El ingeniero Jorge Gurlekian en investigador del CONICET y es coordinador del área de Audición y Habla del Laboratorio de Investigaciones Sensoriales (LIS).

El LIS es un centro dependiente del CONICET y está ubicado en la Escuela de Salud Pública, Facultad de Medicina, UBA.

Este trabajo se ha llevado a cabo en el marco de un proyecto Pict98, dirigido por el ingeniero Jorge Gurlekian, aprobado por la SECYT para el período 1999-2000 (Nro. 11-04177).

*Desde los tiempos de Bacon de Verulam, de los filósofos perfeccionistas del siglo XVIII y de la fe en el progreso del XIX, el mejoramiento humano gracias a la ciencia y la tecnología se tuvo por cosa segura. Aun antes, Platón creyó que se podría encontrar solución a los problemas humanos si los filósofos reinaran o los reyes se trocaran en filósofos. El descubrimiento desalentador de nuestros días ha sido que la razón no conduce automáticamente a la solución de los problemas humanos y que, en general, influye escasamente en el sangriento curso de la historia del hombre. El conocimiento es poder, según Bacon; pero ¿ha aumentado el poderío tecnológico la felicidad humana? ¿Es deseable una sociedad regida científicamente?*

Ludwig von Bertalanffy

# ÍNDICE

# ÍNDICE DE ILUSTRACIONES

# Capítulo 1 INTRODUCCIÓN

Cuando se lee un texto en voz alta se realiza un proceso de conversión de texto a habla.

Los actuales sistemas de conversión de texto a habla necesitan la ayuda de una base de datos para alcanzar la alta calidad. Aquí se construye una base de datos prosódica que contiene información del español hablado en Buenos Aires y sirve para entrenar a un conversor adecuado a nuestra lengua.

El presente trabajo se encuadra dentro del área del conocimiento conocido como Tecnologías del Habla que recibe aportes de la lingüística, la ingeniería, la informática, la física acústica, la fisiología y la psicología. En este marco estudia los sistemas de conversión de texto a habla, para cuyo desarrollo la lingüística brinda los modelos de entonación y los métodos para poder interpretarlos, la ingeniería se ocupa de proporcionar los modelos de producción de habla y los métodos de procesamiento de señales, y es la informática quien brinda los modelos de software junto a los métodos apropiados para su implementación.

En particular se ocupa de las bases de datos que dichos conversores usan como entrenamiento para alcanzar alta calidad en el habla que producen. En los últimos años, el desarrollo de los sistemas de conversión de texto a habla se ha convertido en una tarea interdisciplinaria, con intervención necesaria y fundamental de áreas de la lingüística, la ingeniería y la informática.

Está dirigido a toda persona especializada en el tema de los sistemas de conversión de texto a habla. Sin embargo, algunas partes de los capítulos principales pueden ser de interés más general debido al auge de las computadoras y las tecnologías del habla específicamente.

Surgió a partir de un proyecto de desarrollo de un sistema de conversión de texto a habla, de alta calidad, para el español hablado en Buenos Aires, a realizarse en el Laboratorio de Investigaciones Sensoriales, dependiente del CONICET. En dicho proyecto se propone construir una base de datos prosódica como etapa necesaria en el desarrollo de un sistema conversor de texto a habla. Se propuso que sea el presente trabajo el ámbito de construcción de una base de datos prosódica que cumpla los requisitos del proyecto citado. Una vez construida la base de datos comienza una etapa de entrenamiento y aprendizaje en la que se desarrollan las estrategias apropiadas para generar contornos prosódicos. Dichas estrategias son usadas en la etapa final de desarrollo y puesta a punto del conversor.

Esta base de datos requirió durante su desarrollo la creación de un corpus de oraciones que contuviese las sílabas más representativas del español de Buenos

Aires. Dicha tarea demandó una fuerte inversión, incluido el desarrollo de una aplicación para realizar transcripciones ortográficas y fonológicas en forma automática y una base de datos específica para depurar oraciones. Todo esto ayudó a definir algunas de las estructuras de datos y métodos.

Además de ser un pilar muy importante en el desarrollo de un sistema conversor de texto a habla para el español de Buenos Aires, en la actualidad existen pocos desarrollos soportados por sistemas de gestión de bases de datos. Es una herramienta muy útil para aquellos que se dedican a la investigación de la entonación. Pretende ayudar a encontrar nuevas formas de interpretación prosódica, así como a aclarar algunos conceptos actualmente en estudio, propios de nuestra lengua.

Desde una visión más general, estas bases de datos, también llamadas "corpus de habla", se usan tanto para tareas de investigación como para el desarrollo de aplicaciones tecnológicas. Entre las primeras podemos citar: investigación fonética, sociolingüística, psicolingüística, lingüística en general, audiología, patologías del habla y para la adquisición de la lengua primaria, secundaria, etc. Entre las aplicaciones tecnológicas, además de los sistemas de conversión de texto a habla, encontramos: síntesis del habla, análisis del habla, reconocimiento del habla, sistemas de interacción oral (utilizan reconocimiento y producción de habla), reconocimiento del hablante.

## *1.1. Sistemas de conversión de texto a habla y bases de datos prosódicas*

Los sistemas de conversión de texto a habla reciben un texto cualquiera, antes ingresado en la computadora, y generan el sonido correspondiente intentando pronunciaciones que al percibirlas se asemejen al habla natural. Las áreas de aplicación son muy variadas y entre las principales pueden citarse: servicios que requieren respuesta telefónica hablada, ayuda a discapacitados visuales, juegos y libros con salida hablada, anuncios de viajes, avisos de problemas, etc.

Ellos deben "descubrir" la estructura superficial (fonemas) y profunda (significado, intención) del texto a pronunciar. Los sistemas de conversión de texto a habla tradicionales han evolucionado y se han desarrollado métodos que permiten obtener la estructura superficial, pero la alta calidad del habla por computadora sigue siendo una meta a alcanzar. Los sistemas de conversión de texto a habla actuales sólo pueden acercarse a la estructura profunda con la ayuda de una base de datos.

Han logrado muy buena calidad en la pronunciación de sílabas y otras unidades lingüísticas tomadas individualmente, pero la pronunciación en su conjunto es bastante monótona y sin un ritmo adecuado. Las investigaciones actuales proponen realizar previamente un entrenamiento prosódico previo, con una base de datos, para superar este escollo.

El entrenamiento con la base de datos permite derivar los parámetros que usarán los módulos constituyentes del conversor, y adicionalmente ayuda a definir, elegir y depurar las estrategias más adecuadas que se llevarán a cabo durante la conversión on-line.

Como el producto de la base de datos es dependiente del lenguaje se deben seguir las teorías lingüísticas y modelos prosódicos adecuados para incorporar texto representativo, junto a la información que lo caracteriza, de la lengua en cuestión.

## *1.2. Objetivos del presente trabajo de grado*

El objetivo primario es construir una base de datos, que sirva para el posterior entrenamiento y desarrollo de un sistema conversor de texto a habla, orientado al español hablado en Buenos Aires.

Los objetivos secundarios perseguidos son:

- Explorar y exponer los antecedentes y el estado del arte de los sistemas de conversión de texto a habla y de las bases de datos que estos conversores usan en búsqueda de la alta calidad.

- Aportar los conocimientos sobre las variadas tecnologías y modelos propios de la informática como integrante del equipo del LIS (Laboratorio de Investigaciones Sensoriales) que lleva adelante el desarrollo de un conversor de texto a habla para el español hablado en Buenos Aires.

- Desarrollar el software complementario requerido en las diversas etapas de construcción de la base de datos prosódica.

## *1.3. Consideraciones metodológicas relacionadas con el desarrollo de la base de datos*

Como todo software, el ciclo de vida de la base de datos comienza con una especificación de requerimientos. Luego sigue la etapa de diseño, la de implementación y, finalmente, la de prueba.

Normalmente se pasa por un período de análisis para llegar a la especificación de requerimientos, el cual se basó en la fase de análisis general del proyecto del LIS. De ella se utilizó una parte, que se reformuló para el trabajo a desarrollar. Además, esa parte tuvo que ser adaptada a un nuevo proceso de investigación, necesario para poder concretar los objetivos propuestos.

# Capítulo 2 SISTEMAS DE CONVERSIÓN DE TEXTO A HABLA

En este capítulo se exponen los resultados más importantes de la exploración que debió realizarse para comprender el ámbito que rodea a la construcción de las bases de datos que ayudan al desarrollo de los conversores de texto a habla. Por lo tanto fue indispensable conocer no sólo los antecedentes propios de estos conversores sino también la situación actual en la que se encuentra esta tecnología, detallando los problemas tradicionales y fundamentalmente los que pueden solucionarse con el uso de bases de datos como la construida en este trabajo.

La exploración debió abarcar parte de las áreas del conocimiento que intervienen en el desarrollo de estos conversores; desarrollo que recibe aportes fundamentales de la lingüística, la ingeniería y la informática. Dentro de este marco interdisciplinario se debieron explorar los modelos de producción de habla, los modelos entonacionales y los tipos de conversores de texto a habla que, más recientemente, usan bases de datos para su entrenamiento prosódico. La comprensión acerca del uso que se les da a estas bases de datos y los métodos posibles de interacción con ellas permitió definir detalladamente la información a introducir en la base de datos y su interfaz con el exterior, que se especifican en el capítulo siguiente.

## *2.1. Introducción a las tecnologías relacionadas con el habla*

Hay un amplio rango de tecnologías que caen bajo el título general de tecnologías del habla, que incluye a los sistemas que realizan algún tipo de procesamiento del lenguaje hablado: reconocimiento automático del habla, generación automática del habla (sistemas que hablan, sistemas de síntesis del habla, sistemas de conversión de texto a habla), sistemas con entrada y salida hablada (incluye sistemas que entienden el habla), sistemas que dialogan y sistemas traductores habla-habla, codificación del habla (reducen el tamaño en bytes del habla, manteniendo la inteligibilidad), análisis del habla o procesamiento paralingüístico (incluye verificación/identificación del hablante, verificación/identificación del lenguaje), aplicaciones en general: mejora del habla, conversión de voz, etc.

Muchas de estas tecnologías requieren una cantidad sustancial de datos del habla extraídos de las grabaciones. Estos datos son usados para derivar los parámetros que usan sus modelos constituyentes y para evaluar su comportamiento, repetitivamente, bajo condiciones de prueba controladas.

### 2.1.1 Marco interdisciplinario

El habla es común a todas las actividades del hombre. Sin embargo las principales áreas del conocimiento involucradas en las tecnologías del habla son:

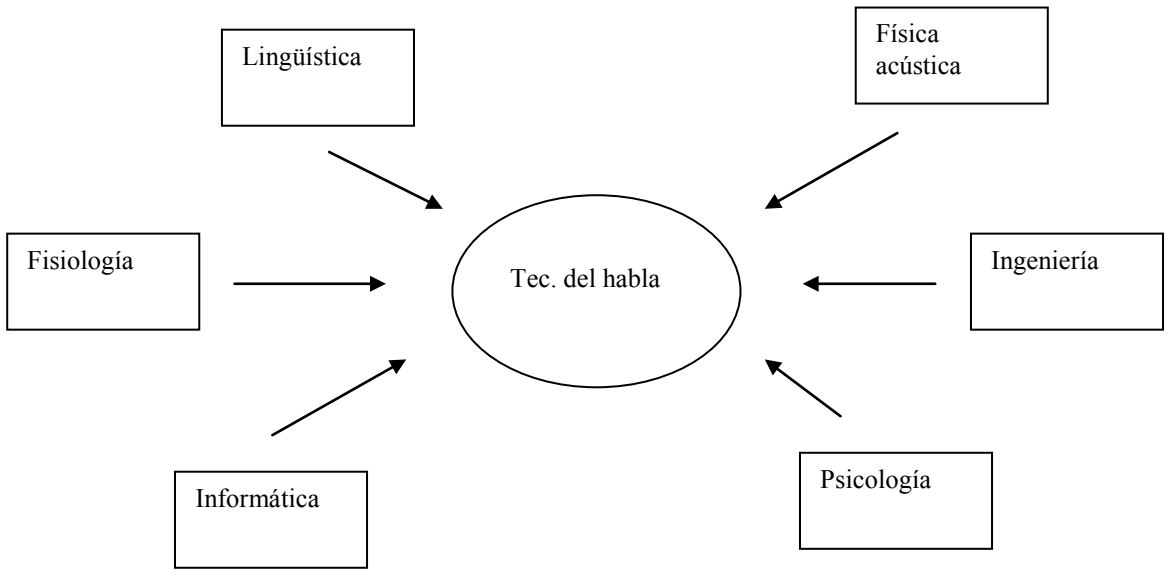


Figura 2.1. Tecnologías del habla y aportes de otras áreas del conocimiento.

La fisiología aporta la caracterización y descripción del funcionamiento del aparato fonador humano. La física acústica explica, a través de métodos formales, cuestiones tales como la propagación del sonido y sus componentes acústicos. La psicología aporta los conocimientos sobre el mecanismo sensoperceptivo que realizamos en el momento de recibir un mensaje oral.

En referencia a los sistemas de conversión de texto a habla, el trabajo interdisciplinario ayuda a objetivar todos los conocimientos teóricos necesarios, pero especialmente los referidos a la lingüística. Tradicionalmente los equipos de lingüistas han trabajado con métodos menos formales que los requeridos para estos desarrollos, en tanto que en los últimos tiempos se han logrado formalizar muchos conocimientos sobre fonética, fonología y prosodia, como por ejemplo los modelos de entonación actuales.

Los modelos de producción de habla, desarrollados tradicionalmente por ingenieros, han sido deficitarios en los aspectos referidos a la calidad del habla sintética que producían. Con el aporte de las teorías, modelos y métodos formalizados aportados por la lingüística se ha conseguido la integración necesaria en el camino de la búsqueda de habla artificial de alta calidad. Ninguno de los intentos de creación de conversores de texto a habla realizados por grupos de

especialistas de esas áreas, trabajando casi aisladamente, ha llegado a producir buenos resultados.

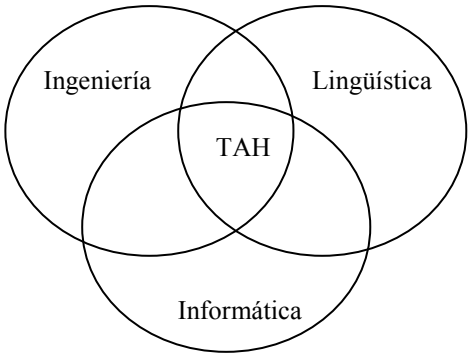


Figura 2.2. Marco interdisciplinario de la tecnología de conversión de texto a habla actual.

La informática ayuda a integrar las estrategias de los lingüistas y los ingenieros, aportando principalmente los modelos de representación de la información, los métodos de desarrollo de software correspondientes y las plataformas básicas de software necesarias durante el trabajo con las computadoras.

### 2.1.2 Técnicas básicas en el procesamiento de la señal acústica

Seguidamente se describen las dos técnicas básicas involucradas en el procesamiento de la señal acústica del habla.

#### *2.1.2.1 Análisis del habla*

El análisis acústico de la señal del habla es un proceso que descompone la señal en sus elementos constituyentes: contornos de frecuencia fundamental (F0), energía total (E), formantes (F1, F2, F3, F4, F5 y F6), etc. Está basado principalmente en la transformada discreta de Fourier (DFT) y en métodos temporales y homomórficos [RABINER, 1979].

La frecuencia fundamental es el indicador cuantitativo de la frecuencia de vibración de las cuerdas vocales. Su continuidad determina la fuente de excitación que intervino en la generación del sonido. El contorno de frecuencia fundamental constituye una de las características determinantes de la entonación.

La energía indica la cantidad total de energía sumada para todas las componentes de frecuencia o formantes.

Los formantes son contornos que representan las mayores concentraciones de energía correspondientes a las frecuencias de resonancia que se generan en las cavidades acústicas del aparato fonador. Ellos representan las características particulares de cada unidad del habla.

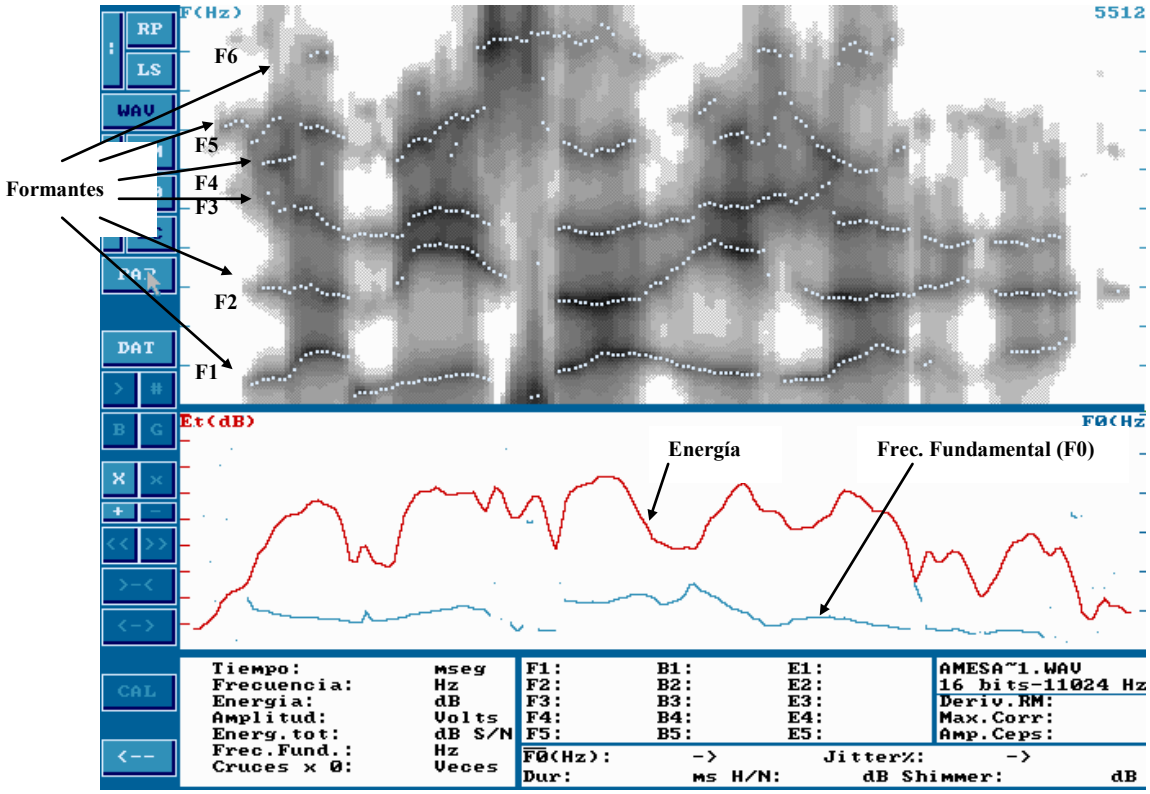


Figura 2.3. Descomposición de la señal del habla (software Anagraf).

### *2.1.2.2 Síntesis del habla*

La síntesis de la señal del habla puede ser vista como el proceso inverso al de análisis, o sea, compone la señal acústica del habla a partir de sus elementos constituyentes.

Este proceso es el que, en su etapa final, llevan a cabo los sintetizadores de habla basados en parámetros acústicos (ver más adelante 2.1.4.1 Sintetizadores y 2.2.2.1 Sintetizadores de formantes), a diferencia de los basados en parámetros articulatorios, para componer diversas unidades de concatenación. El término

síntesis del habla se ha extendido y en muchas ocasiones se lo utiliza para representar a todos los procesos que realiza un sintetizador.

Un sintetizador ideal debería ofrecer una elevada calidad en los enunciados producidos tanto en lo que se refiere a la naturalidad, asociada a las características suprasegmentales, como a la inteligibilidad, asociada a las características segmentales (ver más adelante "Problemas de calidad").

### 2.1.3 Comunicación verbal

El habla es el principal medio de comunicación del hombre. Históricamente su producción ha causado fascinación e intriga. Así lo demuestran los tempranos intentos que hizo el hombre, en la época de los griegos, por simular el habla artificialmente. Pero allí no se hicieron más que trucos para hacer creer al pueblo en general que se trataba de voces que provenían de los dioses, tarea que en realidad realizaban hombres escondidos detrás de las paredes.

Luego, durante el renacimiento se usaron aparatos mecánicos para emitir sonidos, algunos de los cuales resultaron muy ingeniosos, como los de Von Kempelen en 1791 [FLANAGAN, 1972]. Estos primeros sonidos artificiales sólo pretendían parecerse a los sonidos más elementales de la voz humana. La simulación del habla era imposible.

Recién en este siglo se desarrollaron las primeras teorías sobre la producción del habla y la comunicación oral. A partir de entonces se pudo comprender que en un mensaje hablado existe mucha información, la cual puede clasificarse según los siguientes niveles de abstracción mental que hacemos en el momento de la audición, percepción e interpretación del mismo:

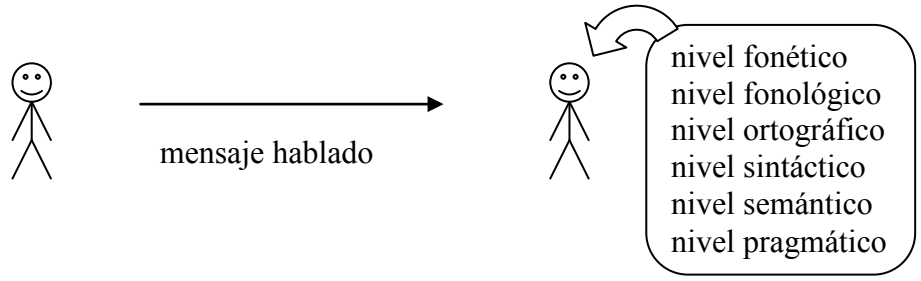


Figura 2.4. Mensaje hablado: niveles de abstracción mental.

Las características fonéticas y fonológicas se describen más adelante. Con respecto a los niveles semántico y pragmático debe decirse que la entonación es un aspecto muy significativo, ya que oraciones que difieren únicamente en la realización

entonacional pueden diferenciarse además en su significado, semántico o pragmático. La misma oración puede ser pronunciada con un tono afirmativo, dubitativo o interrogativo, entre otras posibilidades, cambiando así la intención [SOSA, 1991].

