

Base de datos oral y textual para el Euskera

*Juan M^a Sánchez, Imanol Madariaga, Isaac Amezaga, Mikel Martín,
Eva Navas, Iñaki Gaminde, Inma Hernáez*



Índice

- ❑ Introducción
- ❑ Objetivos
- ❑ Usuarios potenciales y material sonoro
- ❑ Estructura de la base de datos
 - ☞ Estructura general
 - ☞ Formatos de archivos
 - ☞ Segmentación y etiquetado
- ❑ Sistema de consulta prototipo
- ❑ Conclusiones



Introducción

- ❑ Necesidad de material sonoro para aplicaciones multimedia
- ❑ Cantidad de grabaciones para distintos propósitos.
- ❑ Existen muchas variantes dialectales del euskera.
- ❑ Hay un euskera escrito estándar (Batua)
- ❑ Los dialectos distintos del batua pierden terreno.
- ❑ La enseñanza del euskera oral es difícil en estos casos.
- ❑ Se necesita una herramienta que unifique los materiales, adecuada para distintos usos.



Objetivos

- ❑ Recoger, preservar y clasificar material sonoro hablado.
- ❑ Implementar una estructura abierta para la integración sonora.
- ❑ Proveer de un espacio de difusión para las variedades dialectales vizcaínas.
- ❑ Herramienta para investigaciones lingüísticas.
- ❑ Sistema de consulta basado en web:
 - Acceso rápido a los documentos.
 - Independiente de hora y lugar.
 - Multimedia.



Usuarios Potenciales

- Estudiosos :
 - ↳ Lingüistas: dialectología, fonética, fonología, sintaxis,...
 - ↳ Sociólogos, etnólogos y antropólogos.
- Estudiantes:
 - ↳ De la cultura vasca.
 - ↳ Terminología técnica ????.
 - ↳ Estudiantes de Euskera.
- Profesionales de las tecnologías de la lengua.
- Público en general.



Clasificación del material sonoro

4 tipos de archivos:

- ❑ Palabras aisladas.
- ❑ Frases.
- ❑ Textos.
 - ↳ Etnográficos, cuentos, acontecidos.
- ❑ Literatura popular:
 - ↳ Canciones, adivinanzas, oraciones, versos.



Estructura de la base de datos

Información textual

- ❑ Cabecera: información genérica sobre la señal.
- ❑ Textos:
 - Transcripciones.
 - Alineamiento temporal.
- ❑ Enlaces al archivo sonoro.

```
<text> <u who=gizona> <vocal type="exklamazioa"
desc="ene"> </vocal> Karea bai soroatan bai usatzen
izan da <pause></pause> baina haziagaz ez nahastu
<pause></pause></u><u who=lñaki><vocal
type="afirmazioa" desc="um-hum"> </vocal> </u> <u
who=gizona>haziagaz nahastu barik karea
zabaldu<pause></pause>ekarri <long> harria eta
kare harria <pause> </pause></u> <u
who=lñaki><vocal type="afirmazioa" desc="um-
hum"></vocal></u><u who=gizona>eta urtu
<pause></pause>uregaz bota eta urtu <pause>
</pause> </u>...
```

Archivo sonoro



Formato de archivos sonoros

- Para almacenar: formato WAV.

- Se admite cualquier número de bits por muestra.
- Frecuencia de muestreo:

Al menos 16000 muestras/segundo.??????

- Para transmitir:

- WAV: se transmite la señal original.
- MP3: se reducen los tiempos de transmisión.



Formato de archivos textuales

- ❑ TEI SGML.
- ❑ 3 partes:
 - ☞ Cabecera: información descriptiva.
 - ☞ Textos: transcripciones y alineamiento temporal.
 - ☞ Enlaces:
 - A la señal original.
 - Relaciones entre elementos.



Contenido de la cabecera TEI

File Description

- Title Statement
 - Title
 - Author
 - Sponsor
 - Responsibility
- Publication Statement
 - Publisher
 - Distributor
 - Authority
 - Distribution Place
 - Id code
 - Publication date
- Source Description
 - Bibliographic source
 - Recording data
 - Script data

TEI Header Contents (II)

Profile Description

- Text Description
 - Domain
 - SubDomain
- Locutor Description
 - Name
 - Id
 - Sex
- Text Class
 - KeyWords
- Source Description
 - Source Place
 - Source Region



Textos: transcripciones

- ❑ Uno o varios niveles de transcripción:
 - ❑ Euskera batua (estándar)
 - ❑ Euskera vizcaíno
 - ❑ Otros: transcripciones fonéticas, ...
- ❑ Se acepta también otra información: gramatical, sintáctica...
- ❑ Marcas de eventos del habla dentro de las transcripciones.

