LLISTERRI, J.- CARBÓ, C.- MACHUCA, M. J.- de la MOTA, C.- RIERA, M.- RÍOS, A. (2003) "El papel de la lingüística en el desarrollo de las tecnologías del habla", in CASAS GÓMEZ, M. (Dir.) - VARO VARO, C. (Ed.) *VII Jornadas de Lingüística*. Cádiz: Servicio de Publicaciones de la Universidad de Cádiz. pp. 137-191.
http://liceu.uab.es/publicacions/Linguistica_TH_Cadiz02.pdf

El papel de la lingüística en el desarrollo de las tecnologías del habla

Joaquim Llisterri, Carme Carbó, María Jesús Machuca, Carme de la Mota, Montserrat Riera y Antonio Ríos

Departament de Filologia Espanyola
Universitat Autònoma de Barcelona

1. Las tecnologías del habla

El auge de la Sociedad de la Información ha propiciado que aparezcan nuevas necesidades relacionadas con el acceso de un número cada vez mayor de personas a sistemas informáticos que tradicionalmente habían estado sólo al alcance de una minoría. Operaciones tan cotidianas como comprar un billete de tren o de avión en una máquina expendedora o retirar dinero de un cajero requieren un cierto grado de "alfabetización" informática, y prueba de ello es que no son pocos los usuarios que experimentan dificultades cuando se enfrentan por primera vez a un sistema automático. Existen también personas que, por su edad, su formación o por tener alguna discapacidad, no pueden utilizar un teclado y una pantalla, hasta ahora la forma más habitual de interacción con un ordenador.

La extensión del uso del teléfono a capas cada vez más amplias de la población ha favorecido la creación de servicios telefónicos para atender las demandas de los consumidores. Así, han surgido la banca telefónica, los portales de voz desde los que se puede obtener información similar a la que aparece en los portales en la web, los centros de atención al cliente que deben funcionar ininterrumpidamente, los servicios de reserva de entradas para espectáculos, etc. En muchos de estos casos no es factible -o, cuanto menos, no es económico- contar con un operador humano, por lo que es preciso diseñar instrumentos que proporcionen de un modo automático los servicios requeridos.

Además, el pequeño tamaño de aparatos como los teléfonos móviles y de los llamados asistentes personales (PDA) no permite incorporar un teclado y una pantalla en la que se pueda leer o escribir cómodamente un texto, y tampoco es práctico obligar al usuario a llevar consigo -pese a que existen en el mercado- una serie de accesorios para obtener con facilidad la información deseada.

Una posible solución a estos problemas parece venir de la mano de las nuevas interfaces multimodales, como las pantallas táctiles, el reconocimiento de la escritura manuscrita o, muy especialmente, el reconocimiento del habla. No deja de resultar hasta cierto punto paradójico que, siendo el habla el medio de comunicación por excelencia entre las personas, la interacción con los equipos informáticos se lleve a cabo casi siempre mediante la escritura. Las denominadas tecnologías del habla tienen como finalidad facilitar el uso de los ordenadores introduciendo y recibiendo información de modo oral,

y hacer también posible la automatización de los servicios telefónicos como los que anteriormente se mencionaban (Llisterri, 2001a).

Para alcanzar este objetivo es necesario contar, al menos, con tres tecnologías básicas: las que posibilitan que la información escrita se convierta en voz -la síntesis del habla-, las que permiten que un sistema informático realice las tareas que se le solicitan verbalmente -el reconocimiento del habla- y las que facilitan la interacción oral entre una persona y un servicio -los sistemas de diálogo-. Si bien todas ellas se han incorporado a numerosas aplicaciones, debe reconocerse que tienen aún un cierto número de limitaciones: la síntesis, por ejemplo, no reproduce con toda fidelidad el habla humana, el reconocimiento se ve dificultado por los ruidos del ambiente o por factores que dependen del hablante, y los sistemas de diálogo tampoco proporcionan toda la naturalidad propia de una conversación entre dos personas.