Ejemplo:

No, se acabó la yerba.

¿De cuántas maneras diferentes se puede pronunciar esta oración? Claramente se puede distinguir el único significado semántico que posee, sin embargo, la intención (significado pragmático) puede ser disímil.

#### *2.1.3.1 Comunicación oral entre el hombre y la máquina*

Ha existido una gran variedad de proyectos con larga data intentando imitar la voz y el habla del hombre por medios artificiales. Los primeros aparatos comerciales de bajo costo, con posibilidad de uso masivo, aparecieron en Estados Unidos a partir de mediados de los setenta. Muestra de ello son las primeras calculadoras parlantes. A partir de allí se ha producido un aumento vertiginoso de los aparatos que intentan imitar la voz y el habla humana.

Además, se debe señalar la importante evolución en la calidad del habla que se ha venido manifestando desde aquel momento. Aún así hay mucho para mejorar. Ejemplo de ello es el siguiente fenómeno: la generación automática de voz y habla por parte de las computadoras sigue siendo, en el marco de la interacción hombre-máquina mediante el lenguaje natural, una de las principales líneas de investigación tanto teórica como aplicada, involucrando desde las ciencias básicas hasta las problemáticas de diseño de software propios de la informática.

Para que la interacción hablada entre los hombres y las máquinas sea posible, el proceso debe ser bidireccional: las computadoras deben ser capaces de producir mensajes (generación de habla) y de decodificar los que les llegan (reconocimiento del habla). Las investigaciones en ambos campos ya han dado muy buenos frutos, existiendo en la actualidad varias aplicaciones comerciales de buena calidad, orientadas al público en general. Se pueden citar, entre los más difundidos, los conversores de texto a habla de IBM, Dialogic, Bell Labs, y los programas de reconocimiento de habla de Dragon Systems, Lernout & Hauspie, Philips e IBM. En particular para el español existen muy pocos sistemas de conversión de texto a habla de buena calidad, ninguno para el español de Buenos Aires.

A pesar del enorme avance logrado en el campo de las tecnologías del habla, aún están muy inmaduros los proyectos que involucran la idea de conversión de

concepto a voz. Ya sea para que la computadora responda oralmente a preguntas que se le realizan, luego de reconocer la voz y el habla de una persona y el significado de las palabras y oraciones, o para que reemplacen a los actuales traductores humanos, necesarios en todos los eventos y relaciones internacionales.

### 2.1.4 Generación de habla artificial

Desde las primeras máquinas del siglo XVIII, pasando por los primeros ensayos eléctricos como el Voder, los primitivos sintetizadores hasta los complejos sistemas de conversión de texto a habla actuales, se ha recorrido un largo camino que aún no ha llegado a su fin. Muestra del interés que genera este campo son los arrolladores avances que dieron lugar a grandes proyectos nacionales en los países centrales, inclusive megaproyectos que involucran a casi todo un continente, como ocurrió en la comunidad europea con los programas ESPRIT, SAM, EAGLES, etc.

#### *2.1.4.1 Sintetizadores*

A partir de la década del setenta, siguiendo la evolución de las computadoras, se desarrollaron los primeros sintetizadores, sobre la base de modelos de producción de habla elementales. También en esta época se crearon los primeros programas de análisis de la señal acústica del habla.

Un sintetizador, antes del proceso final de síntesis, debe realizar un proceso de elección de los contornos constituyentes de la señal acústica, como se observa en el siguiente esquema:

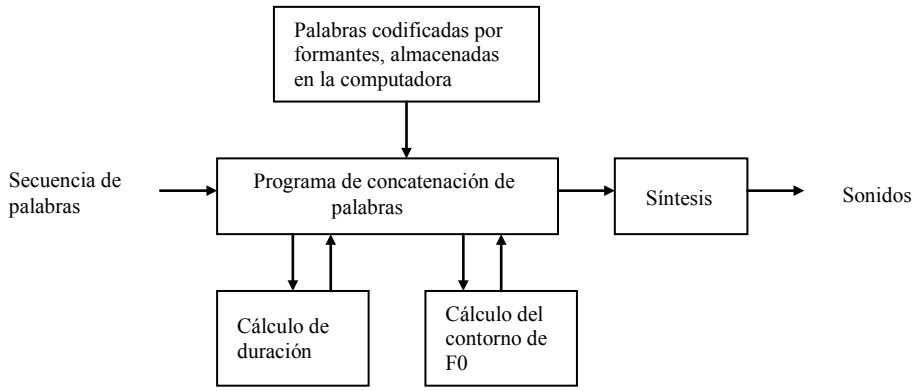


Figura 2.5. Esquema del sintetizador experimental de Bell Labs en 1971.

Los primeros sintetizadores trabajaban de dos maneras diferentes [FLANAGAN, 1972]. Los sintetizadores basados en datos del habla, que obtenían parámetros y símbolos especiales de habla pregrabada, analizada y etiquetada, para indicar las características fonéticas y prosódicas. Los basados en reglas, que no tenían datos pregrabados, sino reglas preestablecidas para la producción del habla. Como superación de las técnicas básicas surgieron estrategias que combinan varias de esas técnicas.

Luego evolucionaron, se hicieron más complejos y pasaron a llamarse sistemas de conversión de texto a habla. Estos sistemas son mucho más sofisticados que sus antecesores pues realizan un procesamiento del texto ingresado, cuyo producto alimenta al proceso de síntesis final. O sea, deben realizar diversos tipos de análisis para deducir y derivar los símbolos fonéticos y prosódicos a partir del texto ingresado, los cuales van a parar al módulo de procesamiento de señal (mucho más complejo que los sintetizadores iniciales), que desarrolla complejas estrategias sobre la base de parámetros y algunas reglas básicas.

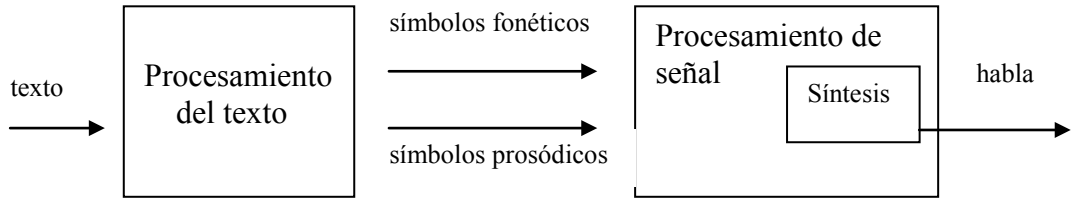


Figura 2.6. Los dos módulos principales de un conversor de texto a habla.

También evolucionaron las técnicas de síntesis y se empezaron a concatenar no sólo parámetros del habla, sino también unidades naturales del habla codificadas con las mismas técnicas que usan los archivos de onda, tradicionalmente identificados por extensiones como wav, au, etc.

### *2.1.4.2 Problemas de calidad*

La calidad del habla artificial ha sido tradicionalmente difícil de medir y valorar. Su percepción auditiva ha sido y sigue siendo el método más eficiente de medición.

Durante esta percepción, que siempre es subjetiva, se pueden distinguir dos niveles abstractos que componen el mensaje recepcionado: el nivel segmental y el suprasegmental.

La información segmental es la asociada a la cadena de sonidos que componen el mensaje. Los sonidos que se pueden producir a través del aparato fonador humano

son variados, aún considerando un único locutor. En cada idioma se han seleccionado una serie limitada de sonidos ideales que son los segmentos básicos que afectan a la calidad segmental. Por ejemplo, la representación segmental de la palabra casa en el español de Buenos Aires puede ser: /k/, /a/, /s/, /a/.

La información suprasegmental es aquella que queda asociada a la prosodia y en particular a la entonación. Refleja tanto elementos lingüísticos (carácter de la frase, pausas, acentos, agrupación en elemento de significado), como elementos no lingüísticos (intención, sentido, estado de ánimo características personales del locutor, etc.). Esta información es la clave para conseguir una alta naturalidad; y generalmente se codifica en el momento de la síntesis a través de tres parámetros acústicos:

a) La evolución temporal de la frecuencia fundamental (F0), que es el correlato físico más importante desde el punto de vista perceptivo.
b) La duración de los segmentos o sonidos que componen la frase.
c) La curva de energía (E) de la señal acústica, que es el parámetro menos importante desde el punto de vista perceptivo.

Los primeros sintetizadores no podían alcanzar buenos niveles de calidad; solamente producían segmentos de habla inteligibles, aunque de baja calidad. Con el correr del tiempo se produjeron importantes avances en el aspecto segmental pero no se lograban sonidos naturales, siempre sonaban a computadora, a robot, o sea, "latosos". Los componentes que afectaban al aspecto suprasegmental eran aún desconocidos, la teoría lingüística no los explicaba. La entonación era uniforme, sin relación con la acentuación ni con el ritmo. Podríamos definir la relación inicial entre la ingeniería y la lingüística, en este campo, como casi inexistente.

Se sabía que la calidad segmental podría ser alcanzada hasta un grado aceptable pero no había muchas esperanzas con respecto a la calidad suprasegmental. Se debió esperar entonces el desarrollo de teorías lingüísticas, fundamentalmente las referidas a la entonación, para albergar esperanzas de mejoras importantes en la prosodia del habla artificial.

### 2.1.5 Teorías lingüísticas y modelos representativos

Las teorías lingüísticas describen los conceptos generales y particulares de la lengua, tanto los referidos a los símbolos escritos como aquellos que se presentan en la emisión acústica, o sea, en el habla. La comprensión y el correcto uso de estos conceptos son la base de partida de todo sistema de conversión de texto a habla que pretenda alcanzar la alta calidad.

En referencia a los problemas tradicionales de calidad suprasegmental derivados de la falta de modelos prosódicos, la lingüística aporta los modelos de entonación y los métodos para poder interpretarlos.

Antes de pasar a explicar los modelos de entonación es conveniente aclarar algunos conceptos básicos sobre fonética, fonología y prosodia.

### *2.1.5.1 Fonética y fonología*

La fonética es la rama de la lingüística que se ocupa del estudio de los sonidos de una lengua. La fonología se refiere específicamente a las representaciones abstractas de los sonidos diferenciables de una lengua. La unidad fonética mínima es el fono, mientras que la fonológica es el fonema.

Ejemplos:

| palabra | fonemas | fonos |
|---|---|---|
| Casa | /k/ /a/ /s/ /a/ | [k] [a] [s] [a] |
| Baba | /b/ /a/ /b/ /a/ | [b] [a] [ß] [a] |

Como se puede observar los símbolos fonológicos se representan entre barras inclinadas y los fonéticos entre corchetes. También se puede destacar que en muchos casos coincide exactamente la representación fonológica mental con las características fonéticas del habla. Pero en algunos casos, se utiliza el mismo símbolo fonológico para representar a dos o más realizaciones fonéticas de la misma unidad (llamados alófonos).

Desde el punto de vista de la entonación, el estudio fonético sirve para descubrir el sistema de relaciones fonológicas que se pone en juego en el momento de la comunicación vía habla.

### *2.1.5.2 Prosodia*

La prosodia es la parte de la gramática que abarca el estudio de las leyes de la estructura métrica, las cuestiones relacionadas con el ritmo y también todos los procedimientos que afectan a la articulación melódica del habla, cuyos factores más determinantes son la entonación y el acento.

Las características prosódicas tienen funciones específicas (conectiva y demarcativa) durante la comunicación oral. Ayudan al oyente en la interpretación de una pronunciación: agrupando palabras en unidades de información más grande

(frases intermedias o frases entonacionales) y dirigiendo la atención a palabras específicas (foco). En el ámbito de los sistemas de conversión de texto a habla la prosodia es la realización en el habla de la estructura profunda del texto, o sea, del significado semántico y pragmático.

El efecto prosódico más fácil de observar es el producido por el énfasis que se le da a ciertas partes del discurso, el cual permite destacar dichas partes que llamaremos "focos". Por ejemplo, ciertas tonalidades aplicadas a una sílaba permiten destacarla e indirectamente la palabra o grupo sintáctico al que pertenece dicha sílaba será también destacado (remarcado) como un componente importante dentro del significado de la pronunciación.

#### *2.1.5.3 Teorías entonacionales y modelos de entonación*

Los modelos entonacionales intentan describir principalmente el comportamiento de los contornos de frecuencia fundamental y de energía. En la actualidad existen varios modelos entonacionales como el propuesto por [FUJISAKI, 1993], el de "subidas y caídas" propuesto por [HART ('t), 1988], el modelo Tilt [DUSTERHOFF, 1997], el modelo de [PIERREHUMBERT, 1988]. El último ha sido adoptado en este trabajo, a través de su implementación por el método ToBI.

*Unidades prosódicas*

El discurso en general y las oraciones en particular se pueden dividir en unidades prosódicas de distinto tamaño. [NAVARRO TOMAS, 1974] se refiere a esta circunstancia de la siguiente manera:

"La división de la frase en unidades melódicas no es un hecho que se produzca siempre de una manera uniforme e invariable. Una misma frase de cierta extensión puede ser dividida en mayor o menor número de unidades, según la intención especial con que en cada caso se actualice su sentido. (…) Influyen en esta división circunstancias de orden lógico y emocional".

La forma en que el habla continua se organiza en un conjunto finito de unidades fonológicas es estudiada por la teoría fonológica que llega a definir hasta siete niveles que van desde la sílaba hasta el enunciado prosódico total. Aquí conviene detenerse en el nivel de la frase entonacional o grupo melódico que es el más relevante.

Se ha establecido que los grupos melódicos (frases entonacionales) son de una importancia capital en las fases de percepción y procesamiento de la cadena de sonidos del habla emitidos y que el oyente se basa principalmente en éstas y posiblemente en otras unidades fonológicas para la decodificación del mensaje y no

en la estructura sintáctica de las oraciones [SOSA, 1991]. Los grupos melódicos proporcionan el eslabón que el oyente necesita para proceder al análisis semántico y sintáctico de una oración, partiendo de la percepción de la secuencia de sonidos que la integra.

Se puede identificar una estructura jerárquica a partir de esos grupos que no necesariamente se corresponde con la estructura sintáctica (también jerárquica) de toda pronunciación.

Muchos factores influyen en el tamaño y composición de estas unidades de la estructura prosódica. Entre estos factores se encuentran las aparentes divisiones obligatorias indicadas ortográficamente como en el caso de las oraciones entre paréntesis y las subordinadas.

*Ejemplo de frase entonacional y frase intermedia*

Una oración compleja puede contener varias "frases entonacionales". A su vez una "frase entonacional" puede incluir dos o más "frases intermedias":

Oración: Ayer leí la novela en la oficina, hoy lo hice a pleno sol.

Frases entonacionales (en este caso claramente diferenciadas por la coma):
- ayer leí la novela en la oficina
- hoy lo hice a pleno sol

Frases intermedias:

Las frases intermedias contienen sílabas muy relacionadas entre sí, que forman la menor unidad de entonación, pudiendo entonces encontrarse distintas realizaciones para la misma frase entonacional.

Caso 1)
- ayer
- leí la novela
- en la oficina
- hoy
- lo hice a pleno sol

Caso 2)
- ayer
- leí
- la novela
- en la oficina
- hoy

- lo hice
- a pleno sol

*Unidades lingüísticas que influyen en la conversión de texto a habla*

Las unidades largas requieren un número más elevado de ellas para la producción de un conjunto de mensajes. La frase es un ejemplo de este tipo de unidades. La codificación de frases completas mediante las técnicas de codificación conocidas (por ejemplo las usadas para producir archivos wav) permiten obtener habla de una inteligibilidad y una naturalidad insuperables, aunque naturalmente, mediante estas técnicas, nunca puede conseguirse la síntesis de un número ilimitado de mensajes, especialmente si éstos son imprevisibles. Sin embargo, muchas aplicaciones de uso doméstico no requieren más que la producción de un número limitado de enunciados, por lo que la frase constituye una unidad adecuada.

Por el contrario, las unidades pequeñas ofrecen mucha mayor flexibilidad cuando se requiere la conversión en habla de un texto sin restricciones, puesto que a partir de ellas pueden formarse unidades mayores. Al reducir el tamaño disminuye también el número de unidades que se necesitan. De hecho, el inventario de fonemas de una lengua natural se sitúa entre los 20 y los 37 elementos [LLISTERRI, 1988] y, mediante la combinación de estas unidades básicas se forma un conjunto infinito de enunciados. Como contrapartida, se trata de unidades abstractas (el fonema no tiene existencia real en la onda sonora) que, en sus realizaciones concretas presentan un grado muy alto de variabilidad y son, por lo tanto, difícilmente identificables.

## *2.2. Arquitecturas actuales de los sistemas de conversión de texto a habla*

Durante mucho tiempo las técnicas de conversión de texto a habla no estuvieron bien definidas pues no había equipos interdisciplinarios que integraran los conocimientos propios de cada especialidad. Se ha evolucionado rápidamente en los últimos años y hoy en día los sistemas de conversión de texto a habla realizan complejos procesos hasta llegar a producir habla artificial.

En la actualidad existe una amplia variedad de sistemas conversores de texto a habla, desde los que trabajan puramente sobre la base de "reglas" hasta los que solamente "concatenan" segmentos de habla previamente grabados. Sin embargo aquellos que trabajan con una metodología mixta han demostrado ser los más eficientes hasta el momento. Estos sistemas, por un lado, concatenan unidades acústicas previamente grabadas, almacenadas y catalogadas (ejemplo: fonos, difonos) y por el otro, generan las características prosódicas sobre la base de complejas estrategias que a su vez combinan métodos probabilísticos con algunas reglas lingüísticas básicas.

Un sistema de conversión de texto a habla se compone de dos módulos claramente diferenciados, que requieren para su realización una metodología y conocimientos de base radicalmente distintos: procesamiento lingüístico prosódico y el procesamiento acústico.

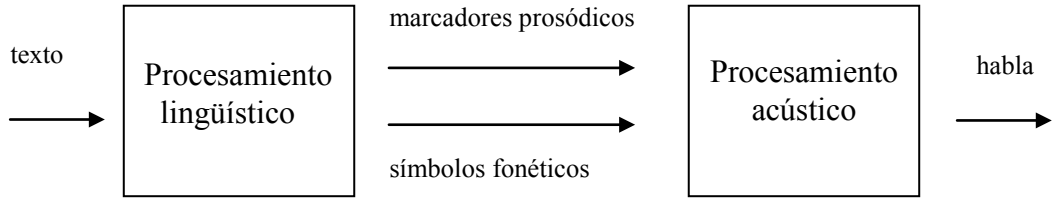


Figura 2.7. Módulos principales de un sistema de conversión de texto a habla actual.

Entre estos dos grandes bloques son muchos los puntos de conexión. El punto medio de encuentro es la representación fonética y prosódica del texto.

### 2.2.1 Procesamiento lingüístico

El objetivo general del procesamiento lingüístico es determinar, a partir de un texto, dos tipos de información necesarios para proporcionar al módulo de procesamiento acústico los datos necesarios en la generación de habla natural. Estos dos tipos de información se conocen como información segmental (por ejemplo los

fonos) e información suprasegmental (por ejemplo marcadores prosódicos representados por símbolos tonales y métricos).

En los sistemas actuales de conversión de texto a habla se sigue la siguiente secuencia de funciones que conforman este módulo:

#### *2.2.1.1 Pre-procesamiento*

La primera tarea a realizar dentro del procesamiento lingüístico es dar formato al texto adecuadamente, representando en forma textual números, abreviaturas, etc. Este proceso se realiza, generalmente, mediante autómatas traductores que utilizan gramáticas regulares elementales [DUTOIT, 1996]. Normalmente este módulo delimita también la unidad dentro del texto que va a ser tratada por el resto de los módulos. Es un módulo bastante dependiente de la aplicación en la que se desenvolverá el conversor.

#### *2.2.1.2 Análisis lingüístico*

Después del pre-procesamiento se realiza un análisis lingüístico, desde los puntos de vista sintáctico y semántico, para tratar de hallar el foco (segmento con mayor contenido semántico) de la oración e intentar modelar aspectos como el énfasis. Esta tarea es de bastante complejidad y muy dependiente del idioma considerado. Se realiza mediante un análisis del texto adaptado a los propósitos de generación prosódica. El análisis gramatical suele realizarse usando un diccionario que posee un léxico relevante (locuciones adverbiales, perífrasis verbales, expresiones idiomáticas) incluyendo raíces, prefijos y sufijos. Normalmente se incluyen reglas gramaticales para determinar las categorías de las palabras que no han sido encontradas en el diccionario.

#### *2.2.1.3 Análisis morfosintáctico-prosódico*

En el análisis morfosintáctico-prosódico se pretende, a partir del análisis anterior, marcar, por un lado, fronteras sintáctico prosódicas y, por otro, los acentos. Una frontera sintáctico-prosódica queda definida por su contexto, el cual determina su relación lógica con las estructuras anterior y posterior (ejemplo: pausa, alargamiento de la sílaba anterior a la marca). Los acentos obedecen más bien a aspectos rítmicos y de énfasis principalmente. Estas marcas prosódicas son

importantes para ayudarnos al entendimiento del mensaje y son las que permiten darle naturalidad.

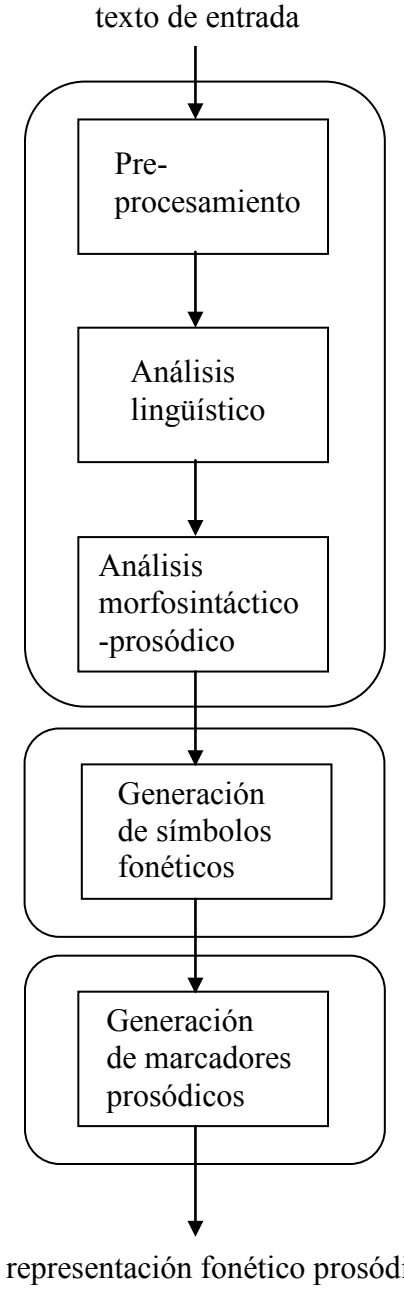


Figura 2.8. Módulo de procesamiento lingüístico de un conversor actual.

