Capítulo III

El corpus de voz

Un corpus es una colección de grabaciones de voz, las cuales están accesibles en forma de lectura dentro de una computadora y las cuales vienen con anotaciones y con documentación suficiente para permitir el reuso de los datos. Los corpus de voz son siempre diseñados para propósitos específicos. Estos propósitos determinan el contenido y el diseño de un corpus. Por esta razón las características de los corpus pueden variar mucho dependiendo para que fue creado.

3.1 Tlatoa Common Questions Corpus

El corpus *Tlatoa Common Questions Corpus* es una base de datos de voz grabada por teléfono de personas que hablan español mexicano. Este corpus es un esfuerzo de colectar grabaciones del español hablado en México en varias regiones del país. Este corpus ha sido diseñado para proveer datos útiles en el desarrollo y evaluación de sistemas de reconocimiento de voz.

Hasta la fecha, el corpus *Tlatoa Common Questions* consiste de voz grabada por teléfono por más de 400 locutores. Ha sido diseñada para cubrir adecuadamente vocabularios comunes como lo son dígitos, números naturales, si/no, horas, días, meses, fechas y nombres y apellidos, etc.. Además se ha incluido en el protocolo una pregunta donde la persona graba 30 segundos de habla espontánea o *stories*.

3.2 Condiciones de grabación:

Las grabaciones por teléfono se realizan a una frecuencia de muestreo de 8000 Hz y los archivos son almacenados en formato RIFF. El hardware que se utilizó consiste en una computadora PC *Gateway* 2000 con procesador pentium, sistema operativo *Windows* NT y una tarjeta de teléfono Dialogic.

Para realizar las grabaciones, se implementó un sistema de recolección de voz o captura usando el CSLU *toolkit*. Este sistema es el encargado de grabar las pronunciaciones de los locutores que llaman al sistema. Entre sus funciones están la detección de la llamada, su atención y descolgado. Además, el sistema guarda en una bitácora información del transcurso de la llamada una vez que esta haya finalizado, suministrando información útil para su posterior análisis. Este análisis incluye saber el número de repeticiones de preguntas por ausencia de respuesta o por adelantamiento del locutor.

¡Gracias por participar!

Nuestro grupo de investigación está recolectando voces para estudiar las características fonéticas del español que se habla en nuestro país. Le recordamos que la información proporcionada será usada únicamente para efectos de investigación. Al finalizar con éxito la grabación se le asignará un número de participante para la gran rifa.

1. ¿Está ud. listo para comenzar la sesión de preguntas?
2. ¿Cuál es su nombre completo?
3. ¿Cuál es su edad?
4. ¿Cuál es su número telefónico?
5. Diga alguna dirección que le sea familiar.
6. ¿Cuál es el código postal de la zona en donde vive?

7. ¿En qué ciudad reside actualmente?
8. ¿Cuántos hermanos tiene?
9. Diga el nombre de algún familiar cercano.
10. Diga el nombre de la ciudad en donde nació.
11. Diga cualquier otro número telefónico que le sea familiar.
12. ¿Cuántos años tiene la persona de más edad que conoce?
13. ¿Qué hora es en este momento?
14. ¿Es usted estudiante de la Universidad de las Américas
    IF YES ----> 15. ¿Cuál es su número de estudiante?
16. Diga el nombre de dos ciudades o estados de la República Mexicana que le gustaría visitar.
17. ¿Cuenta ud. Con algún automóvil?
    IF YES ----->17. Diga cualquier número de placa.
19. Ahora le pedimos que hable durante medio minuto sin parar.

Gracias por haber participado. Su número para el sorteo es _____.

Ejemplo 3.1 Preguntas que contiene el sistema.

3.3 División del corpus

Como ya se mencionó anteriormente es necesario dividir el corpus para entrenamiento, desarrollo y prueba. Para entrenamiento se asignaron 270 locutores respondiendo cada una de las preguntas presentadas en el ejemplo 3.1, excepto la número 19 que son los *stories.* Así mismo, para la partición de desarrollo se reservaron 70 locutores, exceptuando también la frase 19. Finalmente, la partición de prueba contaba con 65 *stories,* ningún locutor aparece en dos diferentes subconjuntos. Cabe señalar que los *stories* son habla difícil de reconocer debido a su espontaneidad, las variaciones en el ritmo de hablar del locutor, entonación, y a la cantidad de disfluencias contenida en ellos, por ejemplo, em, ah esteeee, etc..

