BASE DE DATOS MULTIMEDIA PARA EL EUSKERA

Jon Sánchez	Eva Navas	Imanol Madariaga	Amaia Castelruiz	Xabier Zalbide	Iñaki Gaminde
Electrónica y Telecom.	Electrónica y Telecom.	Electrónica y Telecom.	Electrónica y Telecom.	Didáctica de la Lengua y Literatura	Didáctica de la Lengua y Literatura

Universidad del País Vasco / Euskal Herriko Unibertsitatea

e-mail : ion,eva,imanol,amaia,xabier,igaminde@bips.bi.ehu.es

***Abstract*-This paper describes a multimedia archive of Biscayan, a dialect of Basque Language, involving audio files, video files and related information. It is based on an already developed sound archive. The application can be accesed using a web interface.**

I. INTRODUCCIÓN

La lengua vasca (el euskera) es una de las lenguas más antiguas de Europa. A pesar de tener un número reducido de hablantes, presenta una gran fragmentación dialectal. Desde las últimas décadas del sigo XX, se está llevando a cabo una estandarización de la lengua, lo que ha provocado también una progresiva desaparición de ciertas características dialectales.

Entre estos dialectos, el vizcaíno presenta a su vez una gran variedad de subdialectos. A partir de una recopilación de archivos de audio en distintas variedades de éste dialecto se desarrolló la Fonoteca del Vizcaíno [1][2][3]. Este sistema, actualmente en funcionamiento, permite realizar búsquedas on-line para acceder a las grabaciones y obtener información sobre las mismas.

Con el auge de las tecnologías multimedia, y la mejora en la capacidad de las conexiones de Internet, surge el interés de añadir archivos de vídeo a la base de datos existente, ampliando y mejorando así sus prestaciones. Los archivos de vídeo reflejan de forma más completa la compleja realidad del habla natural, ya que permiten obtener información paralingüística, como pueden ser los gestos del hablante, su lenguaje corporal, o las referencias al contexto en el que se produce el texto. Así, la información suministrada por los archivos de vídeo resulta más atractiva y completa.

Por tanto, se busca un sistema que permita realizar búsquedas sobre archivos de vídeo o audio indistintamente. Uno de los objetivos del proyecto ha sido mantener la compatibilidad con la Fonoteca del Vizcaíno, para integrar en un solo sistema ambos servicios.

El documento se divide en tres partes. En primer lugar se explica el diseño de la base de datos multimedia en sí. A continuación, se presenta el nuevo proceso de generación del material. Por último, se plantean los objetivos del sistema de acceso vía web.

II. DISEÑO DE LA BASE DE DATOS

La base de datos contiene archivos de dos tipos. Por un lado, el material multimedia, que consta de archivos de vídeo y audio. Por otro, los archivos de información relativa a dicho material, o metadatos. Cada archivo de audio o vídeo tiene asociados dos ficheros de información, uno en formato SGML y otro en formato binario.

Estos archivos se organizan en un árbol de directorios, tal y como se ve en la Figura 1. Éste se divide en dos ramas principales: una para los archivos multimedia y otra para los archivos de metadatos. Estas ramas tienen una estructura interna idéntica, organizada según los pueblos en los que se han realizado las grabaciones.

A. Material multimedia

Los archivos de audio se almacenan en formato Windows PCM (*wav*), utilizando muestras de 16 bits, y un único canal.

Los archivos de vídeo se almacenan en formato MPEG. Se establece para ellos una duración máxima de unos cinco minutos, para poder realizar su envío a través de Internet en un tiempo razonable.

B. Archivos de información

1. Archivos SGML

Los archivos SGML almacenan información variada sobre los archivos de audio [2][3] y vídeo, siguiendo la recomendación P4 de TEI [4].

2. Archivos binarios

Los archivos binarios se forman extrayendo la información más relevante de los SGML (aquella que se pueda utilizar como parámetro de búsqueda), y almacenándola en formato binario. De esta manera se consiguen búsquedas más rápidas que procesando el texto de los archivos SGML.

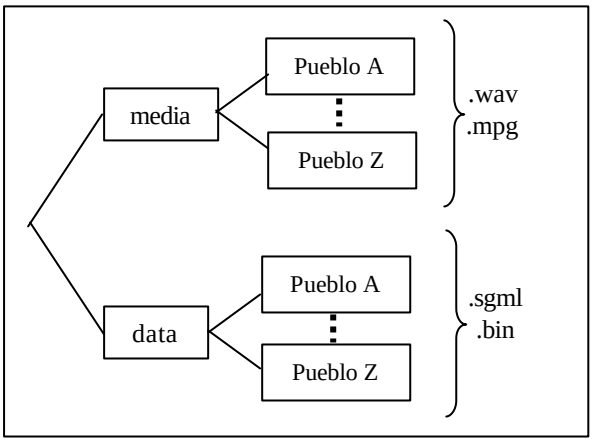

Fig. 1. Estructura de directorios de la base de datos.

III. PROCESO DE GENERACIÓN DEL MATERIAL

Para generar el material de la base de datos se siguen los siguientes pasos:

- Grabación de los archivos multimedia, y anotación del material grabado
- Generación de los archivos TEI SGML y binario

El procedimiento a seguir para la introducción de objetos de audio en la base de datos ha sido descrito con detalle en [3]. Por lo tanto, la explicación se centrará en el proceso de introducción de archivos de vídeo, desde la grabación hasta la generación de los metadatos.