En los sistemas actuales de conversión de texto a habla, las funciones utilizadas para este análisis suelen estar basadas en estrategias que combinan el análisis sintáctico con el uso de reglas sintáctico-prosódicas de contexto gramatical para situaciones ambiguas. Para ello se utilizan diversas técnicas como por ejemplo

las basadas en modelos de Markov, en redes neuronales, en árboles de regresión y clasificación [DUTOIT, 1996]. Es decir, se presupone que estas marcas prosódicas pueden ser deducidas automáticamente del texto de la oración y que gracias a ellas se logrará un mejor entendimiento de las oraciones por parte del oyente.

Finalmente se deben aplicar reglas de acentuación léxica para la determinación de la posición de los acentos.

#### *2.2.1.4 Generación de símbolos fonéticos*

Esta tarea se realiza después del análisis morfosintáctico-prosódico y consiste en realizar una transcripción fonética del texto, también llamada etapa de fonetización automática. Esta etapa es para el español relativamente sencilla. La función a realizar puede verse como una transformación de letras a fonemas y luego de éstos a fonos. Estas transformaciones se realizan, generalmente, mediante reglas dependientes del contexto. Estas reglas deben tener en cuenta la acentuación y principalmente la existencia de pausas en el texto (un espacio en blanco en la escritura raramente se realiza como una pausa).

#### *2.2.1.5 Generación de marcadores prosódicos*

La última tarea a realizar se denomina procesamiento prosódico y recoge la información segmental y suprasegmental generada en los dos últimos pasos (las marcas prosódicas y la transcripción fonética), para traducirla en variaciones de duración segmental (ritmo), de frecuencia fundamental (entonación), e insertar las pausas que existan con una duración adecuada. La función que permite realizar este proceso es una compleja estrategia que asocia el resultado del análisis morfosintáctico-prosódico con estructuras prosódicas almacenadas en una base de datos prosódica previamente grabada y etiquetada, mediante el uso de técnicas probabilísticas o de inducción, a través de los ejemplos contenidos en la base de datos. La calidad y variedad de esta base de datos permite dar un carácter menos monótono al habla sintética. Aquí conviene aclarar que esta base de datos se usa en la conversión on-line, a diferencia de la base de datos construida en este trabajo de grado, que se usa para el entrenamiento del conversor.

Hasta ahora los sistemas de conversión de texto a habla han tratado de obtener patrones prosódicos neutrales para evitar que las oraciones resultaran antinaturales por algún error en el etiquetado. Sin embargo, la eficiencia actual del análisis lingüístico permite procesar patrones prosódicos más complejos (incluyendo varias variaciones de frecuencia fundamental) que son la clave para obtener habla sintética cercana a la producida por un locutor.

### 2.2.2 Procesamiento acústico

El objetivo general que el procesamiento acústico persigue es convertir la cadena fonética y las variables de control prosódico en los componentes de la señal acústica del habla a emitir. Existe una amplia gama de sistemas de conversión texto a habla que abarca desde los sistemas dirigidos por reglas a los dirigidos por los datos. Dicho de una manera concisa, en un sistema "puro" dirigido por reglas, éstas generan la representación paramétrica que alimentará un sintetizador de habla, y en uno dirigido por los datos, éstos representan directamente segmentos de habla. Entre estos dos podemos encontrar multitud de casos intermedios. Así por ejemplo, los segmentos de habla pueden estar parametrizados de acuerdo a un modelo de codificación de habla. En cualquier caso, el modelo de producción de habla debe ser flexible para el control prosódico en busca de la alta calidad del habla sintética.

Actualmente se utilizan modelos que pueden clasificarse en tres grupos principales:

#### *2.2.2.1 Sintetizadores de formantes*

En estos sintetizadores la cadena fonémica y la prosodia controlan las resonancias y la excitación de un sintetizador de formantes. Un exponente claro de este tipo de sistemas lo constituye el sintetizador de Klatt [KLATT, 1980]. El sintetizador de formantes consiste en una composición de filtros que modelan las resonancias y antiresonancias de las cavidades vocal y nasal. Para este modelado se usan filtros que en la configuración más general están conectados en serie y en paralelo. Es un procedimiento de enorme flexibilidad que se pone de manifiesto en la alta calidad del habla sintética que se puede obtener mediante ajuste manual de los parámetros del sintetizador. Sin embargo, se necesitan un número enorme de parámetros en la síntesis automática lo que requiere compiladores cada vez más sofisticados capaces de integrar todo el conocimiento que se adquiere a base de experimentar con el sistema.

#### *2.2.2.2 Sintetizadores basados en concatenación de unidades*

Como su propio nombre indica, en estos sistemas se concatenan un conjunto de unidades extraídas del habla natural. En este tipo de sintetizadores debe estar presente un algoritmo que permita, además de la concatenación de unidades, modificar prosódicamente los segmentos a concatenar. Adicionalmente, en este tipo de sintetizadores se pueden usar técnicas de codificación de habla para reducir el

tamaño de las unidades acústicas almacenadas. También existe la posibilidad de incluir en el modelo de codificación de habla las tareas de concatenación y modificación prosódica, siempre que el codificador parametrice la señal de habla con la suficiente flexibilidad para la adecuación prosódica de las unidades.

Dos son los aspectos fundamentales que se deben abordar para la generación de la señal en un sistema de síntesis por concatenación de unidades:

1) La modificación prosódica: involucra el cambio de los siguientes factores:

- Duración de la emisión: requiere algoritmos de modificación de la escala temporal.
- Contorno melódico: requiere algoritmos de modificación de la frecuencia fundamental.
- Intensidad: requiere algoritmos de modificación de la envolvente energía de la señal de habla.

2) La concatenación de unidades: involucra el suavizado de las unidades a concatenar para evitar las discontinuidades. Puesto que normalmente dichas unidades son extraídas del habla natural, proceden de palabras diferentes. Por lo tanto presentan, en general, discontinuidades en el valor de la frecuencia fundamental, en la envolvente espectral y en la fase de los dos segmentos a concatenar.

La resolución de estos problemas debe realizarse preservando la calidad del habla original en términos de naturalidad, inteligibilidad y características dependiente del locutor.

En la actualidad se ha logrado alcanzar alta calidad segmental usando las técnicas de síntesis por concatenación, sin embargo son muy útiles para el control prosódico, aunque no han logrado tanta calidad, las técnicas basadas en síntesis por formantes.

### 2.2.3 Problemas actuales

Como se ha indicado, en la actualidad se han logrado resolver los problemas de calidad segmental, no así los referidos a la prosodia del habla emitida. Estos problemas, asociados a la falta de calidad suprasegmental, se originan en las dificultades que tiene el módulo de procesamiento lingüístico para descubrir la estructura profunda a partir del texto.

Por lo tanto los sistemas actuales han tratado de obtener patrones prosódicos neutrales para evitar que las oraciones resultaran antinaturales, perdiendo así la

posibilidad de alcanzar la alta calidad. Esta política es la comúnmente encontrada en los sistemas comerciales.

Para solucionar estos problemas se debe formalizar la relación entre sintaxis-semántica-pragmática y los patrones prosódicos que deben derivarse de ellos, o sea, es el módulo de procesamiento lingüístico el que necesita las mejoras. Estas mejoras se logran a través de estrategias para derivar los correctos marcadores prosódicos a partir del texto. Se han propuesto variadas metodologías, las cuales están siendo desarrolladas en distintos laboratorios de investigación. La mayoría propone incluir en la base de datos prosódica de uso on-line los patrones prosódicos característicos de la lengua en cuestión. Estos se obtienen a partir del entrenamiento con una base de datos previamente creada que contiene variedad de oraciones y los rasgos prosódicos correspondientes a ellas. O sea, la base de datos es construida, en varias etapas, mucho tiempo antes de tener funcionando la aplicación definitiva que realiza la conversión de texto a habla. Ellas se utilizan principalmente para entrenar algunos módulos del sistema conversor de texto a habla, ayudando a definir, elegir y depurar las estrategias necesarias (principalmente de derivación y generación de rasgos prosódicos) a la hora de realizar la conversión.

### 2.2.4 Proyecto de desarrollo de un sistema de conversión de texto a habla para el español de Buenos Aires

Este proyecto surgió en el Laboratorio de Investigaciones Sensoriales (LIS) a partir de su amplia trayectoria en el campo de las tecnologías del habla. Durante más de veinte años se ha adquirido una vasta experiencia en investigación y desarrollo de las técnicas de análisis y síntesis de la señal del habla, contando en la actualidad con productos de software que realizan estas tareas: Anagraf y Sinpar respectivamente.

La creación de un sistema de conversión de texto a habla para el español de Buenos Aires se ha dividido en tres partes. La primera etapa consiste en construir una base de datos prosódica de entrenamiento, la segunda etapa es aquella en la que se entrena al sistema y la tercera etapa es la de creación de un sintetizador moderno que integre las estrategias y los datos resultantes del entrenamiento.

La construcción de la base de datos durante la primer etapa es parte de este trabajo de grado y se explica en el siguiente capítulo, como así también las cuestiones referidas al entrenamiento con ella. La tercera etapa aún no está definida.

## *2.3. Bases de datos del habla*

### 2.3.1 Evolución

Desde la visión actual, los primeros sintetizadores usaban unos pocos datos del habla para realizar su trabajo. Con el correr de los años estos datos se convirtieron en archivos cada vez más grandes y más numerosos. Paralelamente, el avance de otras tecnologías del habla requirió el almacenamiento de grandes volúmenes de información. Así surgieron los primeros Corpus de Habla, que guardaban en archivos comunes su conocimiento. Algunos ejemplos son el corpus utilizado para el megaproyecto europeo SAM que funcionó desde 1987 hasta 1993 y el construido para el proyecto alemán VERBMOBIL. En el proyecto SAM se usaron archivos del tipo D-BASE File (extensión DBF) como medio de almacenamiento. Entre ellos podemos citar el archivo corpus.dbf que mantiene información general del corpus y también referencia a otros archivos del mismo tipo y también de texto ASCII [EAGLES, 1997].

Los sistemas de conversión de texto a habla usan parte de estos corpus de habla como bases de datos prosódicas de consulta on-line, de la cual obtienen los símbolos fonéticos y prosódicos que necesitan (ver sección anterior "Generación de marcadores prosódicos" para más detalles). Sin embargo, como ya se comentó, no han alcanzado la alta calidad pues necesitan un entrenamiento previo en el que haya participación del hombre, no sólo durante la etapa de construcción de la base de datos prosódica, sino también durante la etapa de entrenamiento en la cual se produce la definición, elección y puesta a punto de las estrategias a utilizar durante la conversión on-line [ROSS, 1999].

La bases de datos actuales son más difíciles de estudiar pues la mayoría de ellas no están accesibles. Por ejemplo el Linguistic Data Consortium (LDC), fundado en 1992, no ha distribuido públicamente sus bases de datos. Sin embargo existen proyectos que brindan alguna información sobre su forma de trabajo (Festival, IMS, Mbrola), permitiendo deducir las características lingüísticas y prosódicas que almacenan y la interacción necesaria para trabajar con ellas. Algunas son específicas de una lengua particular y otras, de gran envergadura, poseen características multilingües.

### 2.3.2 Información que almacenan

A la hora de considerar qué se carga en una base de datos del habla se debe decidir cuáles son los propósitos específicos y generales que motivan su construcción. No es exactamente la misma información la que se utiliza para crear un sistema de reconocimiento del habla que la que se ha comentado durante la exposición de los sistemas de conversión de texto a habla. Para las tareas de reconocimiento es necesario trabajar con habla espontánea, mientras que para producir habla artificial es conveniente un corpus de habla controlado prosódicamente, que se consigue creando primero el corpus de oraciones y luego grabando y etiquetando las emisiones correspondientes a esas oraciones [EAGLES, 1997].

También deben considerarse los diversos modelos lingüísticos a seguir que determinan los datos a almacenar. Por ejemplo, quienes trabajan siguiendo el modelo Tilt deben considerar los datos referentes a la cresta o pico: inicio de la cresta, máximo de la cresta y fin de la cresta.

Una de las primeras problemáticas a resolver se plantea en el momento de elegir las unidades lingüísticas a almacenar, dependiendo en algunos casos del tipo de síntesis que se pretende utilizar.

#### *2.3.2.1 Unidades lingüísticas*

Se presentan tres factores a contemplar en el momento de escoger una determinada unidad. Estos son:

- el tamaño de la unidad (frase, palabra, sílaba, fonema, fono, demisílaba, difonema, etc.)
- la variabilidad de la unidad
- la relación con modelos de producción de la onda sonora o con teorías lingüísticas

Como consideración general, se puede decir que en ningún caso se sigue solamente uno de ellos, y que las razones para decidir en favor de una u otra están fuertemente condicionadas por la aplicación que se desee dar al sistema y los criterios de flexibilidad, calidad y complejidad de manipulación de las unidades.

*El tamaño de la unidad*

Además de las unidades del habla tradicionales se cuenta también con unidades que han emergido como consecuencia de las necesidades de los sistemas de síntesis y que no habían aparecido en las primeras teorías lingüísticas. Se trata de

segmentos menores que la sílaba como la demisílaba y el difonema. La demisílaba es un fragmento de sílaba comprendido entre el inicio de la sílaba y el centro de la vocal que actúa como núcleo silábico o entre el centro de esta vocal y el final de la sílaba. La síntesis mediante concatenación de demisílabas ofrece la ventaja de reducir el inventario de unidades a utilizar (se calcula que si para una lengua natural el número de sílabas oscila entre 4000 y 10000, el número de demisílabas se reduce a 1000 [LLISTERRI, 1993]). Otra ventaja de la demisílaba es que permite una edición relativamente simple de sus parámetros temporales y facilita también la actuación sobre los elementos suprasegmentales del conjunto del enunciado. Por otra parte, desde el punto de vista de la síntesis a partir de un texto escrito en el que se opere una descomposición morfológica previa, permite tener almacenados de forma unitaria buena parte de los prefijos y sufijos más corrientes. Esta unidad es una de las recomendadas por [DUTOIT, 1996].

Los difonemas son segmentos formados por la parte estacionaria del primer elemento, la transición hacia el segundo y la parte estacionaria de éste último, de modo que puede considerarse un grupo de dos segmentos del que se excluye la transición inicial del primero y la transición final del segundo. La principal ventaja sobre otras unidades es que permite resolver los problemas derivados de la coarticulación, pues en las fronteras de los difonemas se encuentran únicamente estados estacionarios.

Finalmente, cabe considerar los "microfonemas", también llamados tramas, que no son más que fragmentos con características acústicas uniformes en los que puede dividirse un segmento sonoro; ésta es una noción que suele emplearse en el momento de convertir una representación fonética en el conjunto de parámetros acústicos necesarios para el control de la síntesis final.

*La variabilidad de la unidad*

Las realizaciones en la onda sonora de las unidades pequeñas como los fonemas ofrecen un alto grado de variabilidad ya que, debido a las características del mecanismo de producción del habla, éstas se hallan influidas por los segmentos adyacentes. Por tal razón su concatenación es más difícil pues, en las fronteras con las otras unidades (tanto si se trata de fonos, difonemas, demisílabas, sílabas o incluso morfemas) debe modelarse de alguna manera el proceso de coarticulación característico del habla natural. La falta de estudios fonéticos detallados sobre la producción del habla continua y sobre las características invariantes de las realizaciones de elementos abstractos como los fonemas es una de las razones que impiden a los sistemas de síntesis alcanzar la calidad deseada.

*La relación de las unidades con los modelos de producción de habla y con los modelos lingüísticos*

Si las unidades elegidas se definen en términos de modelos bien conocidos de producción de la voz (por ejemplo siguiendo la teoría de la fuente y el filtro tal como se lleva a cabo en la síntesis paramétrica por formantes), el control de la síntesis y la posibilidad de variación de los parámetros que la caracterizan en vistas a obtener una mayor calidad y flexibilidad son mucho más sencillos. Utilizando, por ejemplo, técnicas como la predicción lineal, es difícil relacionar directamente los coeficientes con parámetros conocidos; por ello, para la edición del habla es preciso muchas veces reconvertirlos en datos sobre frecuencia y amplitud de los formantes.

Al mismo tiempo, la utilización de unidades relacionadas con las que se emplean habitualmente en el análisis del habla facilita la extracción automática de los parámetros que las definen.

#### *2.3.2.2 Etiquetas prosódicas y fonéticas*

La información prosódica a almacenar puede ser de distinto tipo, de acuerdo a los métodos utilizados para obtenerla. Estos métodos implementan los modelos y teorías lingüísticas apropiadas para cada lengua. En [EAGLES, 1997] se nombra a los métodos más utilizados para el inglés: MARSEC, ToBI, IPO. [SOSA, 1991] propone una adaptación inicial al método ToBI para el español.

En el proyecto del LIS se propone usar el método ToBI, adaptado al español y extendido, para realizar la caracterización prosódica de las oraciones grabadas. Con respecto a los símbolos fonéticos se propone el método CSLU. Las etiquetas correspondientes a ambas caracterizaciones se obtienen utilizando un software con funciones específicas para esta tarea.

#### *2.3.2.3 Método de etiquetado prosódico ToBI*

ToBI [BECKMAN, 1994] es el sistema de codificación simbólica de la entonación para aplicaciones tecnológicas más utilizado entre los investigadores. En sus inicios fue definido inspirado en el idioma inglés y actualmente es el paradigma relevante para todas las lenguas.

Propone cuatro niveles de transcripción:

a) nivel tonal
b) nivel ortográfico
c) nivel de índices de grado de pausas

d) nivel de misceláneas

a) El nivel tonal considera dos tipos de eventos:

- Los acentos tonales, que se correlacionan con las sílabas tónicas.
- Los tonos de frase, que caracterizan la entonación del final de las frases.

Se utilizan dos símbolos para identificarlos: "H" para los tonos altos (máximos locales de la función de frecuencia fundamental) "L" para los tonos bajos (mínimos locales).

*Acentos tonales*

Para codificar los acentos tonales se alinea el contorno de F0 con la oración origen y se buscan los valores del contorno que se corresponden con las sílabas acentuadas. Es necesario también identificar los máximos y mínimos locales del contorno que no coinciden con las sílabas tónicas, pero las preceden o las siguen inmediatamente.

Los símbolos para codificar estos movimientos son:

- H* si la sílaba acentuada coincide con un máximo local o un segmento ascendente del contorno.
- L* si la sílaba acentuada coincide con un mínimo local o un segmento descendente del contorno.
- Una combinación de dos símbolos L o H concatenados con un signo +, para caracterizar las posibles secuencias formadas por acentos tonales y máximos y mínimos locales que preceden o anteceden a ese acento tonal.

El método original, desarrollado para el inglés propone cinco tipos de acento:
H*, L*, L*+H, L+H*, H+H*

La teoría autosegmental propone siete tipos de acento:
H*, L*, L*+H, L+H*, H*+L, H+L*, H*+H

La adaptación al español propuesta por el LIS considera ocho tipos de acentos:
H*, L*, L*+H, L+H*, H*+L, H+L*, H*+H, H+H*

Algunos ejemplos de estos tipos de acento pueden ser:

i) L+H* (acento en pico ascendente) cuando la sílaba tónica coincide con un máximo local que es precedido por un mínimo local, el cual no coincide con ninguna otra sílaba acentuada.
ii) L*+H cuando la sílaba tónica coincide con un mínimo local que es seguido por un máximo local, el cual no coincide con ninguna otra sílaba acentuada.

*Tonos de frase*

Para codificar los tonos de frase encontramos:

- L% o H% tono de frase que se encuentra al final de la emisión.
- %L o %H tono de frase al comienzo de la emisión.
- L- o H- acento de frase que ocurre al final de una frase, entre la última sílaba tónica y el tono de frase final. Por ejemplo, si existe un mínimo local entre la última sílaba acentuada y el tono de frase final es alto, la codificación es: L-H%. Si hay un máximo intermedio y el tono es bajo: H-L%.

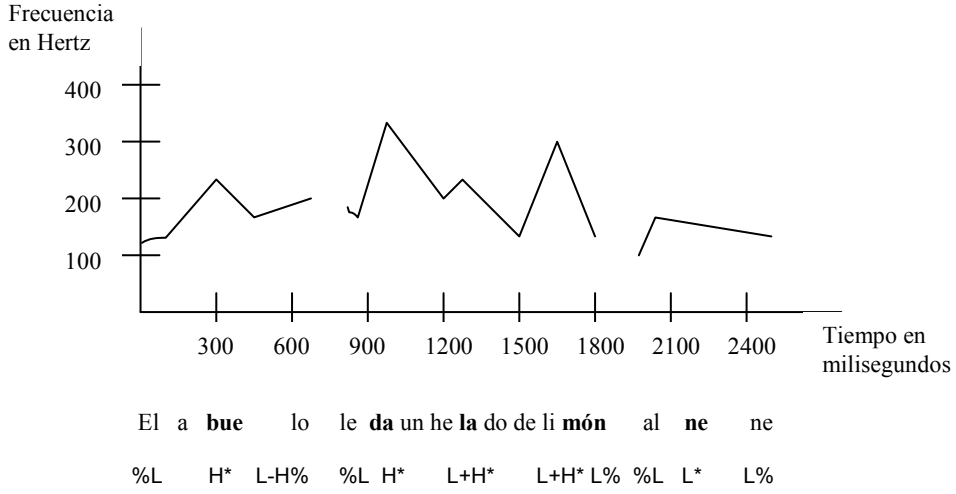


Figura 2.9. Ejemplo de etiquetado ToBI para el español.

