IMPLEMENTACION DE BASE DE DATOS ORAL Y TEXTUAL PARA EL EUSKERA

Juan Mª Sanchez *Imanol Madariaga* *Xabier Zalbide* *Eva Navas* *Iñaki Gaminde* *Inma Hernáez*

Electrónica y Telecomunicaciones | Electrónica y Telecomunicaciones | Didáctica de la Lengua | Electrónica y Telecomunicaciones | Didáctica de la Lengua | Electrónica y Telecomunicaciones

Universidad del País Vasco/ Euskal Herriko Unibertsitatea

{ ion, imanol, xabier, eva, igaminde, inma}@bips.bi.ehu.es

ABSTRACT

This paper presents the implementation of a sound archive of dialectal varieties of spoken Basque. This database contains sound archives with their associated information and it is accessible via a web interface.

1. INTRODUCCIÓN

Durante años, se han venido realizando con distintos fines grabaciones de muy diversa índole de variedades dialectales del euskera. De esta amplia recopilación acústica, surge la necesidad de organizar y conservar este material considerado de alto valor sociolingüístico. El objetivo del trabajo global que aquí presentamos es el diseño y desarrollo de una base de datos que unifique los formatos de las grabaciones y datos asociados a las mismas, junto con el de un software de fácil acceso a ellos. La base de datos desarrollada facilita el estudio lingüístico, el desarrollo de tecnologías del habla, y establece una base sólida para la creación de aplicaciones para el aprendizaje del euskera. Los contenidos y características generales de este sistema se encuentran descritos en [1].

El sistema desarrollado se basa en tres subsistemas:

- Una base de datos, oral y textual, que alberga los archivos sonoros junto con información adicional.
- Un sistema de consulta que permite el acceso a los datos a través de Internet.
- Un software de administración del sistema, que permitirá la continua actualización de la base de datos.

El siguiente apartado describe la base de datos, y a continuación se describe la implementación del sistema (apartado 3) . El artículo termina con una breve discusión sobre el trabajo desarrollado.

2. ARQUITECTURA DE LA BASE DE DATOS

La base de datos se estructura de la siguiente manera: archivos sonoros ubicados en diferentes directorios, y, para cada uno, un archivo de texto en formato TEI SGML [2] y uno binario que contiene los mismos datos en un formato de acceso más rápido.

2.1. Archivos sonoros

Las grabaciones de voz se almacenan en archivos de audio en formato Windows PCM (wav). Todos los archivos sonoros usan muestras de 16 bits, y el sistema acepta diversas frecuencias de muestreo. Todas las grabaciones tienen un único canal (mono).

2.2. Archivos SGML

Estos archivos de texto almacenan, de acuerdo a la recomendación P4 de TEI [2] [3], los siguientes datos:

- Descripción técnica del fichero, con los siguientes datos:
  - Nombre o *path* del fichero de audio.
  - Fecha de actualización del fichero SGML
  - Título e ID de colección (cuando los datos que se utilizan provienen de una recopilación previa)
  - Título de fichero (en el caso de *literatura popular*)
  - Nombre de la persona responsable del etiquetado, del editor o editorial, e información sobre copyright
  - Tipo de material; si éste es *literatura popular* o *texto*, sub-tipo y palabras clave para identificar el tema.
  - Información de locutor: género y nombre si se conoce.
  - Lugar y región de grabación; fecha de ésta y equipamiento utilizado; persona responsable.
- Marcas de tiempo para cada nivel de marcado.
- Transcripción o transcripciones del archivo de audio

Se aceptan múltiples niveles de marcado y transcripciones.

2.3. Ficheros Binarios

El uso de ficheros SGML está muy extendido y resulta útil para realizar el etiquetado, pero su interpretación para realizar las búsquedas requiere mucha capacidad de proceso. Así, para realizar búsquedas eficientes en la base de datos, los mismos datos del archivo SGML se copian en ficheros binarios.

3. IMPLEMENTACIÓN DEL SISTEMA

Con el objetivo de conseguir un motor de búsqueda funcional, la base de datos y el software de búsqueda y gestión se ubican en una máquina Linux, haciendo accesible el interfaz de búsqueda a través de un servidor web Apache.