Textos: Alineamiento temporal

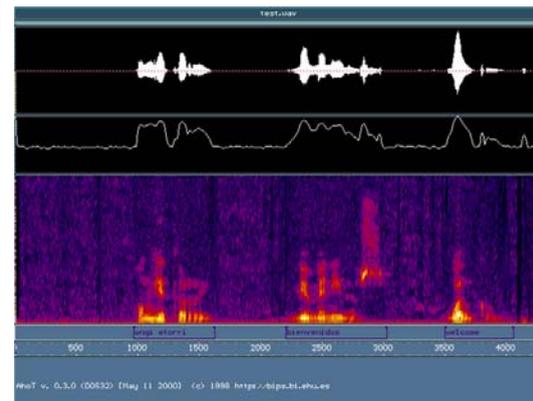
Alineamiento con TEI SGML

- ❑ La etiqueta **<timeline>** permite alineamiento temporal.
- ❑ Alineamiento temporal relativo.
- ❑ Cualquier evento puede ser referenciado en la **<timeline>** **<timeline>**.
- ❑ Cualquier transcripción puede ser referenciada.
- ❑ Hay que declarar la unidad de tiempo (normalmente ms.).
- ❑ La **<timeline>** timeline se enlaza al fichero de sonido.



Segmentación y etiquetado

- Trascripción
- Segmentación



AhoTools

- Etiquetado SGML:

Datu Orokorrak

Izenburua: Iñaki Gaminde-ren Artxiboa

Etiketatzaila: Isaac Amezaga

Argitaratzailea: Iñaki Gaminde

Copyright: Iñaki Gaminde

Testuen egilea: Iñaki Gaminde

Identifikadorea: IGA

Onartu Ezeztatu

Grabaketaren Datuak

Grabatzaila: Iñaki Gaminde

Grabaketaren Data: 1993-1994

Ekipamendua: Kasetea Modua: Irakurritako testua

Mota: Berba isolatuak Azpi mota:

Transkizioak: Batueraz Bizkaieraz Hjztuna

Atzera Onartu Ezeztatu

Parte hartzaile, lekua eta gaiak

Lekuaren Datuak

Eskualdea: Arratia

Herria: Dima

Parte hartzailearen Datuak

Izena: xxx

Identifikadorea (3 hizki): xxx

Sexua

Gizonezkoa Emakumezkoa

Badago parte hartzaile gehiagorik?

Bai Ez

Atzera Onartu Ezeztatu

Gaiaren Datuak: Hitz Gakoak

1

2

3

4

5

Conjunto de macros Visual Basic



Sistema de consulta prototipo

- Interfaz Web :

The screenshot displays a web browser window titled 'Datu Gehiago - Microsoft Internet Explorer'. The browser shows search results for 'Bizkaieraren Fonoteka'. The results are as follows:

Entzuten denaren traskribapena, batueraz:	nardaka
Entzuten denaren traskribapena, bizkaieraz:	nardakia
Grabaketaren egoera legala:	(C)Itzi Gaminde
Grabaketaren data:	1999-2000
Grabaketaren lekua:	Meñaka, Mungialdea
Grabaketa mota:	Berba isolatuak
Grabaketaren gaia eta azpigaia:	Gizakia, Bizilekua

Below the search results, there is a link labeled 'Entzun' and another link labeled 'Irten'.

The web interface also shows a section titled 'Emaitzak' with the following content:

Bilaketa zerbitzariak 3
1.(e)tik 3.(e)ra erakust

Testua	Mate
1 nardaka	Bert
2 nardaka	Bert
3 nardaka	Bert

The browser window also shows the date and time: 'Fri Jul 13 16:34:50 CEST 2001'.

In the bottom right corner, there is a Winamp player window showing the following information:

- Track: 1. SOINU SORTU[3] «0-01»
- Bitrate: 32 kbps
- Frequency: 16 kHz
- Format: mono estéreo

The Winamp player also shows a volume control slider and various playback controls.

<http://bips.bi.ehu.es/fonoteka>

Conclusiones

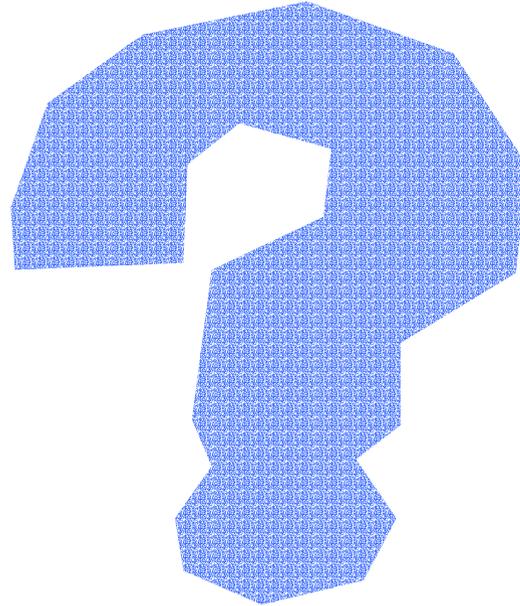
- Sistema abierto:
 - ↳ Permite introducir archivos grabados previamente.
 - ↳ Permite introducir datos marcados previamente.

- Material para investigaciones lingüísticas.

- Base para diferentes productos relacionados con voz:
 - ↳ Productos educativos.
 - ↳ Conversión de texto a voz, reconocimiento.



Preguntas?



■ Bizkaiko Foru
Aldundia
Diputación Foral
de Bizkaia



Este proyecto ha sido subvencionado por
la Diputación Foral de Bizkaia.