*Extensiones al método ToBI definidas en el LIS para el español de Buenos Aires*

El método ToBI describe y codifica los componentes microprosódicos, pero no especifica los macroprosódicos: no diferencia entre tonos altos (o tonos bajos) de distintas alturas entre sí, lo que impide lograr una correcta comparación entre los contornos de entonación que generan similares cadenas codificadas, pero cuyos valores de frecuencia fundamental (F0) están en rangos diferentes de Hertz. Entonces

es difícil la automatización, pues se necesita la opinión de un experto lingüista para generar un contorno a partir de una cadena codificada.

Para enriquecer la descripción de los componentes macroprosódicos se tomaron los valores de los máximos y mínimos locales y se clasificaron en niveles.

El oído no es igualmente sensible a todas las frecuencias. Una diferencia de 25 Hertz es más perceptible si se produce entre los 100 y los 150 Hertz, que entre los 300 y los 350 Hertz [GUIRAO, 1980]. Para solucionar esta situación se deben realizar dos pasos. Primero se normalizan los valores de los máximos y mínimos y luego se determinan los niveles.

Para normalizar se transforman los valores en ERB- rate (equivalent rectangular bandwidth rate scale), cuya fórmula es:

$$Erb(x) = 16{,}7 * \log 10(1 + (x / 165{,}4))$$

con x: frecuencia en Hz
Erb(x): valor en ERB

Luego se clasifican los valores en los siguientes niveles (el rango de F0 se encuentra entre los 60 y 500 Hz aproximadamente en la voz masculina y entre los 80 y 700 Hz en la femenina):

| Valor en Hertz | Nivel |
|---|---|
| 20 | 1 |
| 40 | 2 |
| 70 | 3 |
| 110 | 4 |
| 150 | 5 |
| 190 | 6 |
| 240 | 7 |
| 300 | 8 |
| 370 | 9 |
| 450 | 10 |
| 540 | 11 |
| 650 | 12 |

Nótese que entre los 100 y los 200 Hz hay cuatro niveles, tres niveles entre los 200 y los 300 Hz y sólo dos entre los 300 y los 400 Hz, según la percepción del oído humano.

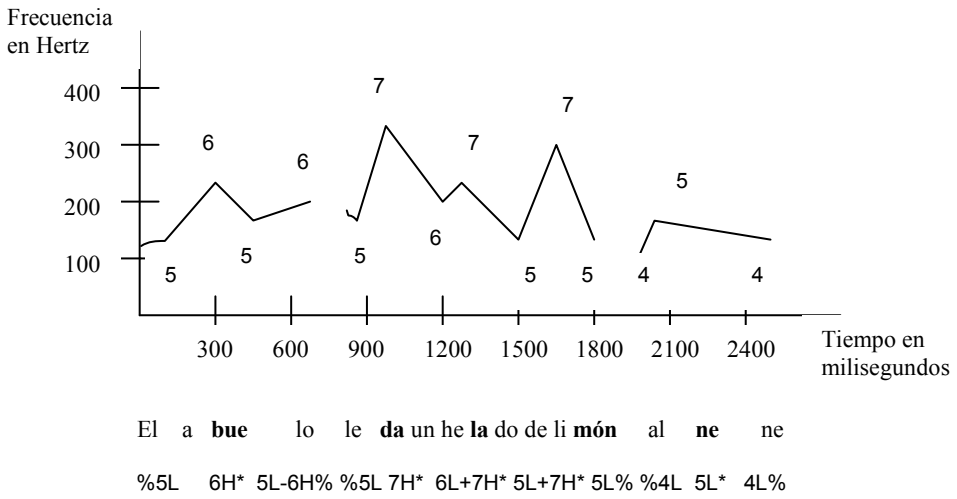


Figura 2.10. Extensión del LIS para el etiquetado ToBI del español.

### *2.3.2.4 Método de etiquetado fonético CSLU*

La guía de etiquetado del CSLU se utiliza como convención para la transcripción de datos fonéticos y fonológicos (o fonémicos), y para los segmentos no pertenecientes al habla. En este trabajo sólo se necesitan las guías de etiquetado fonético y fonológico pues no se trabaja con habla espontánea.

Incluye el método Worlbet, que establece un conjunto de caracteres ASCII para la transcripción de las realizaciones de los fonemas del español.

La alineación temporal se realiza en primer lugar con relación a la forma de onda, y luego teniendo en cuenta el espectrograma (realizado con una ventana pequeña), los contornos de energía y de frecuencia fundamental, vistos simultáneamente y utilizando reglas fijas en casos ambiguos.

# Capítulo 3 BASE DE DATOS PROSÓDICA

En este capítulo se presenta la propuesta de este trabajo conjuntamente con el marco interdisciplinario al que se circunscribe. Luego se describen los resultados del análisis realizado en el capítulo 2 que afectan directamente a la construcción de la base de datos y se detallan las etapas de construcción de la base de datos de acuerdo con la propuesta del LIS. A partir de ahí se describe la creación de un Corpus de Oraciones necesario para construir la base de datos, los pasos seguidos en el desarrollo de la base de datos propiamente dicha y finalmente los pasos seguidos para demostrar su utilidad en el entrenamiento de un conversor de texto a habla de alta calidad.

## *3.1. Propuesta*

Aquí se propone la construcción de una Base de Datos para almacenar características prosódicas y lingüísticas propias del español de Buenos Aires que será usada para entrenar a un sistema de conversión de texto a habla.

En el marco del proyecto interdisciplinario del LIS de desarrollo de un sistema de conversión de texto a habla de alta calidad para el español de Buenos Aires, se propuso que sea el presente trabajo de grado el ámbito de construcción de la base de datos prosódica.

Esta tarea no sólo implica el desarrollo de la base de datos sino también la integración al equipo del LIS y el desarrollo de otras herramientas de software necesarias durante la compleja construcción de la base de datos.

Dentro del equipo del LIS también se pretende aportar los conocimientos sobre tecnologías propias de la informática, parte integrante y necesaria de todo equipo interdisciplinario encargado de crear un sistema de conversión de texto a habla. Entre estos aportes se pueden citar: sugerencias y recomendaciones sobre automatización en el desarrollo de software, integración entre diversas tecnologías actuales, elección de plataformas para el desarrollo y puesta en marcha de productos de software, técnicas de reusabilidad y extensibilidad de aplicaciones preexistentes.

El equipo del LIS cuenta con un análisis general acerca de la construcción de la base de datos prosódica. Este análisis incluye un plan de trabajo a seguir cuyo documento original se adjunta en el apéndice. Se detallan más adelante los pasos de este plan con el objetivo de brindar el contexto de construcción de la base de datos.

Para adecuar el análisis general realizado por el equipo del LIS a las necesidades particulares de la base de datos prosódica se propone realizar un estudio minucioso de los sistemas de conversión de texto a habla que utilizan estas bases de datos (capítulo 2). Estos sistemas se basan en teorías y modelos lingüísticos durante su fase de análisis del texto a ser convertido. Entonces se deben estudiar las teorías más apropiadas para el español de Buenos Aires, y sus métodos representativos para que pueda definirse la información a almacenar. Como los sistemas de conversión de texto a habla necesitan de un entrenamiento prosódico basado en diferentes estrategias de interacción con la base de datos, también se deben entender estas estrategias para definir los procesos y consultas a desarrollar durante este trabajo.

Se propone un diseño orientado a objetos para crear el modelo de datos, debido a la flexibilidad necesaria en este tipo de desarrollos. Las estructuras de datos que se implementen deben permitir almacenar gran variedad de oraciones con sus rasgos prosódicos, sus características acústicas y lingüísticas. Para las oraciones se deben aceptar transcripciones ortográficas, fonológicas, fonéticas y prosódicas que representen unidades segmentales del habla como fonemas y demisílabas y también las características de entonación, ritmo y acentuación del español. Para las sílabas se propone además una representación fonológica con las distintas situaciones de acento y ubicación dentro de la palabra.

En la etapa de prueba, antes de utilizar un sintetizador para intentar alcanzar la alta calidad, se propone un plan tentativo de pasos elementales que deben ser parte del entrenamiento. En estos pasos se obtienen de la base de datos los rasgos prosódicos adecuados para emitir el sonido deseado. Se propone también cargar las oraciones necesarias para probar el funcionamiento de la base de datos.

Además se propone, como aporte a la creación del Corpus de Oraciones citado más abajo en la propuesta del LIS, el desarrollo de los siguientes productos de software:

**Aplicación Sílabas:** software para transcripción automática de oraciones, palabras ortográficas, y sílabas fonológicas clasificadas de acuerdo a la situación de acento y ubicación dentro de la palabra. Se pretende su utilización como parte de un método semiautomático de obtención de un Corpus de Trabajo con sus rasgos ortográficos y fonológicos característicos.

**Base de Datos Corpus:** base de datos para crear un corpus de oraciones. Debe permitir almacenar sílabas fonológicas y su frecuencia de uso en el español de Buenos Aires, un corpus de oraciones de trabajo y también todas las palabras de la Real Academia Española. Deben ser sus funciones: comparar conjuntos de sílabas, generar la información estadística correspondiente y depurar oraciones sobre la base de procesos estadísticos.

*Propuesta del LIS acerca de la construcción de la base de datos prosódica*

La construcción de una base de datos completa, totalmente funcional, representativa del español de Buenos Aires, propuesta por el equipo del LIS, puede resumirse en las siguientes etapas:

- Crear un Corpus de Oraciones estructuralmente representativo de nuestra lengua.
- Desarrollar la Base de Datos Prosódica.
- Cargar el corpus de oraciones en la base de datos.
- Grabar las oraciones del corpus y cargar los datos de las emisiones en la base de datos.
- Analizar y etiquetar las emisiones grabadas, y cargar su producto en la base de datos.
- Probar la base de datos.

A su vez la creación del Corpus de Oraciones involucra:

- Recolectar material de trabajo:
  - obtener las sílabas y su frecuencia de uso en el español de Buenos Aires;
  - obtener corpus de oraciones de trabajo;
  - obtener todas las palabras de la Real Academia Española.
- Desarrollar un método semiautomático para obtener las características ortográficas y fonológicas del material recolectado.
- Crear una base de datos de trabajo para almacenar el material obtenido en los dos pasos anteriores debiendo tener, además, la funcionalidad necesaria para realizar el paso siguiente.
- Cargar el material recolectado en la base de datos corpus.
- Depurar el corpus de oraciones de trabajo: eliminar las oraciones recolectadas redundantes y agregar las necesarias hasta llegar a obtener el Corpus de Oraciones.

## *3.2. Resultados específicos del análisis*

En estos párrafos se detallan los resultados del análisis que afectan directamente al desarrollo de la base de datos prosódica, incluyendo algunas consideraciones de entorno, contenido, funcionalidad general, funciones específicas a cumplir durante el entrenamiento y definiciones realizadas como parte de este trabajo.

### 3.2.1 Consideraciones generales sobre la base de datos prosódica

#### *3.2.1.1 Entorno*

En el siguiente diagrama se muestran los procesos típicos que constituyen el entorno de la base de datos. Muchos de estos procesos están integrados en aplicaciones que realizan varias tareas a la vez. Estas aplicaciones, en la mayoría de los casos, trabajan independientemente de la base de datos, brindando sus resultados a través de archivos. Sin embargo, con las funcionalidades definidas en ella, pueden automatizarse muchas tareas. También se pueden vincular las aplicaciones con la base de datos de manera que, sus resultados sean allí introducidos y, en algunos casos también verificados con la ventaja de poder, desde la aplicación que está corriendo, interactuar con el usuario y corregir los errores, por ejemplo, de inconsistencia.

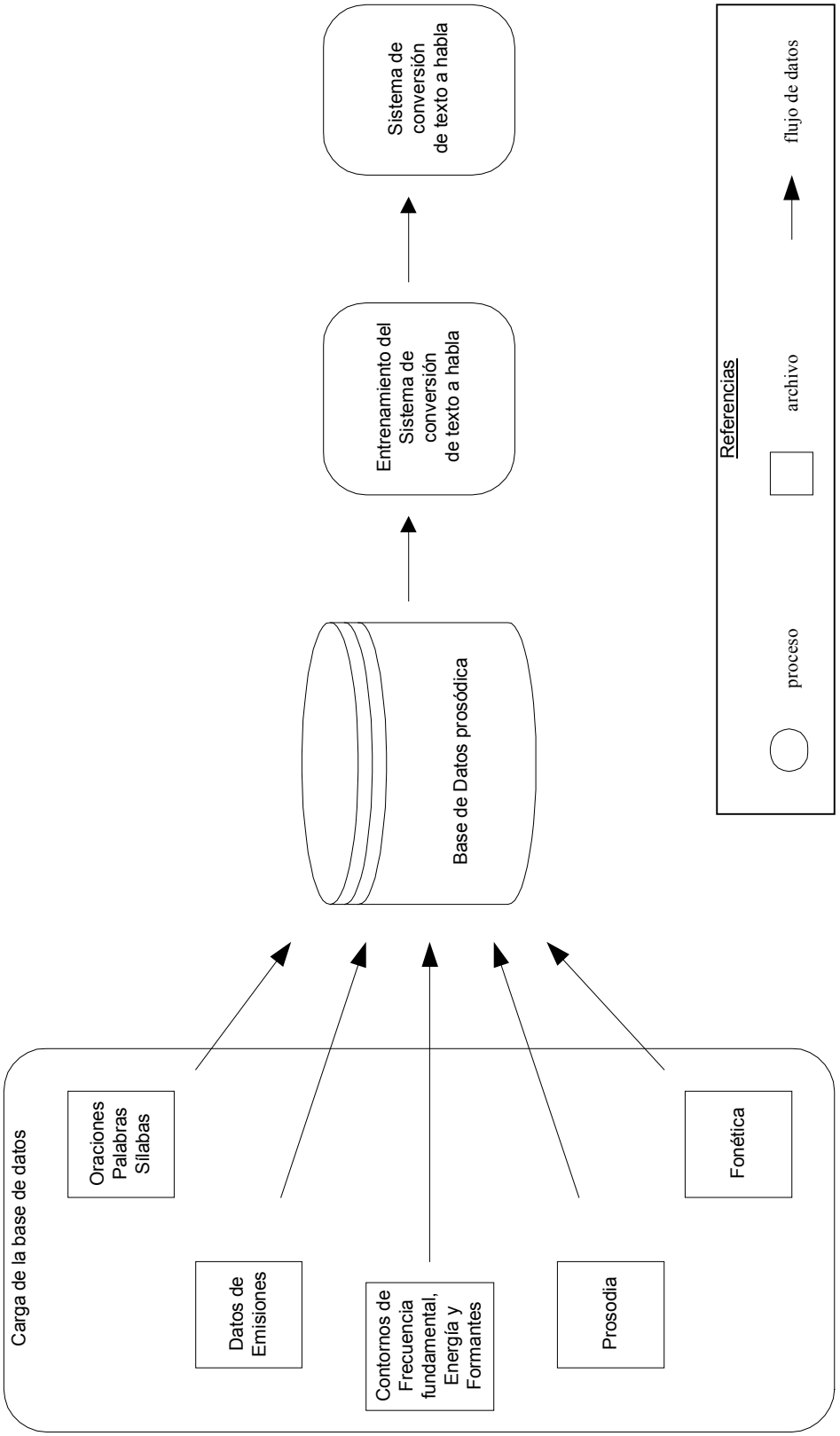


Figura 3.1. Entorno de la base de datos prosódica.

### *3.2.1.2 Definiciones de sílaba fonológica contextualizada y parte de oración*

Se define como sílaba fonológica contextualizada, en adelante "sfc", a la posibilidad de ubicación (inicio, medio y fin) de una sílaba fonológica (acentuada o no) en el contexto de una palabra. Pueden identificarse siete combinaciones de aparición de una sfc en una palabra:

| | | | | | | |
|---|---|---|---|---|---|---|
| monosílabo | acentuada al inicio | acentuada en el medio | acentuada al final | no acentuada al inicio | no acentuada en el medio | no acentuada al final |

Se define parte de oración, en adelante "pado", a una parte de un todo que es el texto de la oración. Una sucesión de palabras no incluye los signos de puntuación que aparecen en la oración a la que pertenecen, pero una pado sí los puede incluir.

Ejemplos:

En la oración "La casa, de todas formas, es linda" se puede reconocer "casa, de" como una pado. Y además la secuencia de palabras: "casa"; "de".

Mientras que en "La casa de mi tía no tiene balcón" aparece la secuencia de palabras anterior pero no la pado "casa, de".

Estas definiciones permiten agregar a los conceptos de conocimiento general (oración, palabra y sílaba) y a los específicos de la lingüística (fono, fonema, morfema, sílaba fonológica), el concepto de unidades no predeterminadas de la oración (pados) y el de abstracción fonológica de una sílaba en su contexto ortográfico (sfc).

### *3.2.1.3 Contenido*

Con respecto a las unidades del habla a ser almacenadas se decidió utilizar sfc y fonos pues son las que permiten trabajar también con unidades mayores como difonos, trifonos y demisílabas. Al ser el fono la menor unidad en que puede segmentarse la emisión acústica, por ejemplo, pueden obtenerse difonos con la posición del borde entre los fonos que son parte de él.

### *3.2.1.4 Funcionalidad*

A la hora de definir las funciones de la base de datos se planteó la necesidad

de ser usada por personas ajenas al campo de la informática. Entonces se definieron procesos que sean fácilmente utilizables y cuyas funciones sean resolver, en forma transparente al usuario, las cuestiones relacionadas con la representación interna de los datos y con el flujo de la información. De esta forma, las personas responsables de la carga de información, los lingüistas que realizan estudios sobre entonación y ayudan a crear las estrategias que se usarán en la conversión on-line de texto a habla, y las personas encargadas del entrenamiento del conversor, se encuentran con una serie de procesos que les permiten realizar su trabajo.

En particular, con respecto a las funciones de carga de la base de datos se planteó la necesidad de automatización que se espera de ellas. Para intentar lograr la máxima automatización posible se definieron procesos que pueden ser invocados fácilmente desde programas externos a la base de datos o desde cualquier herramienta de interacción con bases de datos. Estos procesos se ocupan de invocar a otros procesos encargados de cada tarea específica sin interacción con el usuario; por lo tanto se ocupan de garantizar la integridad de la información que manipulan.

En las sucesivas charlas con los lingüistas del LIS se analizaron los posibles usos que se le dará a la base de datos prosódica en su especialidad, independientemente de los requerimientos necesarios para el entrenamiento del conversor de texto a habla. Durante estas reuniones surgieron algunas de las funciones que se especifican más adelante.

#### *3.2.1.5 Entrenamiento del conversor de texto a habla*

Antes de explicar las consideraciones referidas al entrenamiento es necesario decir que, durante la realización de este trabajo, se decidió que el sistema de conversión de texto a habla a desarrollarse en el LIS cuente con una base de datos complementaria, en la cual se almacene indispensablemente la siguiente información:

1_ Unidades del habla, representadas a través de parámetros o codificadas de la misma forma que los archivos de onda tipo WAV.
2_ Un diccionario de ítems lexicales con las categorías sintácticas y morfológicas correspondientes.
3_ Un conjunto de estructuras oracionales y los diferentes patrones prosódicos asociados a ellas más la información probabilística que ayude a elegir el patrón más adecuado para cada estructura oracional.

La mayoría de estos datos deben obtenerse de la base de datos prosódica durante el entrenamiento.

El entrenamiento del sistema de conversión de texto a habla usando la base de

datos prosódica debe servir para ayudar a definir y elegir las estrategias a usar en la conversión on-line y los datos que dichas estrategias usarán. Un resumen de estas tareas es:

*Con respecto a las estrategias que se usarán en la conversión on-line:*

- Definir diferentes estrategias de generación automática de marcadores prosódicos, que se obtendrán a partir del texto, durante la conversión on-line.
- Definir diferentes estrategias de generación automática de contornos prosódicos, que se derivan de los marcadores prosódicos anteriores.
- Elegir entre las posibles estrategias definidas para generar las características prosódicas y también entre las ya conocidas estrategias para generar características fonéticas, cuáles son las más convenientes para el español de Buenos Aires.
- Elegir las unidades del habla (fonos, demisílabas, difonos, trifonos, etc.) y su modo de almacenamiento (parámetros, WAV) en la base de datos complementaria. Estas unidades serán usadas por las estrategias de generación de características fonéticas.
- Depurar y poner a punto las estrategias elegidas.

*Con respecto a los datos a incluir en la base de datos complementaria:*

- Obtener estructuras oracionales y los patrones prosódicos asociados a ellas.
- Obtener unidades del habla y sus características fonéticas.

