BASE DE DATOS ORAL Y TEXTUAL PARA EL EUSKERA

Juan Mª Sánchez Electrónica y Telecom. UPV/EHU
Imanol adariaga Electrónica y Telecom. UPV/EHU
Isaac Amézaga Electrónica y Telecom. UPV/EHU
Mikel Martín Electrónica y Telecom. UPV/EHU
Eva Navas Electrónica y Telecom. UPV/EHU
Iñaki Gaminde Didáctica de la Lengua UPV/EHU
Inma Hernáez Electrónica y Telecom. UPV/EHU

ion, imanol, isaac, mikel, eva, igaminde, inma@bips.bi.ehu.es

RESUMEN

En este artículo presentamos la estructura y contenido de una base de datos oral en desarrollo para el euskera. La base de datos contiene además de archivos sonoros con grabaciones de muy diferentes tipos y procedencias, las transcripciones correspondientes e información diversa relativa a las mismas, lo cual permite clasificar y por tanto acceder selectivamente al material disponible. El desarrollo de este sistema permitirá el acceso a un gran número de grabaciones, hasta ahora difícil debido a la falta de unificación de formatos de señales y transcripciones asociadas.

1. INTRODUCCIÓN

La disponibilidad de material sonoro para una lengua está adquiriendo en los últimos años gran importancia para el desarrollo de aplicaciones multimedia así como para el desarrollo de las tecnologías del habla. Para la lengua vasca (el euskera) se dispone de una gran cantidad de material recopilado a lo largo de las últimas décadas con fines muy diversos y por tanto potencialmente útil para muy diversas aplicaciones. Sin embargo, debido a la gran variedad del material existente y a la falta de uniformidad en los formatos la mayor parte del material no está accesible para un gran número de potenciales usuarios.

Por otro lado, el euskera tiene la particularidad de encontrarse muy fragmentado dialectalmente. La implantación y desarrollo del euskera estándar (euskera *batua*) ha influido en la evolución de estas variedades, en algunos casos negativamente, en especial en aquellas lingüísticamente más distantes, como es el caso de un gran número de las variedades vizcaínas. Además, el euskera batua no tiene correspondencia directa con ninguna de las variedades dialectales del euskera hablado, lo cual dificulta la enseñanza y aprendizaje de los aspectos orales de la lengua (la pronunciación, el acento, la entonación...).

Considerando que existen grabaciones de muy diversa índole de un gran número de estas variedades, surge la necesidad de organizar y conservar este material que puede considerarse de alto valor socio-lingüístico. El diseño y desarrollo de una base de datos que unifique los formatos de las grabaciones y datos asociados a las mismas, y un software de acceso selectivo a ellos es el objetivo global del trabajo que aquí presentamos.

El siguiente apartado describe las características más generales del sistema en desarrollo, así como los beneficios que pensamos que aportará a la sociedad. El apartado 3 y el 4 describen los aspectos técnicos del sistema.

2. DESCRIPCIÓN DE LA FONOTECA

El proyecto denominado *Fonoteca del Vizcaíno* comprende el desarrollo de tres subsistemas:

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

2.1. Usuarios potenciales

Uno de los propósitos de esta fonoteca es hacer accesible el material sonoro existente a un gran espectro de usuarios. Se han contemplado diversos tipos de usuarios en la definición del sistema, que se verán directamente beneficiados por la fonoteca:

- Los estudiosos lingüistas en sus diversas ramas (estudiosos de la fonética, fonología, sintaxis, morfología, lexicografía) así como dialectólogos y sociolingüistas serán los más directamente beneficiados. Por el carácter cultural de ciertos contenidos de la fonoteca, también etnólogos y antropólogos podrán encontrar en ella material de interés.
- Profesores y estudiantes, desde los niveles más bajos hasta la universidad, ya que les permitirá conocer los diferentes registros de la lengua y la cultura de los pueblos de Vizcaya.
- Profesionales de las tecnologías de la lengua: informáticos e ingenieros encontrarán en la fonoteca material para desarrollar sus investigaciones en los campos de la síntesis y el reconocimiento de la voz y del tratamiento automático de textos.