Con objeto de abordar algunos de los problemas señalados, se ha consolidado una disciplina, conocida como ingeniería lingüística, que se encarga de aplicar los conocimientos lingüísticos en el desarrollo y la mejora de sistemas informáticos, convirtiendo la lengua en una herramienta al servicio del hombre (Llisterri y Martí, 2002; *Ingeniería lingüística*; Gómez, 2000). Las tecnologías del habla forman parte de este ámbito del saber, y tienen por objeto, como se ha visto, el tratamiento de la lengua oral, considerando no sólo los aspectos técnicos, sino también aquellos que se relacionan con el estudio de las propiedades articulatorias, acústicas y perceptivas del habla, y con los demás niveles lingüísticos que deben tenerse en cuenta en la comunicación verbal.

A pesar de que las tecnologías de habla se han asociado generalmente a las telecomunicaciones -y, especialmente, al procesado digital de señales-, en el desarrollo de un conversor, de un reconocedor o de un sistema de diálogo deben contemplarse una serie de factores que hacen que la presencia del lingüista sea imprescindible, no sólo en el diseño, sino también en la evaluación de dichas tecnologías, aportando conocimientos que inciden en la calidad de la síntesis, del reconocimiento o del diálogo, y en el buen funcionamiento de las aplicaciones, mejorando así el grado de satisfacción del usuario (Aguilar *et al*., 1997; Llisterri, 2002a, b; Llisterri *et al*., 1999).

En este trabajo se pretende, precisamente, exponer el papel que desempeña el conocimiento lingüístico en las tecnologías del habla, centrándose en la conversión de texto en habla (apartado 2), el reconocimiento del habla (apartado 3) y los sistemas de diálogo (apartado 4).

2. La conversión de texto en habla

Un sistema de conversión de texto en habla (TTS o *Text-to-Speech System*) tiene como finalidad la transformación automática de cualquier texto escrito y disponible en formato electrónico en su correspondiente realización sonora (Dutoit, 1997; Dutoit y Stylianou, 2003; Olive, 1998). Un conversor es una aplicación informática que ha de reproducir el proceso que realizaría un lector del texto en el momento de su oralización; por tanto, se le deberá dotar de toda la información lingüística que sea necesaria para ello (Llisterri, 2001b; Llisterri *et al*., 2003a).

2.1. La estructura de un conversor de texto en habla

La estructura de un conversor suele ser modular, de manera que cada uno de sus módulos se ocupa de un aspecto de la transformación que convierte la cadena inicial de caracteres ortográficos en una señal sonora (Figura 1). En los siguientes apartados se expone el tratamiento lingüístico que se lleva a cabo en cada una de las etapas de la conversión: procesamiento previo del texto (2.2), análisis lingüístico (2.3), transcripción fonética automática (2.4), asignación de elementos prosódicos (2.5), constitución del diccionario de unidades de síntesis (2.6) y conversión en valores de parámetros acústicos (2.7).