3.1. Aplicación de búsqueda

El interfaz web cumple tres funciones principales: generar, enviar e interpretar los formularios de búsqueda; generar una lista completa de resultados que cumplen los criterios de búsqueda; y recuperar todos los datos correspondientes de la base de datos, incluyendo el propio sonido, y enviarlo al navegador del usuario.

3.1.1. Formularios de búsqueda

Se han desarrollado dos formularios de búsqueda: el formulario de búsqueda estándar (figura 1), donde el propio sistema propone los criterios más adecuados para buscar tanto en las diferentes transcripciones como en los metadatos (lugar de grabación, palabra clave etc.), y todos los resultados deben cumplir todas las condiciones impuestas; el formulario de búsqueda avanzada, donde todos los datos SGML pueden utilizarse como criterio, pudiendo ser combinados usando operadores booleanos.

Figura 1. *Formulario de búsqueda estándar.*

3.1.2. Resultados de la búsqueda

Al enviar un formulario de búsqueda, el navegador cliente recibe como respuesta una página web que mostrará cada registro de la base de datos que cumple las condiciones de búsqueda. Para cada registro, se pueden obtener todos los datos del fichero SGML, incluyendo los metadatos, y las distintas transcripciones, así como oír las grabaciones. El sonido se envía codificado MP3.

3.2. Actualización de datos

Los datos se recopilan en dos pasos. En el primero, se selecciona la grabación y se segmenta y transcribe, utilizando algún software que permita observar características de la señal que útiles para crear marcas que almacenen la segmentación temporal y las transcripciones. Estas marcas y transcripciones se almacenan temporalmente en ficheros de texto. En el segundo paso, se crea un archivo SGML utilizando una aplicación diseñada al efecto, (figura 2) que en un entorno Windows, solicita los datos a incluir para el fichero (meta datos), comprobando que sean correctos, y recoge los ficheros de transcripciones del paso anterior. Por último, se crea el fichero binario a partir del SGML.

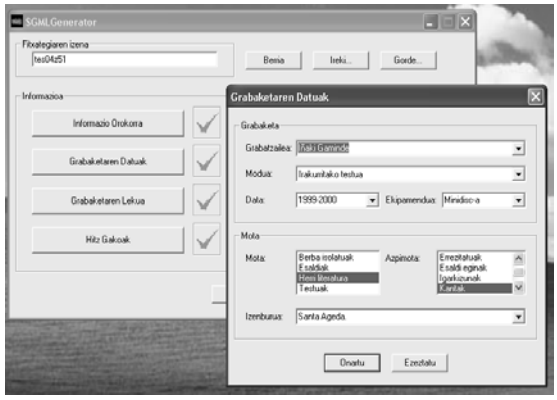

Figura 2. *Programa de generación de código SGML.*

3.3. Herramientas complementarias

Se han desarrollado herramientas basadas en Linux para el mantenimiento de la base de datos y la obtención de estadísticas. Algunas de las funciones que realizan estas herramientas son la regeneración de un archivo binario tras un cambio en el SGML, el cálculo del número de ítems en la base de datos, y chequeos de consistencia.

4. CONCLUSIONES

La estructura desarrollada conforma un sistema abierto, donde es fácil añadir nuevos datos, y se pueden configurar nuevos tipos de datos. Todos los datos sobre grabaciones se almacenan en archivos SGML compatibles con las recomendaciones TEI. El sistema de acceso está basado en web, permitiendo a cualquier usuario con un navegador y sistema de sonido compatible MP3 rastrear la base de datos y escuchar las grabaciones deseadas.

El sistema se ha implementado con éxito, de manera que actualmente la base de datos contiene 16701 grabaciones realizadas en 79 pueblos de 8 regiones distintas. El sistema está accesible en la dirección http://bizkaifon.ehu.es/.

5. AGRADECIMIENTOS

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

6. REFERENCIAS

[1] Sánchez, J. et. al. “Base de datos oral y textual para el euskera” ,URSI 2001 pp. 551-552

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

[3] Real Academia Española (1999). Transcripción y codificación de textos orales.