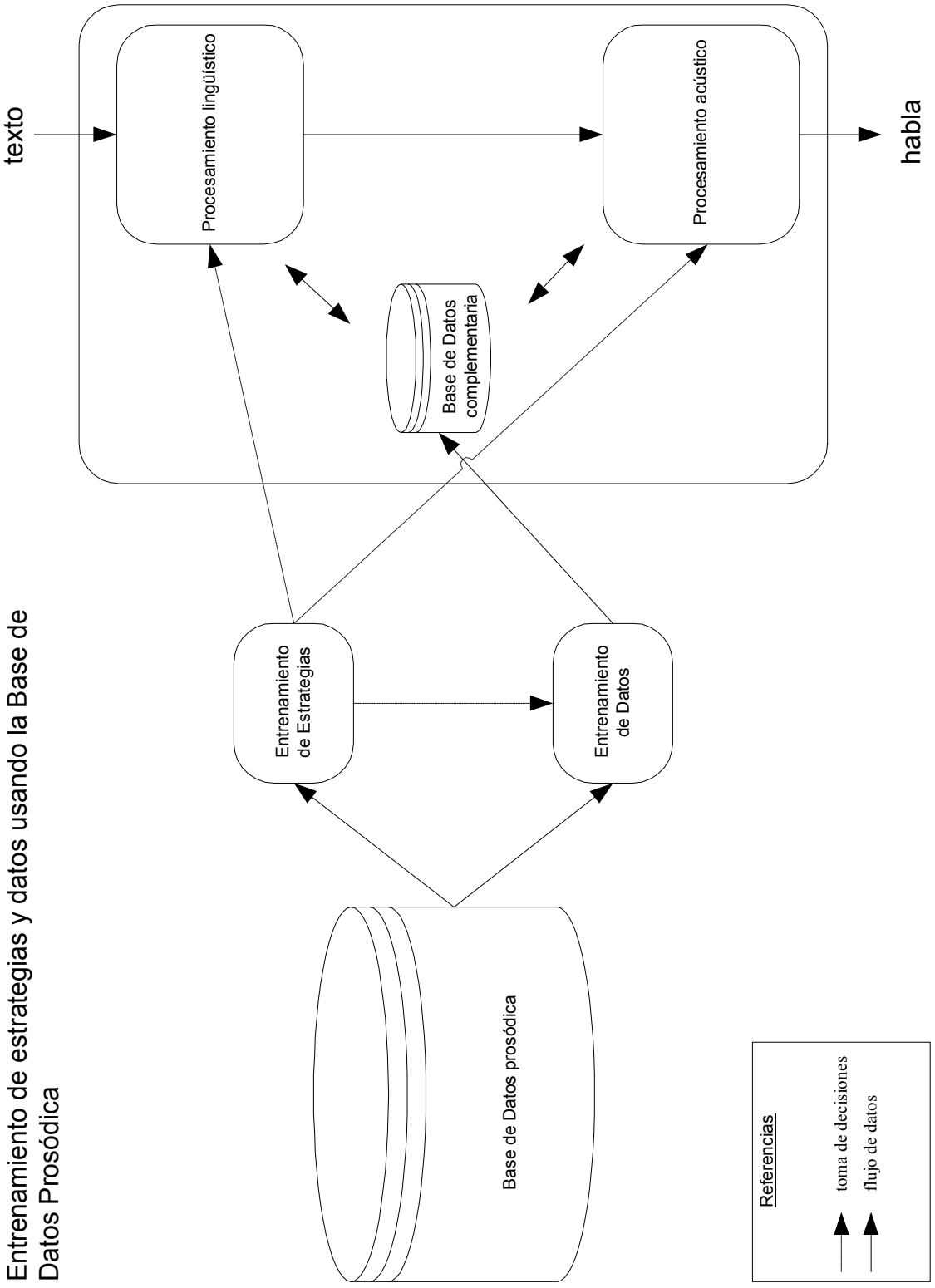


Figura 3.2. Entrenamiento del conversor de texto a habla.

### 3.2.2 Especificación de requerimientos de la base de datos prosódica

Aquí se detalla la información que la base de datos debe almacenar y la funcionalidad que se requiere de ella.

#### *3.2.2.1 Información de entrada*

La base de datos contiene la siguiente información asociada a cada oración:

- Texto de la oración, incluyendo los símbolos de puntuación.
- Transcripción ortográfica de las palabras que la forman.
- Transcripción fonológica de las sílabas que aparecen en cada palabra, con información contextual referente a la ubicación y situación de acento. Cada sílaba puede estar ubicada al principio, en el medio o al final de una palabra, o en posición neutral en el caso de palabras monosilábicas. Se indica acentuación tanto para los acentos ortográficos como para los prosódicos.

También contiene datos de emisiones previamente grabadas, las cuales se corresponden con alguna de las oraciones anteriores. Cada oración puede tener una o más emisiones a través de las que se realiza acústicamente, como producto del habla de un locutor.

De cada emisión se tiene:

- Nombre y apellido del hablante.
- Nombre y ubicación del archivo de onda (por ejemplo c:\bdp\audio\onda40.wav), fecha en que se realizó la grabación, tipo de codificación (WAV, AU, etc.), frecuencia de muestreo (11.025 Hz, 22.050 Hz, 44.100 Hz, etc.), nivel de muestreo para representar la altura tonal (8, 16), tipo de canal (mono, estéreo).
- Fonos observados, con los instantes de inicio y de fin correspondientes a cada uno, expresados en milisegundos. Los fonos se representan a través de símbolos escritos en ASCII. En este trabajo se utilizan aquellos definidos en las reglas de transcripción fonética CSLU, que no exceden los dos caracteres. Sin embargo pueden almacenarse otros símbolos.
- Palabras detectadas, con la información prosódica asociada a cada una.
- Ocho contornos parametrizados correspondientes a frecuencia fundamental (F0), energía (E) y formantes (F1,…,F6).

La información prosódica asociada a cada palabra pronunciada en una emisión, de acuerdo a las reglas de transcripción prosódica ToBI, es la siguiente:

- Transcripción tonal, representada por una secuencia de símbolos escritos en ASCII, cuya longitud máxima es de veinte caracteres.
- Indicación de foco entonacional, en caso de que corresponda.
- Instante temporal en el que finaliza, expresado en milisegundos.
- Tipo de pausa que se produce al final de la palabra, representada por un número del 1 al 4, donde 1 indica la pausa más breve.

La base de datos debe cargarse en el siguiente orden:

1) Oraciones con sus palabras y sílabas asociadas.
2) Datos de emisiones.
3) Transcripción prosódica, transcripción fonética y contornos parametrizados de cada emisión.

#### *3.2.2.2 Formato y tipo de archivos de entrada*

La información a cargar en la base de datos se encuentra en archivos de texto, con formato ASCII. Los campos tienen longitud variable y se encuentran delimitados por el carácter "|".

*Archivos de oraciones, palabras y sílabas*

Estos archivos contienen el texto de cada oración a ser incorporada en la base de datos y los detalles de las transcripciones ortográficas de las palabras y fonológica de las sílabas. Su formato ha sido definido durante el desarrollo del presente trabajo.

Los campos abajo nombrados que contienen número de oración o número de emisión son del tipo numérico (6), donde 6 indica la cantidad de dígitos. Los que contienen número de orden son del tipo numérico (2).

Formato del archivo de oraciones:
campo 1: número de oración;
campo 2: texto de la oración, alfanumérico (255)

Formato del archivo de palabras:
campo 1: número de oración, correlacionado con el campo 1 del archivo de oraciones;
campo 2: número de orden de la palabra dentro de la oración;

campo 3: texto de la palabra, alfanumérico (30)

Formato del archivo de sílabas:
campo 1: número de oración, correlacionado con el campo 1 del archivo de oraciones;
campo 2: número de orden de la palabra dentro de la oración, correlacionado con el campo 2 del archivo de palabras;
campo 3: número de orden de la sílaba dentro de la palabra;
campo 4: texto de la sílaba, alfanumérico (7);
campo 5: situación de acento en el contexto de la palabra aludida en el campo 2 (S: si, N: no);
campo 6: ubicación en el contexto de la palabra aludida en el campo 2 (I: inicio, M: medio, F: final, N: neutral)

Ejemplo del archivo de oraciones:

1|La casa es linda.|
2|El animal come mucho.|

Ejemplo del archivo de palabras:

1|1|la|
1|2|casa|
1|3|es|
1|4|linda|
2|1|el|
2|2|animal|
2|3|come|
2|4|mucho|

Ejemplo del archivo de sílabas

1|1|1|la|S|N|
1|2|1|ka|S|I|
1|2|2|sa|N|F|
1|3|1|es|S|N|
1|4|1|lin|S|I|
1|4|2|da|N|F|
2|1|1|el|S|N|
2|2|1|a|N|I|
2|2|2|ni|N|M|
2|2|3|mal|S|F|
2|3|1|ko|S|I|
2|3|2|me|N|F|
2|4|1|mu|S|I|
2|4|2|tso|N|F|

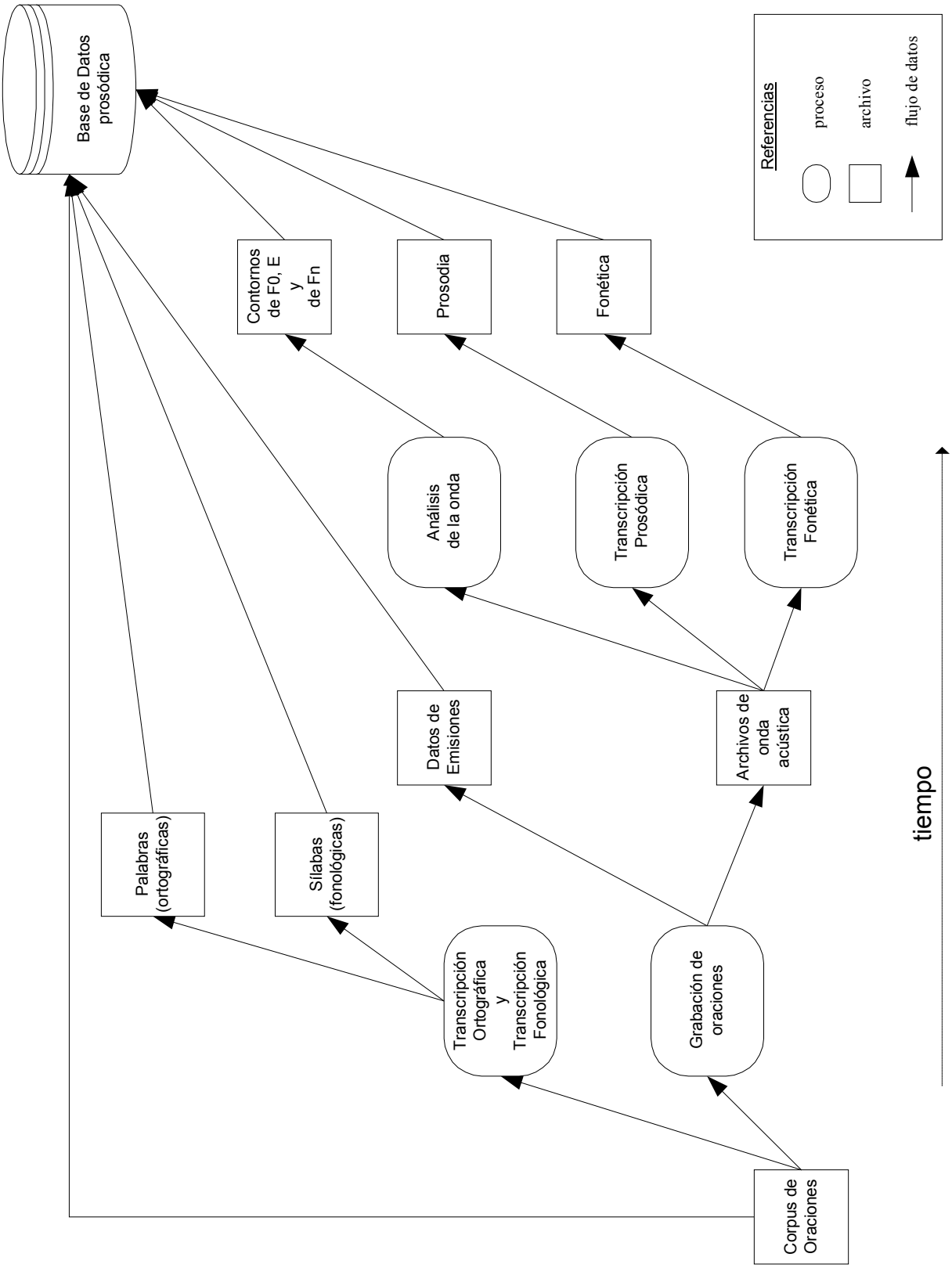


Figura 3.3. Carga de la base de datos prosódica.

*Archivo con datos de emisiones*

Contiene información acerca de uno o más archivos de onda, sus características de grabación y las oraciones pronunciadas en el momento de la grabación.

Formato del archivo de emisiones:
campo 1: texto de la oración pronunciada, alfanumérico (255);
campo 2: nombre y ubicación, dentro del sistema de archivos, del archivo de onda, alfanumérico (255);
campo 3: nombre y apellido del hablante, alfanumérico (40);
campo 4: fecha de grabación;
campo 5: tipo de codificación (WAV, AU, etc.), alfanumérico (3);
campo 6: frecuencia de muestreo en Hertz (11.025, 22.050, 44.100, etc.), numérico (5);
campo 7: nivel de muestreo (8, 16);
campo 8: tipo de canal (M: mono, E: estéreo)

Ejemplo del archivo de emisiones:

La casa es linda.|c:\bdp\audio\onda40.wav|Juan Perez|21/12/1999|WAV|22.050|16|M|
El gato saltó.|c:\bdp\audio\onda56.wav|Juan Perez|19/12/1999|WAV|22.050|16|M|

*Archivos con datos de transcripciones prosódicas*

Estos archivos contienen información prosódica de emisiones. Se tienen dos archivos: en uno se indica la emisión a partir de la cual se hizo la transcripción prosódica y en el otro se detalla la prosodia correspondiente a cada palabra encontrada en la emisión.

El formato de estos archivos está basado en las recomendaciones de etiquetado ToBI [BECKMAN, 1994].

Formato del archivo referente a emisiones:
campo 1: número de emisión;
campo 2: nombre y ubicación del archivo de onda, alfanumérico (255)

Formato del archivo con prosodia de palabras:
campo 1: número de emisión, correlacionado con el campo 1 del archivo anterior;
campo 2: transcripción ortográfica de cada palabra detectada en la emisión, alfanumérico (7);
campo 3: transcripción tonal de la palabra en la emisión, pudiendo no indicarse ningún valor, alfanumérico (20);
campo 4: foco entonacional ('S' o 'N');

campo 5: índice de pausa, numérico (1);
campo 6: instante de finalización, numérico (6)

Ejemplo del archivo referente a emisiones:

```
1|c:\bdp\audio\onda40.wav|
2|c:\bdp\audio\onda56.wav|
```

Ejemplo del archivo con prosodia de palabras:

```
1|la|%2L|N|1|120|
1|casa|6H*|S|1|800|
1|es||N|1|1100|
1|linda|4H*-2L%|N|4|1800|
2|el|%2L|N|1|140|
2|animal|7H*|S|1|1300|
2|come|3H*|N|1|1900|
2|mucho|L%|N|4|2400|
```

*Archivos con datos de transcripciones fonéticas*

Estos archivos contienen rasgos fonéticos de emisiones. Se tienen dos archivos: en uno se indica la emisión a partir de la cual se hizo la transcripción fonética y en el otro se detallan los fonos detectados, incluyendo el instante de inicio y el de fin de cada uno de éstos.

El formato de estos archivos está basado en las guías de etiquetado [CSLU, 1993].

Formato del archivo referente a emisiones:
campo 1: número de emisión;
campo 2: nombre y ubicación del archivo de onda, alfanumérico (255)

Formato del archivo de fonos:
campo 1: número de emisión, correlacionado con el campo 1 del archivo anterior;
campo 2: instante de inicio del fono, numérico (6);
campo 3: instante de fin del fono, numérico (6);
campo 4: símbolo del fono, alfanumérico (2)

Ejemplo del archivo referente a emisiones:

```
1|c:\bdp\audio\onda40.wav|
2|c:\bdp\audio\onda56.wav|
```

Ejemplo del archivo de fonos:

1|20|75|1|
1|70|157|a|
1|155|230|k|
1|230|325|a|
1|322|428|s|
1|420|480|a|
(continúa)

### *Archivos con datos de contornos parametrizados*

Estos archivos contienen los parámetros obtenidos de la emisión, a través del análisis de la onda. Se tienen dos archivos: en uno se indica la emisión a partir de la cual se derivaron los contornos parametrizados y en el otro se detallan los valores de cada contorno incluyendo el instante temporal, expresado en milisegundos, en que halla ese valor.

Formato del archivo referente a emisiones:
campo 1: número de emisión;
campo 2: nombre y ubicación del archivo de onda, alfanumérico (255)

Formato del archivo de contornos parametrizados:
campo 1: número de emisión, correlacionado con el campo 1 del archivo anterior;
campo 2: instante temporal, numérico (6);
campo 3: valor de F0, numérico (4);
campo 4: valor de E, numérico (2);
campo 5: valor de F1, numérico (4);
campo 6: valor de F2, numérico (4);
campo 7: valor de F3, numérico (4);
campo 8: valor de F4, numérico (4);
campo 9: valor de F5, numérico (4);
campo 10: valor de F6, numérico (4)

Ejemplo del archivo referente a emisiones:

1|c:\bdp\audio\onda40.wav|
2|c:\bdp\audio\onda56.wav|

Ejemplo del archivo referente a contornos parametrizados:

1|10|0|20|400|1000|2500|3400|4500|0|
1|20|100|30|600|900|2600|3700|0|0|
1|30|130|25|650|800|2700|3900|4400|5300|
(continúa)

### *3.2.2.3 Interfaz con el exterior y funcionalidad*

Además de existir la posibilidad de utilizar la base de datos a través del lenguaje de consulta y manipulación que proporciona todo DBMS, la base de datos debe contar con la funcionalidad necesaria que requieren aquellas personas que son responsables de las siguientes tareas: carga de la base de datos y entrenamiento de un conversor de texto a habla.

*Funciones a cumplir durante la carga de datos*

La base de datos debe permitir la introducción, modificación y borrado de la siguiente información, cuya carga se realizará en el siguiente orden:

1) Oraciones, palabras ortográficas y sílabas fonológicas.
2) Datos de las emisiones correspondientes a los archivos de onda grabados.
3) Contornos de frecuencia fundamental, energía y formantes a través de sus parámetros representativos.
4) Marcadores o etiquetas prosódicas y fonéticas asociados a las emisiones.

Todas esta tareas deben poder realizarse de manera automática invocando a los procesos pertinentes. Esta característica permite, a los encargados de realizar cada tarea, desentenderse de la problemática de introducción de cada uno de los datos y también de la forma en que ellos se distribuyen y agrupan dentro de la base de datos.

*Funciones asociadas el entrenamiento*

La base de datos debe poder entregar la información solicitada durante el entrenamiento, ya sean los datos que se almacenarán en la base de datos complementaria (ver Entrenamiento del conversor de texto a habla explicado anteriormente) o la información prosódica y fonética necesaria para definir, elegir y depurar las estrategias que usará el conversor durante el trabajo on-line. Los procesos que realicen estas tareas deben ser de fácil uso y lo suficientemente variados como para cumplir con la totalidad de requerimientos que el entrenamiento demanda.

Entre las funciones principales que debe realizar se encuentran aquellas que retornan los siguientes datos:

- Palabras, sílabas fonológicas y sfc asociadas a las oraciones.
- Marcadores prosódicos y fonéticos asociados a las emisiones.
- Contornos parametrizados de frecuencia fundamental (F0) asociados a las emisiones.
- Contornos parametrizados de energía (E) asociados a las emisiones.
- Contornos parametrizados de los formantes (F1, ... , F6) asociados a las emisiones.

Deben contemplarse respuestas que combinen dichos datos, y además, considerando las definiciones de sfc y pado ya explicadas en esta sección, se pueden plantear procesos que respondan tanto a solicitudes basadas en datos de emisiones y en unidades predeterminadas de uso cotidiano (oraciones, palabras y sílabas), como a las basadas en sfc y pado.

#### *3.2.2.4 Restricciones asociadas a los datos y su manipulación*

Las restricciones que se imponen a los datos sirven para mantener la lógica de la información almacenada o integridad de la base de datos. Entre las principales restricciones a cumplir se encuentran:

- Cada oración contiene una o más palabras (palabras ortográficas), y toda palabra tiene una o más oraciones en las que se encuentra. A su vez una palabra puede tener una o más sílabas asociadas, y toda sílaba tiene una o más palabras en las que se encuentra contextualizada. Además una sfc puede aparecer en más de una palabra.
- Cada oración puede tener una o más emisiones en las que se la pronuncia (se realiza acústicamente a través del habla), pero toda emisión tiene una sola oración asociada.
- Al eliminar una oración se deben eliminar también sus palabras y sílabas. De ninguna otra manera se debe permitir el borrado de palabras ni de sílabas.
- Al eliminar una emisión de deben eliminar también los datos prosódicos, fonéticos y los parámetros de los contornos asociados a ella.

## *3.3. Corpus de Oraciones*

Un Corpus de Oraciones es un conjunto de oraciones escritas que cumple con características específicas dependientes del proyecto para el cual es construido. En el ámbito de las tecnologías del habla todo corpus se utiliza para ser grabado y etiquetado (caracterizado desde diversos puntos de vista lingüístico) convirtiéndose en un Corpus de Habla. En los tiempos actuales, la mayoría se encuentra contenido en una Base de Datos, que amplía su funcionalidad y ayuda a su manipulación.

Desde hace años existe una amplia variedad de corpus de habla, construidos con propósitos verdaderamente distintos (ATIS, MBROLA, VERBMOBIL, etc.). Muchos de ellos están orientados a tecnologías específicas y otros son más generales. En algunos se comenzó recolectando habla espontánea y luego se hicieron las transcripciones deseadas: ortográfica, fonética, fonológica, sintáctica, prosódica, etc., mientras que en otros se construyó inicialmente un corpus de texto, que luego fue grabado y finalmente etiquetado. En el ámbito de los sistemas de producción de habla (síntesis, sistemas de conversión de texto a habla, etc.) se privilegia el control prosódico, que puede llevarse a cabo durante las grabaciones, por sobre la variedad infinita intrínseca del habla espontánea.

En el proyecto del LIS se propone crear un corpus de oraciones con aproximadamente mil quinientas sentencias distribuidas de la siguiente manera: quinientas sentencias gramaticales declarativas con palabras simples, quinientas sentencias gramaticales interrogativas, ciento sesenta sentencias gramaticales complejas (con sintaxis más compleja y palabras polisilábicas), etc. La propuesta del LIS se encuentra en el apéndice.

Durante este trabajo se brindó el apoyo informático necesario para la creación del Corpus de Oraciones. En los comienzos se participó activamente en las deliberaciones del equipo del LIS. Allí se decidió crear un **Corpus de Oraciones, lo más pequeño posible, representativo del español hablado en Buenos Aires**, que en este caso es el elemento primario y fundamental para el desarrollo y construcción de la Base de Datos Prosódica.

Entre las características más importantes y novedosas que se propusieron se destacan la búsqueda de un pequeño número de oraciones y la representación fonológica de las sílabas, en distintas situaciones de acento y ubicación dentro de la palabra.