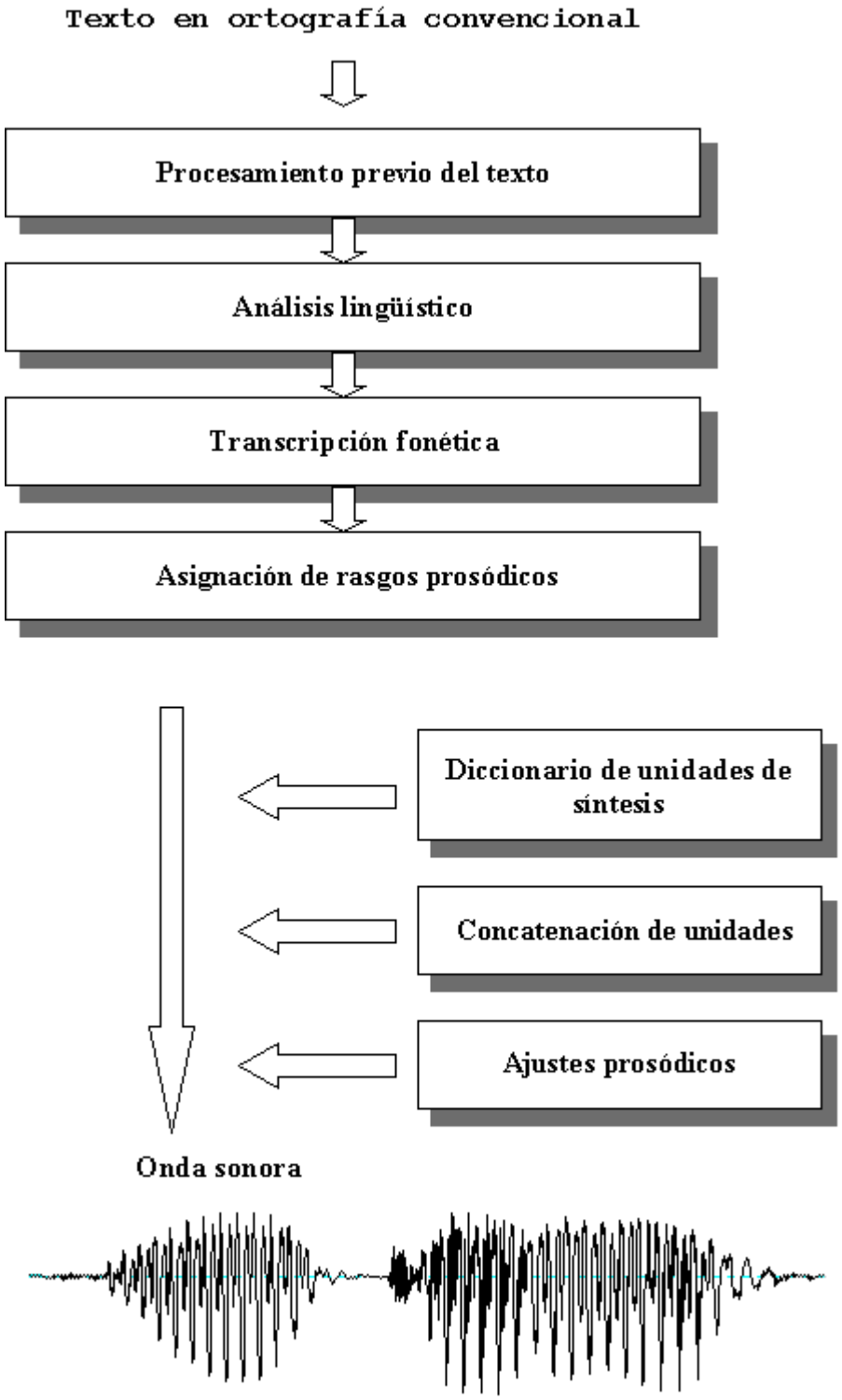

Figura 1: Principales módulos de procesamiento lingüístico en un conversor de texto en habla

2.2. Procesamiento previo del texto

La conversión de texto en habla debe realizarse sobre la representación ortográfica del texto que un hablante leería en voz alta, para lo cual es necesario interpretar y expandir aquellas secuencias de signos que no correspondan a una palabra ortográfica pronunciable: los signos de puntuación con valor no lingüístico, los números, los símbolos, las siglas y las abreviaturas.

Esta labor se lleva a cabo en el módulo de preprocesamiento. La información necesaria es compleja, ya que se trata de reescribir elementos del texto que pueden tener más de una lectura y cuyo uso no siempre está normalizado; por ejemplo, los números telefónicos, que suelen constar de 9 dígitos, pueden aparecer en agrupaciones de dos o tres números separados mediante espacios en blanco, guiones o puntos, o sin que se formen agrupaciones; es posible también, que el prefijo provincial -que en España consta de 2 números en las provincias de Alicante, Asturias, Barcelona, Madrid, Málaga, Sevilla, Valencia y Vizcaya, y de tres números en el resto de las provincias- aparezca escrito entre paréntesis. En cualquier caso, la lectura de los números de teléfono, con sus correspondientes pausas entre grupos de números, vendrá condicionada por cómo estén dispuestos tipográficamente, aunque siempre pueden ser leídos número a número si bien la norma española no lo aconseja (véase la tabla 1).

Expresión numérica	**Expansión**
919998223	nueve uno nueve nueve nueve ocho dos dos tres.
(91) 999 82 23	noventa y uno, nueve nueve nueve, ocho dos, dos tres.
91.999.82.23	nueve uno, nueve nueve nueve, ochenta y dos, veintitrés.
91-999-82-23	noventa y uno, novecientos noventa y nueve, ochenta y dos, veintitrés.
91 999 82 23	

Tabla 1: Lectura de los números telefónicos.