3.4 Protocolo de etiquetado

El proceso de transcripción de los archivos de voz se realiza a tres niveles: a nivel ortográfico, a nivel de palabras alineadas con respecto al tiempo y, por último, a nivel de fonemas alineados con respecto al tiempo. Con las convenciones de etiquetado podemos tener mayor información de cada pronunciación.

3.4.1 Transcripciones ortográficas

Las etiquetas a nivel texto nos indican el contenido de una pronunciación sin tiempo de referencia; es decir, son una transcripción textual de lo que se dijo. Se representan con la ortografía. Por lo regular el contenido de la transcripciones textuales es como el de cualquier texto pero sin puntuación, ni distinción entre letras mayúsculas y minúsculas.

Una herramienta contenida en el *toolkit* que permite hacer transcripciones textuales de una manera muy sencilla es un *script* llamado check_txt_files.tk que revisa los archivos de sonido, los reproduce y nos permite teclear el texto correspondiente.

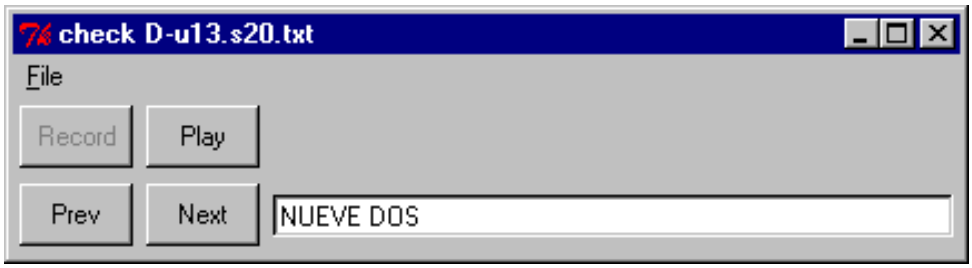

Figura 3.1 Transcripción ortográficas

3.4.2 Transcripciones a nivel palabras

La meta de esta etapa es alinear las transcripciones ortográficas en el tiempo. Además se hace una distinción entre lo que es y lo que no es voz.

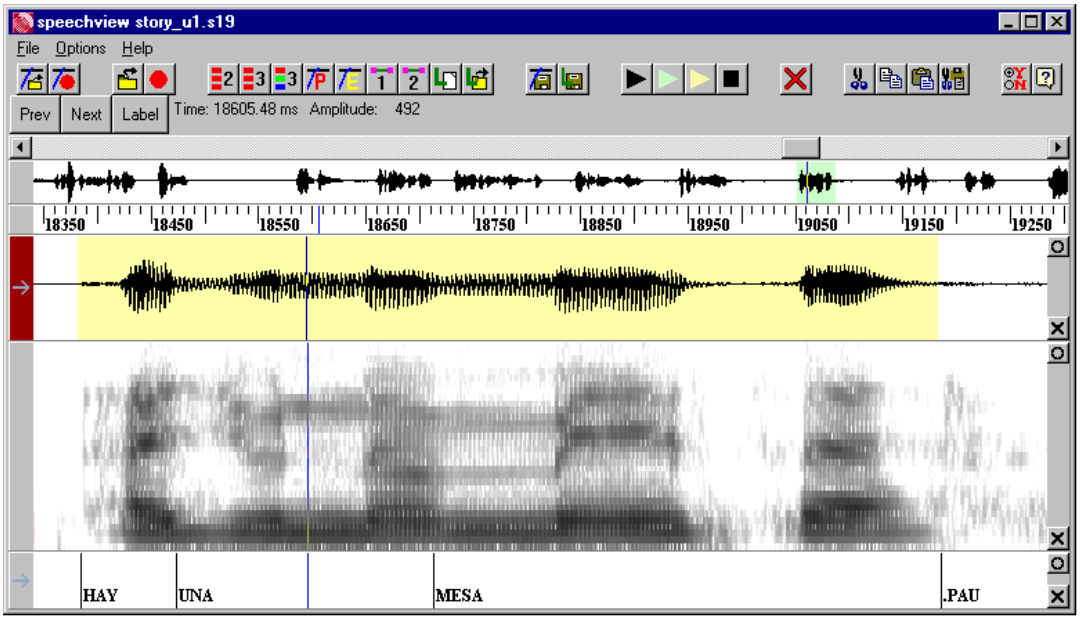

Figura 3.2 Transcripción a nivel palabras

3.4.3 Transcripciones fonéticas

Las transcripciones fonéticas representan el contenido de una pronunciación en cierto nivel de detalle. Como referencia básica del conjunto de fonemas se usó el *Worldbet*, el cual define un conjunto de símbolos fonéticos ASCII que intentan representar todos los fonemas en cualquier idioma en el mundo. A partir de las etiquetas a nivel palabra alineadas con el tiempo se pueden obtener los símbolos fonéticos correspondientes a cada palabra, utilizando un *script*. Por último se alinean los símbolos fonéticos adecuadamente utilizando un proceso manual.

