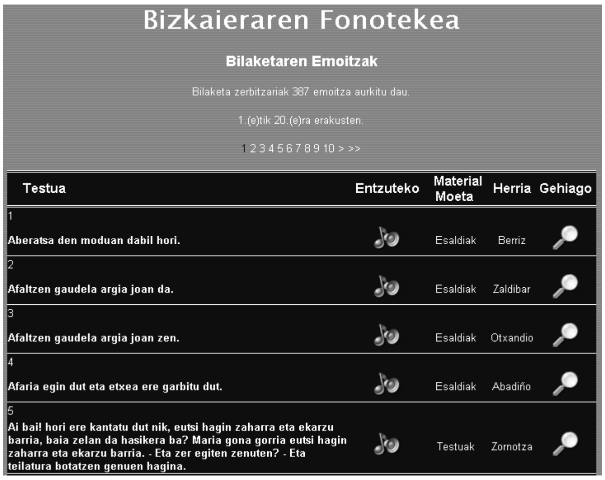
Figure 2: Standard search form.

Then there is the advanced search form, where all data in the SGML files can be taken as a selectable search criterion. Boolean combinations of conditions is allowed (and, or, not).

5.1.2. Result list

When a search form has been submitted, as a result the browser gets a dynamically created web-page containing a table, where, for every data-base record matching the search criteria, a row is added. The format of these rows can be configured via some system files, but usually it contents a result index number, a transcription of the text (or the reference title in case of Popular Literature), the place where the recording took place, and two links: the first one to allow users to access the sound directly, and the other one to grant access to the full information about the recording.

Figure 3 shows the look of the result list.



Bizkaieraren Fonotekea

Bilaketaren Emoitzak

Bilaketa zerbitzariak 387 emoitza aurkitu dau.

1.(e)tik 20.(e)ra erakusten.

1 2 3 4 5 6 7 8 9 10 > >>

Testua	Entzuteko	Material Moeta	Herria	Gehiago
1 Aberatsa den moduan dabil hori.	♫	Esaldiak	Berriz	🔍
2 Afaltzen gaudela argia joan da.	♫	Esaldiak	Zaldibar	🔍
3 Afaltzen gaudela argia joan zen.	♫	Esaldiak	Otxandio	🔍
4 Afaria egin dut eta etxea ere garbitu dut.	♫	Esaldiak	Abadiño	🔍
5 Ai bai! hori ere kantatu dut nik, eutsi hagin zaharra eta ekarzu barria, baia zelan da hasikera ba? Maria gona gorria eutsi hagin zaharra eta ekarzu barria. - Eta zer egiten zenuten? - Eta teilatura botatzen genuen hagina.	♫	Testuak	Zornotza	🔍

Figure 3: Result list

5.1.3. Extended data

When the user clicks on the extend data link, a pop-up window turns up, showing all the data that the SGML file contains, including metadata and several transcriptions. If speaker-dependant segmentation has been performed, the names of the speakers, and the transcription of their respective utterances are also shown. By clicking in a text, the user will get the sound the text refers to. The sound file is sent MP3 encoded. This format has been selected because it allows faster transfer times than the Windows PCM format used in the database, and, though the condition that the client computer must have an MP3-capable sound system is imposed, they are widely spread, so it appears to be appropriate.

5.2. Data addition

The data addition process involves sound recording, digitalization, selection, marking, transcribing and labeling. When a new sound file has been recorded, related data have to be created in order to index it into the database. These data have to be gathered in two parts: first, recording is made and selected, and then, with some signal edition software, sound files are marked and transcribed. A standardized Basque orthographic transcription is mandatory for every file, except of those which are Popular Literature. These files require bizkaian Basque transcriptions.