En lo que concierne a los usos no lingüísticos de los signos de puntuación, el sistema deberá tener en cuenta los contextos en los que éstos aparecen para asignarles una interpretación correcta.

El punto se usa tradicionalmente en las expresiones numéricas para separar los millares, los millones y los miles de millones (999.141.179.000), pero en la norma internacional sirve para separar los decimales de la parte entera (3.1416); en la expresión del tiempo separa las horas y los minutos (12.45); en las fechas, las agrupaciones de días, meses y años (4.4.1946); y en los números telefónicos, las agrupaciones de dígitos (93.781.46.56). Además, el punto puede representar el signo de multiplicación, uso en el que alterna con letra *x* (en ambos casos se lee “por”).

Con la coma se ha separado, en la tradición hispana, la parte entera de los decimales (3,1416), y en algunos países americanos los millares, los millones y los miles de millones (999,141,179,000), aunque la norma internacional prefiere la ausencia de signos para este último fin y recomienda escribir los números en agrupaciones de tres cifras separadas por un espacio en blanco (999 141 179 000). Además, la coma puede

aparecer también en la expresión del tiempo para separar las horas de los minutos (12,45).

Los dos puntos representan el signo de división ("entre"), pero se pueden usar para separar las horas de los minutos (12:45), función para la que también se utiliza -además del punto y de la coma, antes citados- el apóstrofo (12'45).

El guión indica el intervalo entre dos números, como es el caso de la numeración de páginas (pp. 23-98), y separa los dígitos de los números telefónicos y las cifras componentes de las fechas (4-4-1946), para lo que también se puede usar la barra (4/4/1946); con este signo se expresan, además, los cocientes de las magnitudes y unidades de medida (60 Km/h).

Para expandir una expresión numérica, el módulo de preprocesamiento primero ha da identificarla: un número cardinal u ordinal (1, 2, 1º, 2ª), una fecha (12-10-1492), una hora (16:30), un número telefónico (91 999 82 23), una cantidad de unidad de medida (10 m., 5 l., 25 €), una cifra escrita en números romanos (MMIII) o un número que forma parte de un sigla (TV3, A3). Una vez identificada la expresión se le ha de asignar una lectura, teniendo en cuenta que no siempre hay una única posible; por ejemplo: 3,40 € = tres euros con cuarenta céntimos / tres euros con cuarenta / tres con cuarenta / tres coma cuarenta.

Así mismo, en el caso de las siglas, las abreviaturas y los símbolos es necesario también identificar el tipo de expresión y resolver los casos de ambigüedad (la abreviatura *col.* puede corresponder a "colección" o a "columna"; la sigla UVI puede referirse a Unidad de Vigilancia Intensiva o a Universidad de Vigo).

Para expandir las abreviaturas y los símbolos se ha de contar con una lista donde consten las palabras que representan. Deberá tenerse en cuenta, además, la concordancia de género y número cuando dichas expresiones acompañan una cantidad numérica (1 cm. = "un centímetro"; 3 cm. = "tres centímetros"; 1 £ = "una libra"; 1 $ = "un dólar").

La lectura de las siglas puede realizarse también mediante su expansión, como se hace con los símbolos y abreviaturas (CCOO = "Comisiones Obreras"), aunque lo más frecuente es deletrearlas (UGT = "u ge te") o, si la sigla tiene una estructura fónica y silábica posible en la lengua, leerla como una palabra (ONU = "onu"). En ciertos casos se podrán combinar ambas formas -lectura y deletreo- para realizar la expansión (CSIC = "ce sic"). Incluso existen ejemplos de siglas que pueden tener más de una expansión (PSOE = "pe soe" / "soe" / "psoe" / "Partido Socialista Obrero Español").

2.3. Análisis lingüístico

La conversión correcta de un texto en habla no es posible sin un análisis morfológico, categorial, sintáctico, semántico y pragmático. El análisis lingüístico es imprescindible, por ejemplo, para proponer la transcripción fonética adecuada en algunas palabras derivadas o compuestas, para determinar el acento en palabras homógrafas, para asignar pausas no marcadas ortográficamente o para generar los contornos melódicos. Es decir, para que un texto pueda ser oralizado apropiadamente debe presentarse como una unidad lingüísticamente coherente y cohesionada, sin ambigüedades involuntarias. La calidad de un conversor mejora si se incorporan las técnicas y herramientas

desarrolladas dentro del ámbito del procesamiento del lenguaje natural, así como sus resultados, en el módulo de análisis lingüístico.