2.2. Clasificación del material sonoro

Aunque la fonoteca será capaz de albergar cualquier tipo de material, el objetivo de este proyecto en este aspecto se limita a la recopilación de variedades dialectales del vizcaíno, y con particular interés en la organización, clasificación y etiquetado de grabaciones ya existentes. En concreto, se dispone de al menos (ya que resta aún mucho material por inspeccionar) 100 horas de

material sonoro procedente de 85 localidades de 8 variedades dialectales vizcaínas consideradas. Utilizando este material disponible como punto de partida, se han considerado tres tipos básicos de material sonoro:

- Palabras pronunciadas de forma aislada, que fueron recogidas en su momento para analizar algún aspecto lingüístico de la variedad dialectal, como puede ser lexicones, estudios sobre el acento u otros. Por lo general, este material se encontraba ya digitalizado y transcrito (aunque sin asociación temporal entre señales y transcripciones) en la mayoría de los casos.
- Textos, que de acuerdo con su contenido, han sido a su vez clasificados en *etnográficos* (ofrecen información sobre las formas de vida local); *lingüísticos* (frases recopiladas para analizar la sintaxis, la entonación...); *narraciones* tales como chistes, sucesos; y finalmente *cuentos*. Una pequeña parte de este material se encuentra digitalizado y transcrito.
- Literatura popular, que incluye cantos, versos, adivinanzas, rezos, juegos y otras manifestaciones literarias populares.

Además, se considera una clasificación jerárquica a dos niveles del tema tratado en la grabación, y del lugar en donde se realizó la grabación (coincidente con la variedad dialectal en el caso de grabaciones de campo).

3. ESTRUCTURA DE LA BASE DE DATOS

3.1. Estructura General

La base de datos almacenará además de los archivos sonoros toda la información disponible sobre ellos. Esta información es de dos tipos:

- Información de *cabecera* o genérica sobre la grabación: características del locutor, fecha de la grabación, lugar (que definirá la variedad dialectal empleada), tipo de material de que se trata según la clasificación descrita en el apartado 2, procedencia de la grabación información sobre derechos de autor, etc.
- Información de *etiquetado* asociada a la señal: etiquetas de texto asociadas a un segmento de señal. Las etiquetas pueden corresponderse con transcripciones (ortográfica, fonética) o cualquier otro tipo de etiquetado, como pueda ser sintáctico, de entonación...

Actualmente se está incluyendo la transcripción ortográfica en euskera batua (salvo en el caso de la literatura popular) y la transcripción ortográfica literal correspondiente al dialecto hablado, mucho más próxima a la pronunciación realizada que la transcripción en euskera estándar.

Así, las señales podrán encontrarse segmentadas y etiquetadas a diferentes niveles. El sistema contempla la posibilidad de añadir tantos niveles de etiquetado como se desee.

3.2. Formatos

Para las señales se ha elegido el formato *wav* para su almacenamiento (una vez seleccionado el segmento de señal requerido por el usuario, se procede opcionalmente a su conversión a formato *mp3* para reducir los tiempos de transmisión).

En lo que se refiere a la información general de cabecera y etiquetado, se ha elegido el sistema estándar de etiquetado SGML siguiendo la recomendación TEI [1], y además, fijándonos en los trabajos realizados en [2] y [3]. Ello nos permitirá disponer de un corpus etiquetado y segmentado de aplicación general, que podrá encontrar múltiples usos además de la aplicación actualmente en desarrollo de la fonoteca. Por cada fichero de audio, existe un fichero de tipo SGML. Este fichero contiene tanto la información general (cabecera) como las etiquetas. Para las particularidades de nuestra fonoteca, se han definido algunas etiquetas nuevas, como por ejemplo, una etiqueta que permite definir los niveles de segmentación y etiquetado disponibles para ese fichero. Además, para facilitar la labor de etiquetado SGML, se ha desarrollado un conjunto de macros de Visual Basic. La segmentación de las señales y la parte del etiquetado de señales síncrona con el tiempo se realiza con la ayuda de un editor de señales propio (disponible en http://bips.bi.ehu.es)[4].

4. SISTEMA DE CONSULTA PROTOTIPO

Actualmente existe un sistema prototipo para consultar la base de datos, la cual contiene únicamente una pequeña muestra de los datos que en un futuro albergará la fonoteca. En este sistema, accesible en http://bips.bi.ehu.es/fonoteka, la búsqueda se realiza únicamente en el campo de la transcripción ortográfica, permitiendo realizar el filtrado de los resultados según valores asignados a los campos correspondientes a las cabeceras de los ficheros.

5. AGRADECIMIENTOS

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

6. REFERENCIAS

[1] Text Encoding Initiative: “TEI Guidelines for Electronic Text Encoding and Interchange (P3)”: Electronic Text Center at the University of Virginia, 1994

[2] Listerri, J.: “Transcripcción, etiquetado y codificación de corpus orales ” http://liceu.uab.es/~joaquim/publicacions/FDS97.html, 1997

[3] Real Academia Española: “Transcripción y codificación de textos orales“, 1999

[4] B. Etxebarria, y otros: “Tools and Basque language databases developed in the AhoLab Laboratory” Workshop Proc. LREC May 2000 pp. 62-70