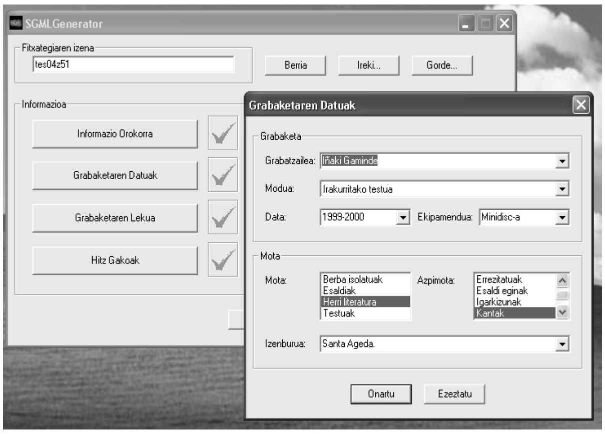
When transcriptions are ready, next step is creating the SGML file using the SGML code generator software; this software will ask for the necessary data, and take the transcription files to create the file. Finally, a binary file is created from the SGML text file.

5.2.1. Time labeling

Special signal analysis and labeling software has been developed. The AhoTools software (Etxebarria et al., 1998) allows the labeler to watch certain characteristics of the signal file, such as the waveform itself, the energy of the signal or its spectrogram. This information is valuable in order to generate mark files which store the time stamps and the desired transcription. The data in these files will be included in the TEI SGML files when they are created.

5.2.2. SGML labeling

In order to simplify and accelerate the generation of correct TEI P4 SGML files, some specific software has been developed. In a Windows friendly interface, the labeler is asked for the necessary data, restricting the data that can be typed to correct values (i.e. it is not possible to type a non-existent place or region name) When all the data have been introduced, the typed items as well as the mark files described in the previous chapter are used to generate correct TEI P4 SGML files.



Fitxategiaren izena: tes04z51

Grabatzailea: Iñaki Gaminde

Modua: Irakurritako testua

Data: 1999-2000

Ekipamendua: Minidisc-a

Mota: Berba isolatuak, Esaldiak, Herri literatura, Testuak

Azpimota: Errezitatuak, Esaldi eginak, Igarkizunak, Kantak

Izenburua: Santa Ageda.

Figure 4: Windows-based SGML Generator software

5.3. Complementary tools

Some Linux-based tools have been developed in order to maintain the database and keep a track of the existing files. These are the functionalities they allow:

- Generating a new binary file with the data contained in a text SGML file.
- Regenerating a binary file when there has been a change in the SGML.
- Getting the whole amount of recordings in the data-base.
- Changing the look of the web interface.
- Checking that every sound file has the related necessary files.
- Checking that the SGML code is right according to TEI recommendations.

6. Conclusions

The developed structure drives to an open system, where new data can be easily included, and new data types can be configured. All the data regarding the recordings is stored in TEI recommendations compatible SGML files, as well as in precompiled binary files that improve the system performance. The access system is web-based, allowing every user with a web browser and an MP3-capable sound system to browse the data-base, and hear the recordings.

This system has been successfully implemented, and the amount of recordings in the database is still increasing. Actually, the database keeps 16701 sound records, structured as follows:

- 292 popular literature recordings.
- 13049 isolated words.
- 818 texts.
- 2542 sentences.

Recordings have been made in most of the towns from Bizkaia where Basque is spoken, exactly in 79 towns from 8 different regions.

The implemented system is yet working and available at the following address: http://bizkaifon.ehu.es/ .

7. Acknowledgements

This project was funded by the Bizkaia's Regional Council under contract named "Fonoteca del Vizcaíno".

8. References

Etxebarria, B. and others. (2000). *Tools and Basque language databases developed in the AhoLab Laboratory*. In Workshop Proc. LREC May 2000. (pp. 62-70).

Listerri, J. (1997). *Transcripción, etiquetado y codificación de corpus orales*. http://liceu.uab.es/~joaquim/publicacions/FDS97.html

Real Academia Española (1999). *Transcripción y codificación de textos orales*.

Text Encoding Initiative (1994). *TEI Guidelines for Electronic Text Encoding and Interchange*. Electronic Text Centre at the University of Virginia