También, durante la tarea de análisis del Corpus de Oraciones, se definieron las dos etapas que deben cumplimentarse para su construcción. La primer etapa involucra las siguientes tareas que fueron llevadas adelante, simultáneamente, por distintos grupos de trabajo del LIS: recolección del material de trabajo, desarrollo de

un método semiautomático para obtener las características ortográficas y fonológicas del material recolectado, creación una base de datos de trabajo para almacenar el material obtenido en los dos pasos anteriores que, además tenga la funcionalidad necesaria para poder realizar el paso siguiente. La segunda etapa consiste en cargar en la base de datos, creada en la etapa anterior, todo el material recolectado.

El aporte de este trabajo de grado durante la creación del corpus de oraciones consistió, no sólo en la participación antedicha sino también en el desarrollo de los siguientes productos de software: **Aplicación Sílabas** y **Base de Datos Corpus**. El primero, también llamado "transcriptor silábico fonológico", surgió a partir de la necesidad de derivar las sílabas fonológicas contenidas en las palabras y de la búsqueda, sin éxito, de algún software similar que realizara la transcripción fonológica adecuada, informando la situación de acento y ubicación. La Base de Datos Corpus está especialmente construida para ser usada durante esta etapa solamente, sin embargo no se descartan futuros usos que puedan llegar a darle otros grupos de trabajo, especialmente aquellos con objetivos dentro del campo de la lingüística.

A continuación se detallan las tareas realizadas en cada una de las dos etapas.

### 3.3.1 Etapa 1: recolección del material de trabajo

Durante esta etapa se llevó a cabo la recolección de la materia prima para la obtención del corpus de oraciones buscado.

Se propuso entonces un método semiautomático para obtener las características ortográficas y fonológicas de un corpus de oraciones de trabajo: el primer paso consistió en obtener archivos de texto cuyo contenido se desea desmenuzar y caracterizar. El segundo paso es el que realiza automáticamente la aplicación "Sílabas", que genera un archivo con las oraciones que se detectan en el texto, un archivo de palabras contenidas en esas oraciones y un archivo de sílabas. Este último contiene las sílabas fonológicas derivadas de sus homónimas ortográficas más la clasificación según situación de acento y ubicación dentro de la palabra.

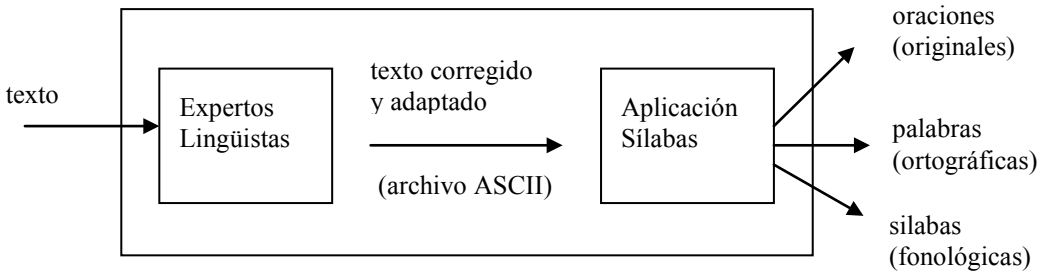


Figura 3.4. Método semiautomático para obtener oraciones, palabras y sílabas.

La adquisición del material recolectado involucró las siguientes tareas en función de sus distintos orígenes:

a) Obtener las sílabas usadas en el español de Buenos Aires y su frecuencia de uso.

Esta tarea se realizó a partir del trabajo de [GUIRAO, 1993], tomado como base para la obtención de las sílabas más usadas en el español de Buenos Aires. En particular se extrajeron las sílabas fonológicas allí presentadas junto con su frecuencia de aparición. Cabe aclarar que se decidió trabajar con sílabas fonológicas pues ellas constituyen la abstracción mental correspondiente a las sílabas ortográficas y fonéticas, de manera que son las más apropiadas para este tipo de desarrollo.

b) Obtener un corpus de oraciones de trabajo de aproximadamente 1500 oraciones.

La recolección de este corpus de trabajo la hizo un grupo de expertos lingüistas, quienes produjeron alrededor de mil quinientas oraciones tomadas mayormente de los diarios Clarín y La Nación, abarcando diversos temas como economía, política, turismo, deportes, sociedad, etc. Entre ellas debían seleccionarse fundamentalmente las declarativas simples. Aunque también debían incluirse oraciones complejas, interrogativas y exclamativas. Además debieron eliminarse los signos de puntuación y símbolos no necesarios en esta etapa. Entre otras tareas se debieron convertir los símbolos numéricos y las abreviaturas en texto escrito, y se eligieron principalmente oraciones de hasta veinte sílabas, de acuerdo con el criterio que indica que esa es normalmente la cantidad de sílabas que pronunciamos en cada grupo respiratorio.

Las oraciones producidas contienen frases entonacionales y frases intermedias, que son las necesarias para representar las variaciones prosódicas representativas de una lengua. Así, algunas oraciones pueden contener dos o más frases entonacionales y a su vez una frase entonacional puede contener dos o más frases intermedias.

c) Obtener todas las palabras de la Real Academia Española.

El diccionario de la Real Academia Española se encuentra disponible en un disco compacto. Sin embargo, la recolección de las palabras de este diccionario debió hacerse de manera semiautomática, dejando de lado las explicaciones y demás características asociadas a ellas como desinencias, posfijos, sufijos, etc. La característica subyacente que poseen estas palabras es la de contener todas las posibles combinaciones de sílabas fonológicas de acuerdo a su acentuación y ubicación dentro de la palabra.

### 3.3.2 Etapa 2: depuración de oraciones hasta obtener el corpus

Esta etapa incluyó las tareas de carga de la base de datos y de depuración de las oraciones de trabajo. Ambas tareas se realizaron usando la funcionalidad de la Base de Datos Corpus.

La Base de Datos Corpus incluye procesos que permiten una carga sencilla. Estos procesos pueden invocarse desde cualquier herramienta de software que acepte conexión a bases de datos y permita ejecutar sentencias usando el lenguaje SQL.

La fase de depuración consistió en eliminar, sobre la base de estrategias estadísticas, las oraciones recolectadas redundantes y a la vez agregar las necesarias para alcanzar el objetivo propuesto. La base de datos cuenta con procesos que retornan información estadística acerca de las sfc existentes y faltantes, en comparación con las de la Real Academia Española por un lado y con las del Estudio Estadístico del Español por el otro.

La estrategia de depuración utilizada sobre el corpus de oraciones de trabajo almacenado en la Base de Datos Corpus puede resumirse así:

*Búsqueda de sfc faltantes*

1) Obtener estadísticas sobre sílabas, palabras y sfc existentes.
2) Buscar y separar las sfc más usadas en las palabras de la Real Academia Española, que no están en el corpus de oraciones de trabajo. De cada sfc indicar las palabras de ese diccionario en las que se encuentra.
3) Buscar y separar las sfc obtenidas de Estudio Estadístico del Español, que no están en el corpus de oraciones de trabajo.

*Eliminación de oraciones redundantes*

4) Eliminar oraciones con todas sus palabras repetidas.
5) Eliminar oraciones con todas sus sfc repetidas.
6) Eliminar oraciones con todas menos una sfc repetidas y guardar esas sfc.

*Agregar nuevas oraciones*

7) Construir nuevas oraciones que incluyan las sfc separadas en los puntos 1), 2) y 3) intentando generar la menor cantidad de oraciones posible.
8) Utilizar la aplicación Sílabas con las oraciones del punto anterior.
9) Cargar en la base de datos el resultado de Sílabas.

*Repetir los pasos del 1) al 6), y luego agregar el siguiente*

10) Eliminar oraciones con todas menos dos sfc repetidas y guardar esas sfc.

*Repetir los pasos del 7) al 9)*

*Obtener estadísticas finales y Corpus de Oraciones*

11) Calcular porcentajes de representatividad alcanzada con el corpus de oraciones de trabajo actual.
12) Crear un archivo ASCII con todas las oraciones del corpus de trabajo. Ellas conforman el buscado Corpus de Oraciones.

### 3.3.3 Aplicación Sílabas

La aplicación Sílabas es un software que realiza una transcripción ortográfica y fonológica de un texto escrito almacenado en un archivo ASCII.

Al ser invocada, toma el archivo de texto ASCII que se le indica e identifica las oraciones que contiene. De cada oración obtiene las palabras que la componen y de cada palabra obtiene sus sílabas, trabajando en todo momento con los caracteres ortográficos que conforman el texto ingresado. Luego se realiza la transcripción fonológica que consiste en convertir cada sílaba ortográfica en su correspondiente sílaba fonológica. Finalmente se clasifican las sílabas fonológicas de acuerdo a la situación de acento (Si/No) y ubicación (Inicio, Medio, Fin o Neutral en caso de monosílabos) dentro de la palabra, que en este trabajo se denominan sílabas fonológicas contextualizadas (sfc).

Las oraciones y palabras identificadas más las sfc generadas, se almacenan en archivos ASCII. El formato de estos archivos es el mismo que utilizan como entrada los procesos de la base de datos prosódica cuya tarea es cargar oraciones, palabras y sílabas. Un ejemplo de los archivos de salida es el siguiente.

Ejemplo del archivo de oraciones:

1 | La casa es linda. |
2 | El animal come mucho. |

Ejemplo del archivo de palabras:

1|1|la|
1|2|casa|
1|3|es|
1|4|linda|
2|1|el|
2|2|animal|
2|3|come|
2|4|mucho|

Ejemplo del archivo de sílabas:

1|1|1|la|S|N|
1|2|1|ka|S|I|
1|2|2|sa|N|F|
1|3|1|es|S|N|
1|4|1|lin|S|I|
1|4|2|da|N|F|
2|1|1|el|S|N|
2|2|1|a|N|I|
2|2|2|ni|N|M|
2|2|3|mal|S|F|
2|3|1|ko|S|I|
2|3|2|me|N|F|
2|4|1|mu|S|I|
2|4|2|tso|N|F|

Este software fue desarrollado sobre la base de las rutinas preexistentes de separación en sílabas con que contaba el LIS. Ellas fueron adaptadas y ampliadas de acuerdo a los objetivos perseguidos.

### 3.3.4 Base de Datos Corpus

Es una base de datos que sirve para almacenar oraciones con sus palabras ortográficas y sílabas fonológicas, o sea, todo el material generado por los expertos lingüistas y por el método semiautomático descripto anteriormente en la Etapa 1 del corpus (recolección del material de trabajo).

Sirve como base de datos de trabajo y tiene la funcionalidad necesaria para depurar el corpus de oraciones de trabajo antes citado permitiendo por un lado, eliminar las oraciones recolectadas redundantes y por otro, agregar las oraciones necesarias hasta llegar a obtener el Corpus de Oraciones definitivo.

#### *3.3.4.1 Desarrollo de la Base de Datos Corpus*

El desarrollo de esta base de datos contempló el análisis realizado por el equipo del LIS. En la etapa de diseño también se colaboró con el equipo del LIS, el cual inicialmente creó un modelo de entidades y relaciones e implementó un prototipo relacional de la base de datos.

A partir de ahí se extendió el modelo de datos y se agregaron muchas funciones que debe llevar a cabo esta base de datos. Por lo tanto se definieron y se diseñaron los procesos necesarios para realizar todas las tareas involucradas en las dos etapas antes señaladas, conjuntamente con las restricciones de integridad que deben respetarse en todo momento. La implementación de los procesos se hizo usando vistas, procedimientos almacenados y disparadores, mientras que las restricciones de integridad se implementaron a través de los conceptos de clave primaria, clave única y clave externa. Para agilizar las búsquedas se crearon índices sobre los campos clave.

El diagrama de entidades y relaciones correspondiente a la base de datos corpus es el siguiente.

## *3.4. Diseño e implementación de la base de datos*

### 3.4.1 Consideraciones sobre las técnicas de diseño

Se hicieron dos diseños de la base de datos: uno basado en el paradigma de la programación orientada a objetos y otro basado en el tradicional paradigma de entidades y relaciones.

El diseño orientado a objetos permite implementaciones futuras usando lenguajes de programación que aceptan persistencia o que se acoplan a sistemas de gestión de bases de datos orientadas a objetos. Se utilizó el método OMT (Rumbaugh) para representar la estructura estática de las clases y las relaciones entre ellas. Previamente se trabajó con la técnica informal CRC (Class Responsibility Collaboration), que aportó las bases necesarias para lograr una buena jerarquía de clases y las responsabilidades de cada una.

Durante el diseño se trabajó en forma conjunta con el equipo del LIS. Habiéndose ya creado el Corpus de Oraciones bajo el paradigma relacional el grupo del LIS solicitó la implementación de la base de datos prosódica dentro del mismo paradigma, descartando las bases de datos orientadas a objetos y las relacionales orientadas a objetos. Se basó para ello no sólo en el conocimiento que poseen de la tecnología relacional y en la reusabilidad de las tablas, vistas, procedimientos almacenados y disparadores de la Base de Datos Corpus, sino también en la difusión e implementación masiva de los sistemas de gestión de bases de datos relacionales; propiciando así la integración con:

- Aplicaciones existentes y futuras para entrenamiento del conversor de texto a habla.
- Aplicaciones y proyectos del área lingüística del LIS.
- Programas internacionales de conversores de texto a habla multilingües.

Se arribó a esta situación contando con el modelo de entidades y relaciones creado para la Base de Datos Corpus y el diseño de los procesos para manipular su información. Entonces se contemplaron dos posibilidades para llegar a la implementación relacional:

1) Mapear el modelo orientado a objetos a un esquema de base de datos relacional.
2) Adaptar y extender el modelo de entidades y relaciones creado para el corpus de oraciones y mapearlo a un esquema de base de datos relacional.

Se estudiaron las dos alternativas y se concluyó que, en este caso, se llega a una mejor implementación con la segunda alternativa pues esta opción, además, permite reusar el diseño e implementación de los procesos de la Base de Datos

Corpus agregando los necesarios para esta base de datos. Para llegar a esta conclusión se contemplaron por un lado, las técnicas automáticas ampliamente difundidas de mapeo desde el modelo de entidades y relaciones al modelo relacional y, por otro lado, las posibles técnicas de mapeo del modelo orientado a objetos al modelo relacional. En relación con este último mapeo, se estudiaron las técnicas actuales que se proponen en el paper "Mapping Object to Data Models with the UML", publicado en el sitio en Internet de la firma Rational Rose, y se probaron dichas técnicas usando el software de diseño orientado a objetos que provee la misma empresa.

A continuación se presentan los dos diseños creados.

### 3.4.2 Diseño orientado a objetos

Inicialmente se trabajó con la técnica informal Class Responsibility Collaboration (CRC). Luego se utilizó el método OMT (Rumbaugh) para representar la estructura estática de las clases y las relaciones entre ellas, para cuya diagramación se utilizó UML (Unified Modeling Language).

#### *3.4.2.1 Técnica Class Responsibility Collaboration (CRC)*

Las clases resultantes de la aplicación de esta técnica son las siguientes:

*Clase: Base de datos prosódica (concreta)*

Esta clase representa a una base de datos prosódica que contiene información sobre la entonación, ritmo y acentuación del español de Buenos Aires y sirve para entrenar a un sistema de conversión de texto a habla. Tiene características fonéticas, fonológicas y parámetros asociados a la información acústica. Sabe responder a variadas consultas sobre todos los datos que posee.

| | |
|---|---|
| Conocer a los objetos contenidos en ella | |
| Cargar oraciones con sus palabras y sílabas | actualizador de oraciones |
| Cargar datos de emisiones | actualizador de emisiones |
| Cargar marcadores prosódicos | actualizador de prosodia |
| Cargar rasgos fonéticos | actualizador de fonética |
| Cargar parámetros de contornos | actualizador de contornos |
| Borrar todas las oraciones | actualizador de oraciones |
| Borrar todas las emisiones | actualizador de emisiones |
| Responder consultas prosódicas, fonéticas, fonológicas y sobre los parámetros de las unidades | |

del habla almacenadas consultor
Responder consultas generales y estadísticas

*Clase: Cargador de archivo (abstracta)*

Esta clase define el comportamiento común a todos los cargadores de archivos que transfieren datos ASCII hacia la base de datos.

Abrir archivo
Leer archivo
Cerrar archivo

*Clase: Cargador de archivo de oraciones (concreta)*

Esta clase representa la tarea de transferencia de oraciones.

Leer oraciones de un archivo
Crear colección temporaria de oraciones oración

*Clase: Cargador de archivo de palabras (concreta)*

Esta clase representa la tarea de transferencia de palabras.

Leer palabras de un archivo
Crear colección temporaria de palabras palabra

*Clase: Cargador de archivo de sílabas (concreta)*

Esta clase representa la tarea de transferencia de sílabas.

Leer sílabas de un archivo
Crear colección temporaria de sílabas sílaba

*Clase: Cargador de archivo de emisiones (concreta)*

Esta clase representa la tarea de transferencia de los datos de emisiones.

Leer los datos de emisiones de un archivo
Crear colección temporaria de datos de emisiones emisión

*Clase: Cargador de archivo ToBI (concreta)*

Esta clase representa la tarea de transferencia de transcripciones prosódicas ToBI.

| | |
|---|---|
| Leer los datos ToBI de un archivo | |
| Crear colección temporaria de elementos ToBI | elemento ToBI |

*Clase: Cargador de archivo de fonos (concreta)*

Esta clase representa la tarea de transferencia de transcripciones fonéticas CSLU.

| | |
|---|---|
| Leer los datos CSLU de un archivo | |
| Crear colección temporaria de elementos CSLU | elemento CSLU |

*Clase: Cargador de archivo de formantes (concreta)*

Esta clase representa la tarea de transferencia de contornos de formantes parametrizados.

| | |
|---|---|
| Leer los parámetros de los contornos de un archivo | |
| Crear colección temporaria de contornos parametrizados | contorno |

*Clase: Consultor (concreta)*

Esta clase representa el proceso de consulta de información contenida en la base de datos, el cual se lleva a cabo durante el entrenamiento del conversor de texto a habla.

| | |
|---|---|
| Retornar las características prosódicas de una emisión | emisión |
| Retornar las características fonéticas de una emisión | emisión |
| Retornar la oración correspondiente a una emisión | emisión |
| Retornar los focos entonacionales de una emisión | emisión |
| Retornar las emisiones asociadas a una oración | oración |
| Retornar el contorno de frecuencia fundamental de una emisión | emisión |
| Retornar el contorno de energía de una emisión | emisión |
| Retornar los contornos de los formantes de una emisión | emisión |

*Clase: Actualizador (abstracta)*

Esta clase define el comportamiento común a todos los actualizadores de información, que agregan, modifican y eliminan datos.

Comenzar actualización de la base de datos

Finalizar actualización

*Clase: Actualizador de oraciones (concreta)*

Esta clase representa al proceso que se lleva a cabo para cargar y eliminar oraciones con sus palabras y sílabas.

| | |
|---|---|
| Eliminar una oración | oración |
| Cargar oraciones desde un archivo | cargador de archivo de oraciones |
| Cargar palabras desde un archivo | cargador de archivo de palabras |
| Cargar sílabas desde un archivo | cargador de archivo de sílabas |

*Clase: Actualizador de emisiones (concreta)*

Esta clase representa al proceso que se lleva a cabo para cargar y eliminar los datos de la emisiones correspondientes a oraciones.

| | |
|---|---|
| Eliminar una emisión | emisión |
| Cargar datos de emisiones desde un archivo | cargador de archivo de emisiones |

*Clase: Actualizador de prosodia (concreta)*

Esta clase representa al proceso que se lleva a cabo para cargar los marcadores prosódicos de las emisiones.

| | |
|---|---|
| Cargar marcadores prosódicos desde un archivo | cargador de archivo ToBI |

*Clase: Actualizador de fonética (concreta)*

Esta clase representa al proceso que se lleva a cabo para cargar los rasgos fonéticos de las emisiones.

| | |
|---|---|
| Cargar rasgos fonéticos desde un archivo | cargador de archivo de fonos |

*Clase: Actualizador de contornos (concreta)*

Esta clase representa al proceso que se lleva a cabo para cargar los contornos parametrizados de las emisiones.

| | |
|---|---|
| Cargar contornos parametrizados desde un archivo | cargador de archivo de contornos |

*Clase: Sílaba (concreta)*

Esta clase representa a una sílaba fonológica que esta contenida en una o más palabras.

| | |
|---|---|
| Retornar su texto | |
| Conocer las palabras que la contienen | palabra |

*Clase: Palabra (concreta)*

Esta clase representa a una palabra ortográfica que esta contenida en una o más oraciones y esta formada por una o más sílabas con indicación del acento y ubicación dentro de ella.

| | |
|---|---|
| Conocer las oraciones que la contienen | oración |
| Retornar su texto | |
| Mantener el orden de sus sílabas | |
| Eliminar sus sílabas | sílaba |
| Conocer sus sílabas y sfc | sílaba |

*Clase: Oración (concreta)*

Esta clase representa a una oración típica del español de Buenos Aires, esta formada por una o más palabras con de la categoría gramatical dentro de ella, y constituye la transcripción ortográfica de una o más emisiones.

| | |
|---|---|
| Conocer sus palabras y la categoría gramatical correspondiente | palabra |
| Retornar su texto | |
| Conocer las emisiones asociadas | |
| Mantener el orden de sus palabras | |
| Eliminar sus palabras y sílabas | palabra |

*Clase: Emisión (concreta)*

Esta clase representa a la emisión acústica de una oración de texto. Contiene los datos característicos de la grabación, conoce su transcripción prosódica y fonética, y los parámetros de sus contornos asociados de frecuencia fundamental, energía y formantes.

| | |
|---|---|
| Conocer el texto de la oración pronunciada | oración |
| Retornar los datos propios de la grabación | |
| Conocer sus elementos ToBI | elemento ToBI |
| Conocer sus elementos CSLU | elemento CSLU |
| Conocer su contorno de energía | contorno de energía |

| | |
|---|---|
| Conocer su contorno de frecuencia fundamental | contorno de F0 |
| Conocer sus contornos de formantes | contorno de formante |
| Eliminar sus datos prosódicos, fonéticos y de contornos | elemento ToBI,<br>elemento CSLU,<br>contorno de energía,<br>contorno de F0<br>contorno de formantes |

*Clase: Contorno (concreta)*

Esta clase representa la evolución en el tiempo de un componente de la señal del habla. Esta formado por un conjunto de parámetros o datos temporales.

| | |
|---|---|
| Componerse | |
| Conocer sus parámetros | dato temporal |
| Conocer los parámetros que se encuentran dentro de un período de tiempo | dato temporal |
| Eliminar sus parámetros | dato temporal |

*Clase: Contorno de energía (concreta)*

Esta clase representa al contorno parametrizado de energía de una emisión.

