Base de datos oral y textual para el Euskera

Juan Mª Sánchez, Imanol Madariaga, Isaac Amezaga, Mikel Martín,

Eva Navas, Iñaki Gaminde, Inma Hernáez

http:\\bips.bi.ehu.es

Índice

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.

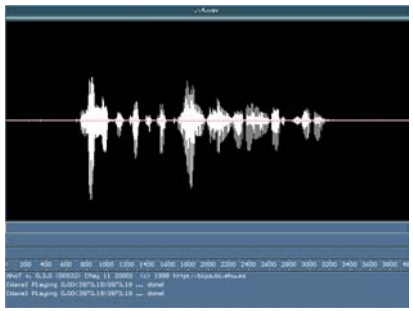
Estructura de la base de datos

Información textual →

<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=Iñaki><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=Iñaki><vocal type="afirmazioa" desc="um-hum"></vocal></u><u who=gizona>eta urtu <pause></pause>uregaz bota eta urtu <pause> </pause> </u>...

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

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
  - Responsability
- Publication Statement
  - Publisher
  - Distributor
  - Authority
  - Ditribution 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
- Source Description
  - Source Place
  - Source Region
- Text Class
  - KeyWords

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.

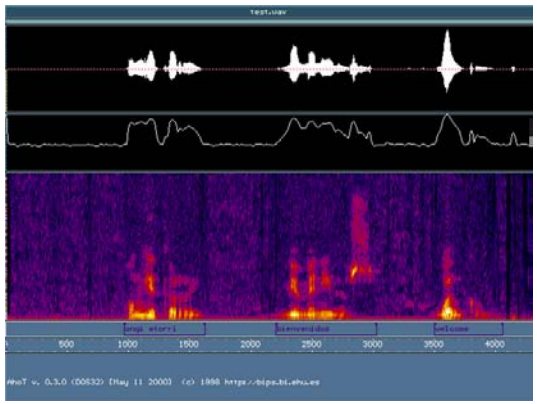
Textos: Alineamiento temporal

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.

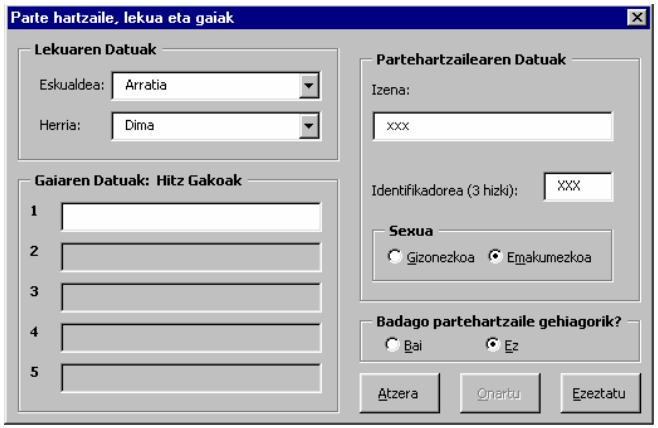
Segmentación y etiquetado

- Trascripción
- Segmentación



AhoTools

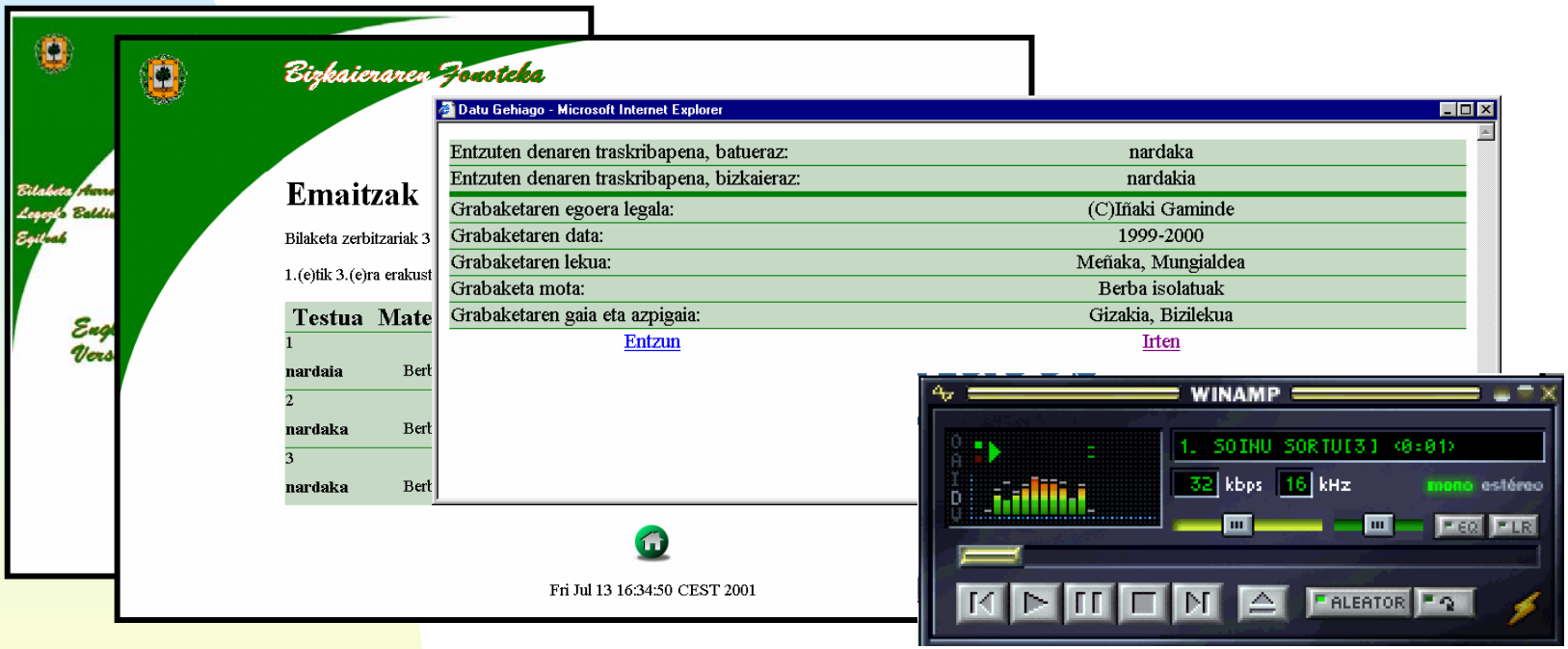
- Etiquetado SGML:



Conjunto de macros Visual Basic

Sistema de consulta prototipo

- Interfaz Web :

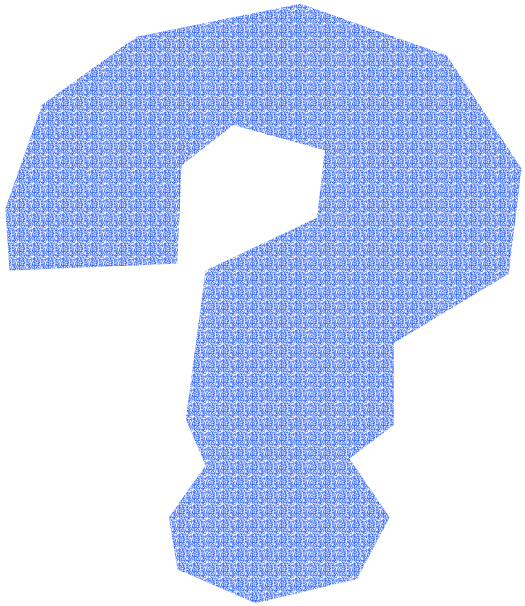


http://bips.bi.ehu.es/fonoteka

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.