A. Archivos de vídeo

Para la recogida de los datos se plantean sesiones de grabación continua de una hora de duración. El equipamiento a utilizar es el siguiente:

- Una videocámara digital: recoge vídeo y audio digitales.
- Un Minidisc: recoge solamente audio digital, con mayor calidad que la videocámara. Es recomendable utilizarlo, pero no siempre será posible.

Mediante la conexión de la videocámara a un PC, se vuelcan estas grabaciones de vídeo a archivos en formato AVI. De éstos, se seleccionan las secuencias más interesantes, utilizando criterios de prioridad lingüísticos, antropológicos y didácticos. Llamamos *clips* a estas secuencias seleccionadas.

Cuando ha sido posible realizar una grabación en Minidisc se sustituye el audio original, de menor calidad. Para ello se ha utilizado el editor de vídeo *Pinnacle Studio 8*.

Los segmentos de vídeo seleccionados son cortados y almacenados, en formato MPEG-1. Junto a cada archivo de vídeo se almacena también otro archivo con el audio correspondiente. Cuando no se disponga de la grabación del Minidisc, se extraerá este audio de la grabación de vídeo. Para dicha extracción se utilizan los programas *GraphEdit* [5] y *WinAmp* [6], ambos de libre distribución.

A continuación se realiza el marcado y etiquetado de los *clips*, utilizando el programa de libre distribución *ELAN*, del *Max Planck Institute of Psicholinguistics* [7][8]. Mediante este proceso se establece la correspondencia temporal entre los *clips* y su transcripción. Como resultado se generan unos archivos en formato XML (archivos *eaf*).

B. Archivos SGML

Los archivos SGML se generan con un editor SGML propio. Este programa, realizado en entorno Windows, solicita al usuario los distintos datos que precisa para generar el archivo.

En estos ficheros SGML es necesario considerar el tipo de datos que contiene el archivo multimedia correspondiente. Según sea vídeo o audio, el código SGML generado presentará diferencias en su contenido, siempre siguiendo la recomendación P4 de TEI [4].

En el caso de que el archivo sea de vídeo, es necesario extraer la información de marcado y etiquetado, así como las transcripciones, de los archivos *eaf XML* antes mencionados. Una vez obtenida esta información, se introduce en el archivo TEI SGML con el formato correspondiente.

C. Archivos binarios

Como se ha indicado anteriormente, los archivos binarios contienen los parámetros más relevantes de los archivos SGML (los que se van a utilizar para realizar las búsquedas), almacenados en formato binario.

En estos archivos también debe constar el dato referente al tipo de grabación (audio o vídeo), mencionado en el apartado anterior.

IV. DISEÑO DEL SISTEMA DE ACCESO VÍA WEB

La aplicación de acceso a la base de datos vía web tiene dos partes diferenciadas. En primer lugar, el formulario de búsqueda, en el que el usuario introduce los parámetros en base a los que desea que se realice la búsqueda de material. En segundo lugar, la página de resultados, en la que se presenta un listado de los archivos que coinciden con las condiciones de búsqueda impuestas por el usuario.

En el diseño del sistema de acceso a la base de datos multimedia se han planteado los siguientes objetivos.

Realización de la búsqueda:

- Se debe permitir la búsqueda de material según el tipo de archivo. Así, se deben poder obtener:
  - Sólo archivos de audio
  - Sólo archivos de vídeo
  - Ambos tipos de archivos

Resultados de la búsqueda:

- Búsqueda de archivos de audio. Deben aparecer, como resultados de la búsqueda:
  - Los archivos de audio en sí
  - Los archivos que contienen el audio correspondiente a los vídeos

- Búsqueda de archivos de vídeo: Junto a cada archivo de vídeo, se proporcionará acceso al archivo que contiene sólo el audio.
- Cuando el resultado de la búsqueda incluya archivos de vídeo, se indicará al usuario que la descarga del mismo puede llevar cierto tiempo. También se recomendará que escuche el audio correspondiente a dicho vídeo antes de descargarlo, para asegurarse de que el archivo es realmente de su interés.

Fig. 2. Ejemplo de página de resultados de la búsqueda, mostrando resultados de audio y vídeo.

V. CONCLUSIONES

La base de datos desarrollada, junto con los interfaces de usuario que la complementan, permiten acceder a material de gran valor lingüístico, etnográfico y cultural, tanto de vídeo como de audio. Actualmente la base de datos contiene todas las grabaciones de audio de la Fonoteca del Vizcaíno realizadas en 78 localidades, y 320 *clips* de vídeo obtenidos en 40 de ellas. Por otra parte, el sistema esta aun en desarrollo, y se sigue enriqueciendo añadiendo nuevos materiales y creando una base de datos cada vez más completa.

AGRADECIMIENTOS

Este proyecto ha sido financiado por la Diputación Foral de Vizcaya - Bizkaiko Foru Aldundia.

REFERENCIAS

[1] J. Sánchez, et. al.,"Base de datos oral y textual para el euskera", URSI 2001 pp. 551-552
[2] J. Sánchez, et. al., "Implementación de base de datos oral y textual para el euskera", URSI 2002 pp. 485-486
[3] I. Hernáez, et. al., "BIZKAIFON: A sound archive of dialectal varieties of spoken Basque", LREC 2002 pp. 865-866
[4] Text Encoding Initiative (1994). TEI Guidelines for Electronic Text Encoding and Interchange. Electronic Text Centre at the University of Virginia.
[5] http://www.digital-digest.com/dvd/downloads/graphedit.html
[6] http://www.winamp.com
[7] http://www.mpi.nl
[8] H. Brugman, et. al., "Multimedia Annotation witth Multilingual Input Methods and Search Support" LREC 2002 pp. 378-383