Conocer la potencia sonora de la emisión

*Clase: Contorno de F0 (concreta)*

Esta clase representa el contorno de frecuencia fundamental de una emisión.

Conocer el nivel tonal de la frecuencia fundamental

*Clase: Contorno de formante (concreta)*

Esta clase representa a un formante de la emisión.

Conocer el nivel tonal del formante

*Clase: Dato temporal (concreta)*

Esta clase representa a un punto que indica el valor de un contorno en un instante de tiempo.

Retornar valor

Conocer su instante temporal

*Clase: Etiqueta temporal (concreta)*

Esta clase representa el período de tiempo en el cual se realiza una unidad del habla.

Retornar su instante de inicio
Retornar su instante de fin
Conocer la duración de la unidad del habla representada

*Clase: Elemento ToBI (concreta)*

Esta clase representa a los marcadores prosódicos de una palabra en una emisión, que incluye información tonal y métrica.

Conocer si posee un foco
Retornar símbolos tonales ToBI
Retornar índice de pausa de fin de palabra
Conocer la palabra etiquetada prosódicamente palabra

*Clase: Elemento CSLU (concreta)*

Esta clase representa a un fono de una emisión.

Retornar símbolo fonético

### *3.4.2.2 Diagrama de clases UML*

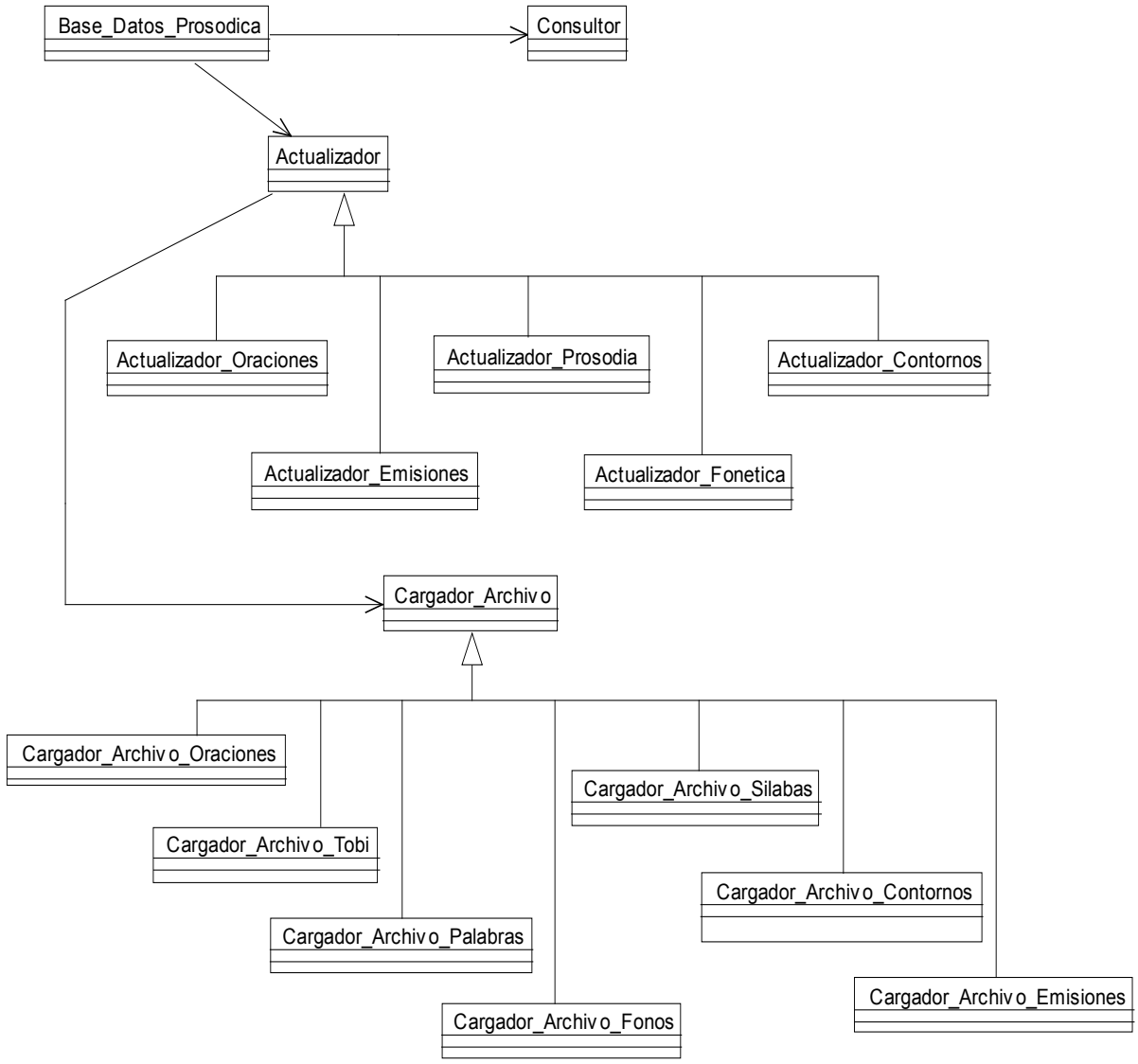


Figura 3.5. Diseño orientado a objetos: parte A.

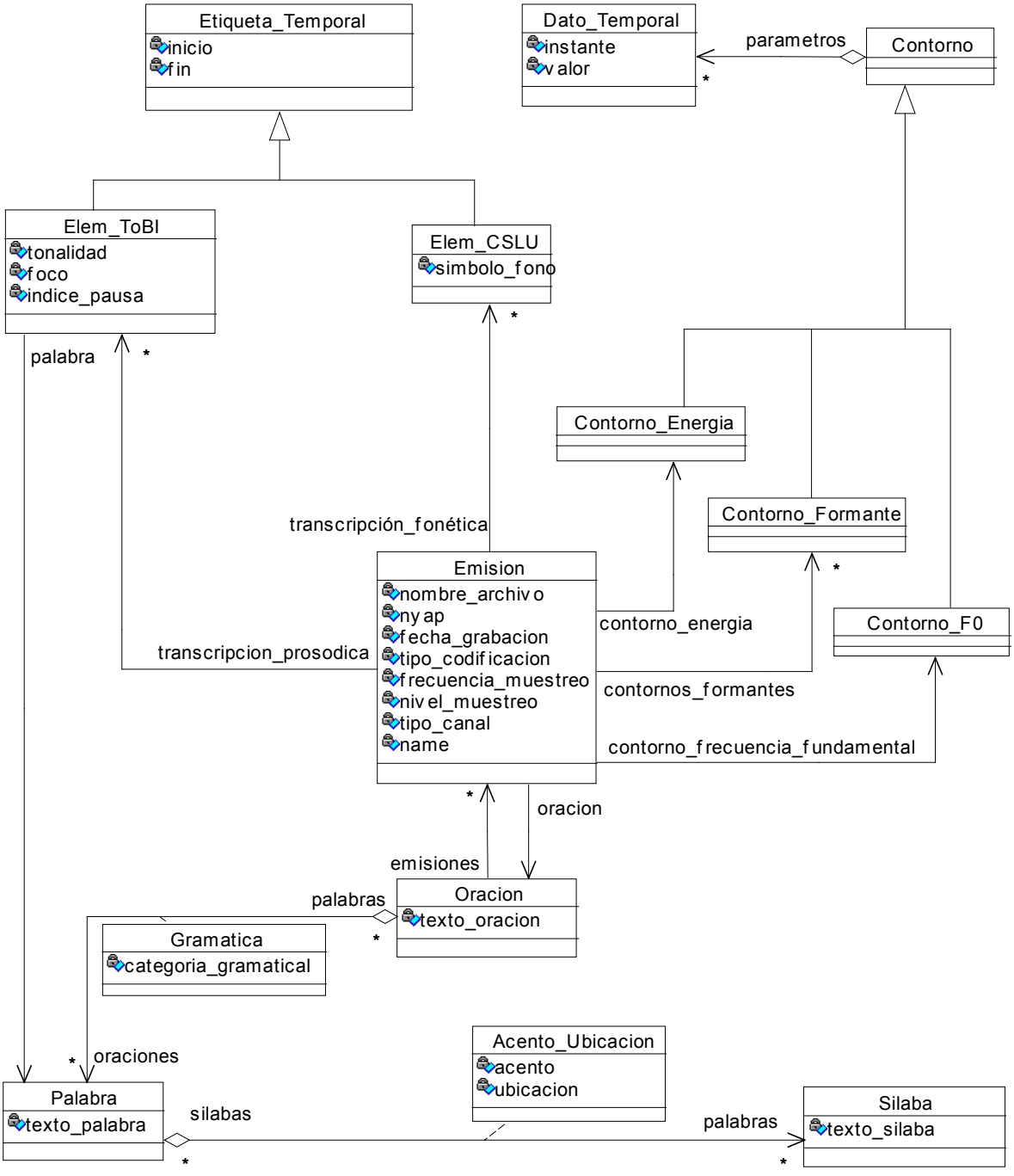


Figura 3.6. Diseño orientado a objetos: parte B.

### 3.4.3 Diseño Entidades y Relaciones

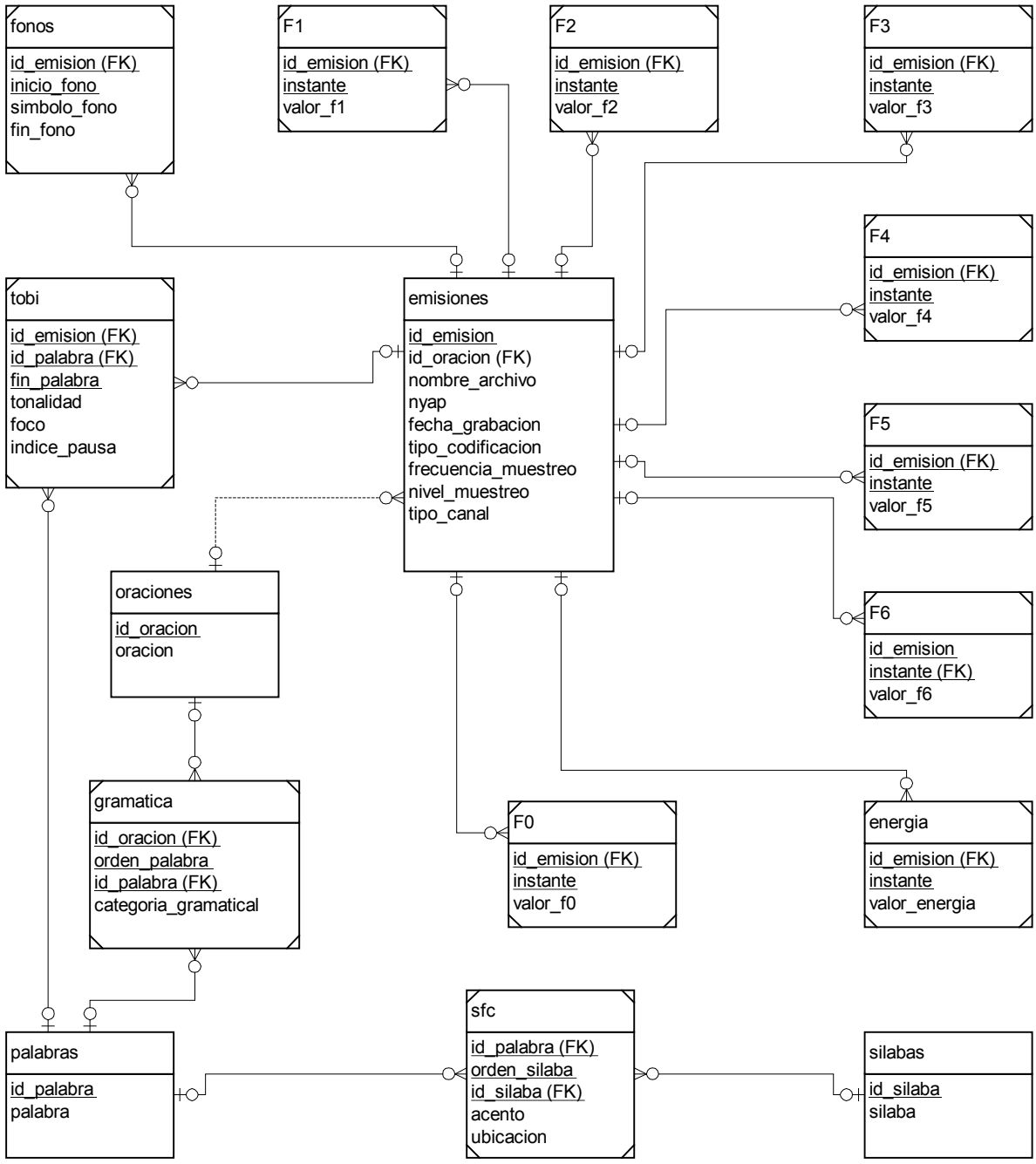


Figura 3.7. Diseño Entidades y Relaciones ya mapeado al modelo relacional.

Considerando las relaciones entre las tablas y, para mantener la integridad de la base de datos, se definieron restricciones de clave primaria, de clave única, de clave externa y restricciones sobre el dominio de los datos.

Además, para agilizar las operaciones de búsqueda de datos se crearon índices para las claves primarias y externas. También se crearon índices sobre otros campos que son utilizados en consultas específicas.

### 3.4.4 Procesos y funciones de la base de datos

Se definieron varios procesos para ser usados durante las tareas de carga de información y entrenamiento del conversor de texto a habla. Además se crearon procesos para realizar tareas de administración general de la base de datos y tareas de resumen y estado de la información contenida.

Los procesos definidos que dan funcionalidad a la base de datos y permiten obtener el conocimiento que de ella se espera, son:

#### *3.4.4.1 Procesos relacionados con la carga de oraciones, palabras y sílabas*

- Cargar oraciones: carga oraciones, con sus palabras y sílabas fonológicas asociadas que se encuentran en tres archivos ASCII: 'oraciones.txt', 'palabras.txt' y 'sílabas.txt' (archivos que pueden ser generados, por ejemplo, con la aplicación Sílabas creada durante este trabajo).
- Borrar una oración: no sólo debe eliminar la oración que se indica de la base de datos sino también sus palabras y sílabas asociadas, y eliminar también aquellas palabras y sílabas que no tienen oración que las contenga.
- Borrar una oración con sus emisiones: a partir de la oración que se le indica debe invocar al proceso 'Borrar una oración', y luego, para cada emisión correspondiente a esa oración, invocar al proceso 'Borrar una emisión'.

#### *3.4.4.2 Procesos relacionados con la carga de datos de las emisiones*

- Cargar datos de emisiones: carga los datos de las emisiones referidas en el archivo ASCII 'emisiones.txt'.
- Borrar una emisión: se le indica el nombre de un archivo de onda, el cual es eliminado junto con sus datos prosódicos, fonéticos, de frecuencia fundamental,

energía y formantes.
- Borrar las emisiones asociadas a una oración: se indica una oración y, para cada emisión correspondiente a ella, invoca al proceso 'Borrar una emisión'.
- Mostrar los datos de una emisión: se le indica el nombre de un archivo de onda. Este proceso retorna los datos asociados al archivo indicado que están en la base de datos.

#### *3.4.4.3 Procesos relacionados con la carga de información prosódica, fonética y de contornos parametrizados*

- Cargar transcripciones de prosodia: carga los marcadores prosódicos que están en el archivo ASCII 'tobi.txt'.
- Cargar transcripciones de fonética: carga los datos fonéticos que se encuentran en el archivo ASCII 'cslu.txt'.
- Cargar contornos parametrizados que se encuentran en el archivo ASCII 'contornos.txt'.

#### *3.4.4.4 Procesos generales y estadísticos*

- Mostrar la cantidad de oraciones, palabras, sílabas y sfc contenidas en la base de datos: cuenta el total de esos elementos y retorna el resultado.
- Mostrar todas las sfc y la cantidad de repeticiones de cada una de ellas: busca todas las sfc contenidas en la base de datos y para cada una cuenta la cantidad de veces que se repite.
- Mostrar todas las sílabas y la cantidad de repeticiones de cada una de ellas: busca todas las sílabas y para cada una de ellas cuenta la cantidad de veces que se repite.
- Mostrar todas las palabras y la cantidad de repeticiones de cada una de ellas: busca todas las palabras y para cada una de ellas cuenta la cantidad de veces que se repite.
- Mostrar la cantidad de repeticiones de una pado: se indica una pado, la busca en todas las oraciones y cuenta la cantidad de veces que se repite.
- Mostrar las oraciones en las que aparece una pado: se indica una pado, la busca en todas las oraciones y retorna aquellas en las que aparece.
- Mostrar las oraciones en las que aparece una sfc: se indica una sfc, la busca en todas las oraciones y retorna aquellas en las que aparece.
- Borrar todas las oraciones: borra todas las oraciones, sus palabras, sílabas y emisiones con prosodia y fonética.

- Borrar todas las emisiones: elimina todas las emisiones con sus datos prosódicos, fonéticos y de contornos parametrizados.

#### *3.4.4.5 Procesos relacionados con el entrenamiento de un conversor de texto a habla*

- Mostrar los focos entonacionales de una emisión: retorna las palabras donde se encuentran los tonos más altos de cada frase entonacional contenida en la oración pronunciada
- Mostrar la oración correspondiente a una emisión: se indica una emisión y retorna la oración correspondiente a ella.
- Indicar cuáles son las emisiones asociadas a una oración: se indica una oración y retorna las emisiones en las que se la pronuncia.
- Mostrar los marcadores prosódicos de una emisión: se indica una emisión y retorna los marcadores prosódicos correspondientes a ella.
- Mostrar las características fonéticas de una emisión: se indica una emisión y retorna los rasgos fonéticos correspondientes a ella.
- Mostrar el contorno de frecuencia fundamental de una emisión: se indica una emisión y retorna los parámetros del contorno de F0 correspondiente a ella.
- Mostrar el contorno de energía de una emisión: se indica una emisión y retorna los parámetros del contorno de Energía correspondiente a ella.
- Mostrar los formantes de una emisión: se indica una emisión y retorna los parámetros de los contornos de los formantes correspondientes a ella.

### 3.4.5 Modos de acceso a la base de datos

Se distinguen dos modos de acceso a la base de datos:

- invocando los procesos arriba citados
- a través del lenguaje SQL

Ambos modos pueden llevarse a cabo mediante las aplicaciones genéricas de consulta de base de datos (ejemplos: ISQL, Microsoft Query, etc.), los cuales brindan un ambiente interactivo y fácil de usar. Todo DBMS proporciona una o más herramientas de este tipo.

También pueden utilizarse aplicaciones desarrolladas por programadores que con diversos objetivos necesitan acceder a la base de datos. Pueden ejecutarse localmente o en forma remota (ejemplo: vía el estándar ODBC).

Ejemplos:

1) A través de sentencias SQL

```
database bdp
Select count(*)
  from oraciones
```

2) Invocando a los procesos desarrollados en este trabajo

```
database bdp
execute mostrar_prosodia("La casa es linda.")
```

Ambos ejemplos también pueden aparecer como parte del código de una aplicación con acceso a bases de datos:

1)

```
_____ (código de la aplicación) ____
_____ (código de la aplicación) ____

conectarse a una base de datos (bdp)
Select count(*)
  from oraciones

_____ (código de la aplicación) ____
_____ (código de la aplicación) ____
```

2)

```
_____ (código de la aplicación) ____
_____ (código de la aplicación) ____

conectarse a una base de datos (bdp)
execute mostrar_prosodia("La casa es linda.")

_____ (código de la aplicación) ____
_____ (código de la aplicación) ____
```

### 3.4.6 Soporte de la base de datos

La base de datos está soportada por un sistema de gestión de bases de datos (DBMS), el cual incluye una herramienta de interfaz con el usuario y un lenguaje de consulta muy potente, para realizar gran variedad de tipos de consultas.

Considerando estos dos productos asociados al DBMS, los tipos de consultas prosódicas propuestos en este trabajo pueden ser ampliados y modificados en el futuro sin gran esfuerzo de programación. Por ejemplo, supóngase que se desea incorporar información sobre las categorías gramaticales de las palabras y luego hacer consultas a partir de dichos datos. Con sólo indicar al lenguaje del sistema de gestión de bases de datos cómo resolver el nuevo tipo de consultas, se puede obtener aún más conocimiento de la base de datos.

Todas las sentencias del lenguaje de consulta podrán ser usadas para definir nuevas estructuras de datos y manipular la información contenida en ellas (agregar, eliminar, modificar y consultar datos).

Utilizando el software de interfaz con el usuario se puede invocar a los procesos propuestos y creados en este trabajo, que facilitan el uso de la base de datos, aunque también se pueden ingresar sentencias escritas en el lenguaje de consulta proporcionado junto al DBMS.

### 3.4.7 Implementación

La implementación en el modelo relacional se realizó en primera instancia utilizando la herramienta CASE con la que se creó el modelo de entidades y relaciones, que permitió contar con el esquema de tablas básico y con las restricciones de integridad e índices asociados. Luego se crearon las vistas necesarias y se programaron los disparadores y procedimientos almacenados.

Para implementar los procesos que implican tareas que no pueden realizarse de manera incompleta se utilizaron transacciones. Un ejemplo es el proceso que carga oraciones con sus palabras y sílabas, el cual no produce efectos si no se completa toda la operación.

Para implementar las tareas de borrado en cascada, como por ejemplo el proceso de borrado de oraciones con sus palabras y sílabas, se usan las características avanzadas de restricciones de integridad para clave externa y también disparadores. Los disparadores correspondientes a este ejemplo se introdujeron en las tablas "oraciones", "gramatica", "palabras" y "sfc".

## *3.5. Prueba de la base de datos prosódica*

Se desea comprobar si la base de datos construida en este trabajo sirve para entrenar a un sistema de conversión de texto a habla.