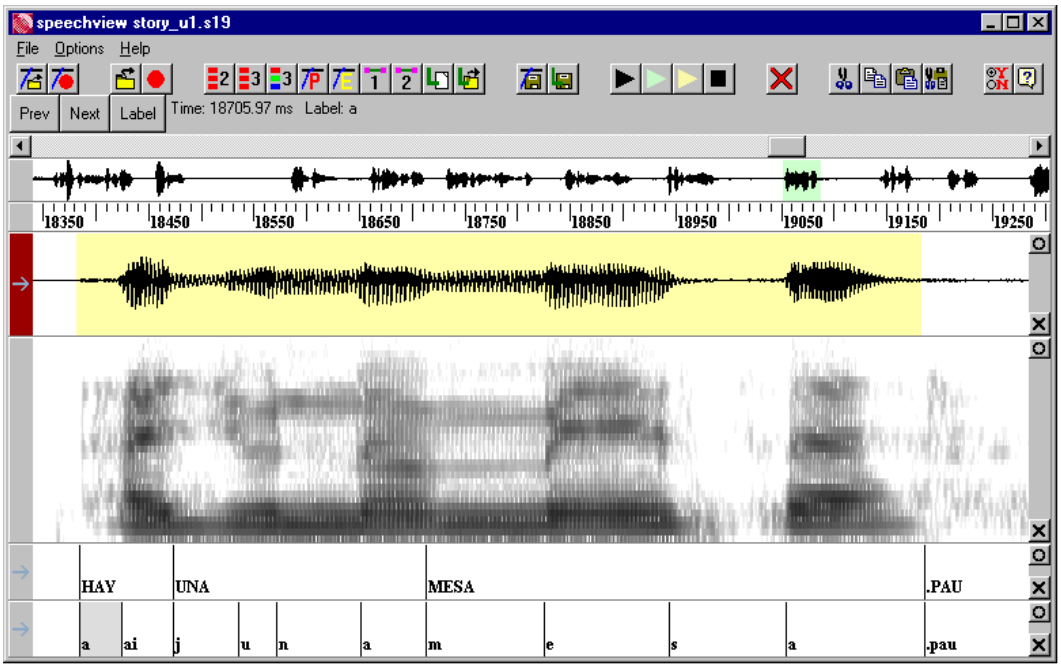

Figura 3.3 Transcripción a nivel fonemas

3.4.4 Conjunto de símbolos

En esta sección se presentan las unidades que se usan para etiquetar el corpus. Estas comprenden tres conjunto de símbolos: los usados para definir los sonidos que pertenecen al lenguaje, aquellos usados para definir sonidos que están fuera del lenguaje y por último, el uso de diacríticos que dan información detallada.

La mayoría de los diacríticos se refieren a un conjunto de sonidos. Por ejemplo, [bn] puede ser música, ruido de personas que hablan o el ruido de la televisión encendida.

Diacrítico	Tipo de diacrítico
_h	Aspirado
_x	Fonema muy corto
_fp	Fonema muy largo
_ln	Ruido en la línea
_bn	Ruido de fondo

Tabla 3.1 Ejemplo de algunos diacríticos usados en el etiquetado para dar información detallada

Existe otro tipo de etiquetas que identifican partes de la señal que no son voz, a las cuales les asociamos un símbolo. Estas etiquetas se usan tanto en transcripciones a nivel palabra como a nivel fonéticas y están precedidas por un punto.

Descripción	Símbolo
Pausas entre palabras	.pau
Respiraciones	.br
Ruido de fondo en la grabación	.bn
Ruido en la línea	.ln
Elementos de la señal no identificables	.unk

Tabla 3.2 Etiquetas que representan la parte de la señal sin voz.

3.5 Formato y estructura de los archivos.

Cada frase grabada se almacenó en la computadora en cuatro formatos específicos. El formato de los archivos de voz es .wav, el de las transcripciones a nivel de texto .txt, a nivel de palabra .wrd y a nivel de fonemas .phn. Estos formatos fueron elegidos siguiendo las especificaciones establecidas en la documentación en línea del CSLU. Los símbolos utilizados para las etiquetas corresponden al alfabeto *Worldbet.*