La realización de la prueba de la base de datos prosódica debe plantearse a partir de las dos principales funciones que ella cumple: cargar la información que se desea almacenar y retornar los datos que se requieran durante el entrenamiento del conversor de texto a habla.

Sin embargo, aquí también se realiza un experimento piloto para generar habla artificial de alta calidad, utilizando para ello dos sintetizadores diferentes. A estos sintetizadores se les introducen los rasgos prosódicos, en este caso representados por el contorno de frecuencia fundamental que se obtienen del entrenamiento antes mencionado.

### 3.5.1 Carga de información

La funcionalidad requerida durante la carga de información se probó utilizando los procesos de incorporación automática de datos implementados para:

- Cargar todas las oraciones que son parte del Corpus de Oraciones, anteriormente creado, con sus palabras ortográficas y sílabas fonológicas.
- Cargar los datos de las emisiones correspondientes a algunas de esas oraciones.
- Cargar las características prosódicas y fonéticas.
- Cargar los contornos de frecuencia fundamental (F0), energía (E) y formantes (F1,…,F6).

El correcto funcionamiento de los procesos de carga se verificó posteriormente a través de consultas manuales usando SQL y de otros procesos que retornan la información antes incorporada

### 3.5.2 Entrenamiento del conversor de texto a habla

Las tareas principales que se realizan durante el entrenamiento son:

- Ayudar a definir, elegir y depurar las estrategias de generación prosódica y fonética que se usarán durante la conversión on-line.
- Obtener las estructuras oracionales con sus patrones prosódicos, y los datos de unidades del habla que se desean almacenar en la base de datos complementaria que será parte del sistema conversor.

Para probar que la base de datos sirve para entrenar a un conversor de texto a habla fueron invocados los procesos que retornan los siguientes datos:

- Palabras, sílabas fonológicas y sfc asociadas a las oraciones.
- Símbolos tonales y de índice de pausas asociados a las emisiones.
- Símbolos fonéticos asociados a las emisiones con indicación de su realización temporal.
- Parámetros de F0 asociados a las emisiones.
- Parámetros de energía asociados a las emisiones.
- Parámetros de F1,..., F6 asociados a las emisiones.

A través de variadas combinaciones en la utilización de estos procesos se obtuvieron diferentes patrones prosódicos y rasgos fonéticos, como también una gran variedad de símbolos tonales y métricos y contornos parametrizados, concluyendo entonces que se puede llevar a cabo con gran sencillez la parte del entrenamiento en la que se requiere interacción con la base de datos.

### 3.5.3 Generación de habla sintética

Consideremos un conversor de texto a habla que realiza los siguientes pasos básicos una vez indicado el texto que debe pronunciar:

1) Obtener unidades ortográficas.
2) Generar los contornos parametrizados (E y formantes F1,…, F6) de las unidades del habla derivadas de 1).
3) Obtener la estructura sintáctica.
4) Generar los símbolos tonales y métricos derivados de 3).
5) Generar los contornos parametrizados de frecuencia fundamental (F0) y energía (E) derivados de 4).
6) Realizar proceso de síntesis usando los contornos obtenidos en 2) y 5).

Cabe aclarar que la energía final debe ser calculada, durante la síntesis, sobre la base de la energía propia de las unidades del habla y de la que está asociada al contorno de frecuencia fundamental.

Considerando que el proyecto de desarrollo del conversor del LIS se encuentra en la etapa inicial sólo se cuenta con aplicaciones que pueden llevar a cabo el último paso. Estas aplicaciones son dos sintetizadores diferentes que fueron utilizados para realizar un experimento piloto y así verificar la alta calidad de habla que pueden producir si se los alimenta con la información prosódica adecuada que se deriva de

los símbolos tonales y métricos. Para obtener estos símbolos se invocó a los procesos correspondientes que se encuentran en la base de datos.

Uno de los sintetizadores es el software Sinpar del LIS, que realiza un proceso de síntesis por Klatt (método que usa formantes y parámetros generales). El otro, software Praat, es un sintetizador que permite el uso del método PSOLA para la sustitución del contorno de frecuencia fundamental de una emisión. Para utilizar el primero de ellos se introdujeron en el sintetizador los contornos generales obtenidos a partir de la aplicación Anagraf del LIS, luego se eligieron parámetros adecuados para la síntesis por formantes y finalmente se agregó el contorno de frecuencia fundamental que se construyó manualmente a partir de los símbolos tonales y métricos que se obtuvieron de la base de datos. En el caso de la síntesis por el método PSOLA, se sustituyó el contorno de frecuencia fundamental original por el mismo que se introdujo en el anterior sintetizador, con los ajustes correspondientes.

En ambos casos, luego del proceso de síntesis, se observó la alta calidad de habla artificial que se puede lograr usando datos obtenidos del entrenamiento con una base de datos prosódica.

# Capítulo 4 CONCLUSIONES

## *4.1. Trabajos realizados*

Se construyó una base de datos prosódica muy sencilla y a la vez poderosa, con un diseño simple y fácil de entender. La base de datos soporta gran cantidad de información prosódica, y a la vez es flexible y fácilmente extensible. Cuenta con la funcionalidad requerida para entrenar a un conversor de texto a habla de alta calidad para el español de Buenos Aires. Lingüistas dedicados a la entonación de nuestra lengua pueden trabajar con ella de manera confortable, a la vez que pueden ampliar sencillamente su contenido usando las distintas funciones previstas para ello.

Los objetivos secundarios propuestos y los resultados alcanzados son:

- Explorar y exponer los antecedentes y el estado del arte de los sistemas de conversión de texto a habla y de las bases de datos que estos conversores usan en búsqueda de la alta calidad.

Se realizó una investigación interdisciplinaria para el entorno actual de los sistemas de conversión de texto a habla y las bases de datos prosódicas. En particular se describieron los modelos y estrategias más adecuados para el español de Buenos Aires. Estos luego influyeron en la especificación detallada del contenido de la base de datos y las funcionalidades de la misma.

- Aportar los conocimientos sobre las variadas tecnologías y modelos propios de la informática como integrante del equipo del LIS que lleva adelante el desarrollo de un conversor de texto a habla para el español hablado en Buenos Aires.

Se participó en las reuniones de trabajo del equipo y se propusieron los métodos y técnicas más adecuadas en la concepción de los problemas asociados a la informática y para la implementación de los mismos.

- Desarrollar el software complementario requerido en las diversas etapas de construcción de la base de datos prosódica.

Se desarrolló la aplicación Sílabas y la base de datos Corpus:

**Aplicación Sílabas:** software para transcripción automática de oraciones y palabras ortográficas, y de sílabas fonológicas clasificadas de acuerdo con la situación de acento y ubicación dentro de la palabra. Se pretende su utilización como parte de un método semiautomático de obtención de un Corpus de Trabajo con sus rasgos ortográficos y fonológicos característicos.

**Base de Datos Corpus:** base de datos para crear un corpus de oraciones. Debe permitir almacenar sílabas fonológicas y su frecuencia de uso en el español de Buenos Aires, un corpus de oraciones de trabajo y también todas las palabras de la Real Academia Española. Deben ser sus funciones: comparar conjuntos de sílabas, generar la información estadística correspondiente y depurar oraciones sobre la base de procesos estadísticos.

Considerando la velocidad de los cambios que se producen en el ámbito de las tecnologías del habla se utilizaron técnicas que permiten la reusabilidad de los trabajos realizados y el fácil y rápido entendimiento para aquel que se integre al equipo de trabajo.

La presente base de datos prosódica contiene la información que permite mostrar los patrones estructurales de la entonación, acento y ritmo del español de Buenos Aires y puede ser extendida fácilmente con datos derivados de emisiones del habla espontánea.

## *4.2. Conocimientos y habilidades adquiridos*

A continuación se presenta un resumen de los conocimientos y habilidades adquiridos durante el desarrollo del presente trabajo:

*Conocimientos adquiridos*

- Funcionamiento interno de un conversor de texto a habla.
- Modelos de producción del habla.
- Anatomía y fisiología del aparato fonador.
- Voz normal y anormal.
- Lingüística: discurso, fonética, prosodia, entonación.
- Metodología de la investigación científica.

*Habilidades adquiridas*

- Investigar en la aldea global (libros, revistas, Internet, relación con centros de investigación de otras latitudes).
- Digitalizar el habla humana y analizar la señal acústica.
- Sintetizar el habla.
- Técnicas de reconocimiento automático del habla.
- Técnicas de verificación del hablante.

Durante este trabajo se tomaron cursos de:
- Metodología de la investigación científica (Facultad de Bellas Artes, UNLP).
- Técnicas de análisis y síntesis (LIS, CONICET).
- Laboratorio de la Voz (Hospital Alemán).

Además se inició una investigación sobre "disfonías emocionales" conjuntamente con el Dr. Jorge N. González y otros profesionales del Hospital Alejandro Korn, sito en la localidad de Melchor Romero. También se definió la estrategia de identificación de personas vía habla (identificación del hablante) para su uso en la Suprema Corte de Justicia de Buenos Aires, oficina pericial La Plata.

## *4.3. Proyecciones*

Las proyecciones de este trabajo pueden dividirse entre específicas y generales. Entre las primeras se encuentran las que se refieren a las actividades relacionadas con los sistemas de conversión de texto a habla y por extensión a las tecnologías del habla. Entre las proyecciones generales se encuentran aquellas que exceden estas áreas y sirven a la sociedad en general. Además se enumera una serie de extensiones que pueden realizarse fácilmente.

Proyecciones específicas:

- Aplicaciones de síntesis del habla.
- Aplicaciones de reconocimiento del habla.
- Aplicaciones de reconocimiento/verificación del hablante.

Proyecciones generales:

- Investigaciones sobre fonética.
- Investigaciones sobre lingüística.
- Desarrollo de métodos de adquisición de la lengua materna.
- Desarrollo de métodos de adquisición de una segunda lengua.

- Investigaciones sociolingüísticas.
- Investigaciones psicolingüísticas.
- Aplicaciones de fonoaudiología.
- Terapias de voz y habla.

Posibles extensiones de la base de datos:

- Incorporar un sistema de consultas sobre la base de las distintas combinaciones de categorías gramaticales y palabras ortográficas, con indicación posicional dentro de la oración.
- Agregar estructuras para almacenar diferentes tipos de caracterización prosódica y fonética, de acuerdo con métodos que siguen diversas teorías similares o distintas.
- Incorporar características de agrupamiento de la información por tema y tipo de habla.
- Integración con programas internacionales de creación de corpus de habla multilingües.

# APÉNDICE PROYECTO DEL LIS

## Etapas de desarrollo de un sistema de conversión de Texto a Habla

Se desarrollarán distintas estrategias adecuadas para el español para generar los contornos prosódicos a partir del texto escrito. Se comienza con una etapa de aprendizaje en una base de datos que contiene los contornos prosódicos observados en un gran número de oraciones.

### *Base de Datos*

La base de datos contendrá un total de 1500 sentencias formadas por: 500 sentencias gramaticales declarativas con palabras simples, 500 sentencias gramaticales interrogativas, 160 sentencias gramaticales complejas (con sintaxis más compleja y palabras polisilábicas), 50 sentencias de obras de teatro, 26 sentencias, equivalentes a números desde 0 a 100 y fracciones, 1 sentencia, equivalente al alfabeto, 50 sentencias sobre información de negocios y la bolsa, 23 sentencias de noticias (divididas en tres notas cortas), 40 sentencias de meteorología y de información sobre tráfico, 150 sentencias, equivalentes a 1000 palabras aisladas.

### Sonidos

Las formas de onda serán grabadas por un locutor masculino y otro femenino, digitalizadas y almacenadas en un disco rígido y segmentadas, para su etiquetado, a 16 bit usando una frecuencia de muestreo superior a los 16 kHz

### Etiquetado de la Base de Datos

Expertos con entrenamiento musical marcarán en forma manual las características melódicas de las oraciones. Se desarrollará un curso para entrenar a estos etiquetadores, para que conozcan y apliquen idóneamente las reglas de transcripción, para que se uniformen los criterios de etiquetado y practiquen con las herramientas disponibles para el estudio del habla. A partir de este curso serán

seleccionados dos etiquetadores. El tiempo de trabajo de los etiquetadores se calcula usando como base que un etiquetador puede rotular cuarenta (40) oraciones por semana trabajando 4 horas diarias. Entonces una base de datos (de 1500 oraciones) podrá ser completada en diez meses de trabajo de un etiquetador. Las bases se rotularán dos veces para cada uno de los hablantes. Los etiquetadores recibirán apoyo de un supervisor para la solución de las dificultades.

### Verificación del etiquetado

Se realizarán dos verificaciones: una automática y otra manual. La verificación automática servirá para controlar las transcripciones, observar errores fonéticos e inconsistencias a nivel tonal. Para ello se utilizarán programas especiales, los cuales detectarán los errores, que luego serán corregidos manualmente. Se pondrá especial énfasis en las variaciones alofónicas más frecuentes del castellano. La verificación manual consistirá en evaluar oración por oración después de usar la estrategia de aceptación o rechazo para grandes errores. Los errores menores serán corregidos in situ. Un supervisor fonético y métrico orientará a los etiquetadores y resolverá problemas de transcripción que se presenten sobre la marcha. Un primer control se realizará semanalmente para brindar información rápida a los etiquetadores sobre su desempeño y mantener un promedio de menos del diez por ciento de oraciones observadas.

### Análisis acústico y programa de etiquetado

Se adaptará el programa para el Procesamiento Digital de Señales del LIS (Anagraf). Este muestra simultáneamente la forma de onda, el espectrograma, la energía, y los contornos de F0, para producir etiquetas acordes con la guía de etiquetado de CSLU. Este programa producirá archivos de texto con las correspondencias entre los niveles fonético y fonémico y el instante temporal del archivo de onda. El mismo programa se adaptará para realizar la transcripción de acuerdo al ToBI, y para el Worldbet, que representa los símbolos fonéticos del IPA (International Phonetic Alphabet) en caracteres ASCII. El programa producirá archivos de acuerdo a las correspondencias acústicas a tonos y junturas de palabras.

### El método de etiquetado para los rasgos fonológicos

La guía de etiquetado del CSLU se utilizará como convención para la transcripción de los datos fonémicos y fonéticos, tanto como los segmentos no pertenecientes al habla. El Worldbet establece un conjunto de caracteres ASCII para la transcripción de las realizaciones de los fonemas del español. El alineamiento temporal se realizará en primer lugar con relación a la forma de onda, y luego teniendo en consideración el espectrograma (realizado con una ventana pequeña), los contornos de Energía y el F0 vistos simultáneamente y utilizando reglas fijas en casos ambiguos.

### Método de etiquetado para los rasgos métricos y tonales.

Para el etiquetado métrico se adoptará el sistema ToBI. Esta perspectiva ha mostrado ser adecuada para oraciones no espontáneas del español.

## *Actividades relacionadas a la generación de contornos prosódicos a partir del texto.*

### Diccionario

Con el corpus de texto de la base de datos se construirá un diccionario de los ítems lexicales y las categorías sintácticas morfológicas, acento lexical e información fonética. Este diccionario compuesto con material de entrenamiento podrá ser ampliado en futuros trabajos.

### Análisis del texto

Se propone construir un analizador que obtendrá una representación sintáctica de la oración y una representación prosódica de categorías fonológicas abstractas como la estructura métrica y pie, las sílabas y sus estructuras silábicas, a partir del corpus de texto y la información que provee el diccionario. Para tal fin se construirá una gramática de reconocimiento y generación en un lenguaje declarativo que será implementado en Prolog. Las representaciones sintácticas y prosódicas estarán sincronizadas lo que permitirá predecir la ubicación de los acentos tonales,

los acentos de límite de los grupos melódicos (frases intermedias), y los acentos de límite de frase entonacional, además de la ubicación del foco oracional –si lo hubiera- y la modalidad oracional (interrogativa, exclamativa, etc.).

### Generación de los marcadores tonales y métricos

Se construirán las reglas para la generación de los contornos prosódicos (acústicos) de las oraciones. Se construirá una gramática de generación de la estructura tonal de las oraciones de acuerdo a la estructura sintáctica, morfológica, métrica, silábica y modalidad oracional aplicable a las oraciones del corpus.

### Generación de los contornos prosódicos

Definida la estructura tonal de la sentencia a sintetizar se evaluará un método paramétrico híbrido original basado en el de prominencias/interpolación de Pierrehumbert que el grupo del LIS ha presentado en el Congreso de la ESCA´99 (EUROSPEECH '99).

Otras investigaciones explorarán las asociaciones entre el conjunto de símbolos tonales abstractos y los contornos de entonación observados mediante técnicas probabilísticas. Para ello se definirán modelos basados en los acentos tonales y acentos de frase. Se asociarán los contornos prosódicos acústicos con la secuencia de símbolos tonales y métricos presentes en la Base de Datos.

En todos los casos la aproximación a los contornos observados será evaluada mediante un criterio de error convencional como lo son las diferencias cuadráticas medias entre el contorno observado y el estimado.

# BIBLIOGRAFÍA

BECKMAN, M. y AYERS, G. 1994. **Guidelines for ToBI Labelling**. The Ohio State University Research Foundation.

BIRD, S. y LIBERMAN, M. 1999. **A Formal Framework for Linguistic Annotation.** Linguistic Data Consortium. University of Pennsylvania.

CLARK, J. y YALLOP, C. 1995. **An Introduction to Phonetics and Phonology**. Blackwell.

CSLU. 1993. **CSLU Labelling Guide**. Center for Spoken Language and Understanding, Oregon Graduate Institute.

DUSTERHOFF, K. y BLACK, A**.** 1997. **Generating F0 contours for speech synthesis using the Tilt entonation theory**. ESCA '97 Workshop. Atenas.

DUTOIT, T. 1996. **A Short Introduction to Text-to-Speech Synthesis**. TCTS Laboratories, Mons.

EAGLES. 1997. **Handbook of Standards and Resources for Spoken Language Systems**. Ed. por Gibbon, D. y otros, Mouton de Gruyter.

FLANAGAN, J. 1972. **Speech Analysis, Synthesis and Perception.** Second Expanded Edition. Berlín, Springer Verlag.

FLANAGAN, J. 1972. **Voices of Men and Machines.** En: The Journal of the Acoustical Society of America. Número 51. Marzo 1972.

FUJISAKI, H. 1993. **From Information to Intonation**. LIS - CONICET, Buenos Aires.

GUIRAO, M. y GARCÍA JURADO, M. A. 1993. **Estudio estadístico del español**. Buenos Aires, LIS - CONICET.

GUIRAO, M. 1980. **Los sentidos, bases de la percepción.** Alhambra Universidad, Madrid.

GURLEKIAN, J. y otros. 1983**. El hombre dialoga con la máquina.** Quid - El de la ciencia, la tecnología y la educación argentina. Tomo 11. Número 14.

GURLEKIAN, J. 1997**. Laboratorio de audición y habla del LIS.** En: Procesos Sensoriales y Cognitivos. Buenos Aires, Ediciones Dunken.

GURLEKIAN, J. y otros. 1999**. A semi automatic method for the characterization of spanish intonation contours.** En: Proc. EUROSPEECH. Budapest.

KLATT, D. 1980. **Software for a cascade/parallel format synthetizer**. En: The Journal of the Acoustical Society of America. Vol. 67.

HART ('t), J. y otros. 1988. **A Synthesis Scheme for British English Intonation**. En: The Journal of the Acoustical Society of America. Vol. 84. Número 4.

HIERONYMUS, J. L. 1994. **ASCII Phonetic Sysmbols for the World's Languages: Worldbet**. AT&T Bell Laboratories, Journal of the International Phonetic Association.

HUNT, A. y BLACK, A. 1996. **Unit selection in a Concatenative Speech Synthesis System using a Large Speech Database**. ATR Laboratories, Kioto.

LLISTERRI, J. 1988. **La síntesis del habla: estado de la cuestión.** En: Procesamiento del Lenguaje Natural. Boletín número 6. Barcelona.

LLISTERRI, J. 1993. **Aplicaciones de la fonética a la tecnología del habla en español: conversión de texto a habla y bases de datos acústicos**. En: III Simposium Nacional de Hispanistas Polacos.

LOPEZ GONZALO, E. y otros. 1997. **Automatic Corpus-Based Trining of Rules for Prosodic Generation in Text-to-Speech**. En: Proc. EUROSPEECH. Atenas.

LOPEZ GONZALO, E. y otros. 1999**. A Mixed Strategy Approach to Spanish Prosody for Text-to-Speech**. En: Proc. EUROSPEECH. Budapest.

MANRIQUE, A. M. B. 1980. **Manual de fonética acústica**. Buenos Aires, Hachete.

NAVARRO TOMÁS, T. 1974. **Manual de entonación española.** Madrid, Guadarrama.

PIERREHUMBERT, J. 1988. **The phonology and phonetics of English intonation**. Ph. D., MIT, IULC.

RABINER y SCHAFER. 1979. **Procesamiento digital d señales y habla**. New Jersey, Prentice Hall.

RENATO, A. 1997**. Boundary tones pitch and ending movements in Argentinean Spanish**. Workshop in intonation. ESCA. Atenas.

RICCILLO, M. 1998. **Desarrollo de un método semiautomático para la caracterización de contornos de entonación.** Tesis de licenciatura. Universidad de Buenos Aires, Fac. Cs. Ex. y Naturales.

ROSS, K. y OSTENDORF, M. 1999. **A Dynamical System Model for Generating Fundamental Frecuency for Speech Synthesis**. En: IEEE Transactions on Speech and Audio Processing. Vol. 7. Número 3. Mayo 1999.

SAKURAI, A. y otros. 1998. **A Linguistic and Prosodic Database for Data-Driven Jananese TTS Synthesis**. Texas Instruments, Tokio y University of Tokio.

SAMAJA, J. 1995. **Epistemología y Metodología**. Buenos Aires, EUDEBA.

SILBERSCHATZ, A. y otros. 1997. **Fundamentos de bases de datos**. Madrid, Mc Graw Hill.

SOSA, J.M. 1991. **Fonética y fonología el español hipanoamericano**. Ph. D. University of Massachusetts.

WIRFS-BROCK, R. y otros. 1990. **Designing Object-Oriented Software.** New Jersey, Prentice Hall.