La incorporación de un categorizador y un procesador morfológico, también llamados *POS Tagger* o *Word Class Tagger,* mejora notablemente la conversión de un texto en habla. Se encarga, por una parte, de etiquetar las unidades según la clase de palabras a la que pertenezcan (nombre, verbo, adjetivo, preposición…) y, por otra, de analizar en morfemas su estructura interna.

Básicamente se pueden distinguir dos tipos de procesadores morfológicos: los "supervisados" y los "no supervisados" (van Guilder, 1995). Los procesadores supervisados se basan en un corpus etiquetado. Tanto en un tipo como en otro, la asignación de categorías puede realizarse mediante reglas (*Rule Based Tagging)* o mediante frecuencias de probabilidad (*Stochastic Tagging*). En el primer caso, las etiquetas se asignan según el contexto lingüístico; es decir, se basan en reglas de coaparición de las palabras (*context frame rules*); también son necesarias listas detalladas de términos que constituyen excepciones a las reglas. En el segundo, se utilizan métodos que calculan la frecuencia de aparición de una palabra determinada en un contexto específico. Una variante son los procesadores en árbol (*Tree based Taggers*), que emplean etiquetas complejas como "nombre-verbo-adjetivo" o "verbo-adjetivo". Con esta estrategia es posible aplicar algoritmos de desambiguación que tienen en cuenta información lingüística de distintos niveles.

Incluir un procesador morfológico en los módulos de un conversor de texto en habla permite mejorar la trascripción fonética de determinadas palabras, pues en ocasiones ésta depende de la estructura morfológica. Por ejemplo, en inglés, la pronunciación de la <s> de una palabra que empiece por *dis-* es consecuencia de su estructura morfológica. La grafía <s> del prefijo corresponde a un sonido sordo (*dismiss* [dɪsˈmɪs]); en el resto de casos depende de cada palabra (*dismal* [ˈdɪzməl], *dismay* [dɪsˈmeɪ]). En catalán ocurre algo similar con las palabras que empiezan por "s" y van precedidas de prefijos como *re-*, *contra-* o *pre-*. La grafía <s> se realiza como sorda cuando la palabra empieza por uno de estos prefijos (*resituar* [rəsituˈa]) y como sonora cuando *re-* no es un prefijo (*resultar* [rəzulˈta]).

Debe determinarse también de forma inequívoca la categoría de palabras homógrafas. Por ejemplo, el acento de formas inglesas como *abstract*, *present, progress*, etc. depende de su categoría, ya que como nombres son llanas y como verbos son agudas. En catalán, la palabra *sa* se pronuncia como [ˈsa] cuando es nombre (*sano*) y como [sə] cuando se trata del determinante posesivo femenino (*su*), pues la asignación del timbre vocálico depende del acento. Otro tanto sucede con *segons*, que puede ser nombre ("segundos"), numeral ("segundos") o preposición ("según"). Nuevamente, el tipo de acento y la melodía de la oración dependen de la categoría.

La función de un estructurador, agrupador o analizador sintáctico es "descubrir" las relaciones jerárquicas y funcionales entre las palabras de un texto con el fin de mejorar los resultados del modelado prosódico (Selkirk, 1984; Abney, 1992). Incluso mediante mecanismos relativamente sencillos como el algoritmo *chink and chunk,* que se basa en

la distinción entre elementos funcionales (*chinks*) y elementos léxicos (*chunks*), y que puede aplicarse por simple heurística, se logran resultados muy positivos (Liberman y Church, 1992). Como no todos los límites prosódicos se predicen con facilidad y es imprescindible identificarlos para poder asignar la prosodia, a menudo se requiere un análisis sintáctico más profundo, capaz de emplear contextos amplios y de resolver problemas de ambigüedad estructural.

Por ejemplo, la interpretación y la manifestación fonética de una respuesta como *Si te parece bajo el armario* depende -en el caso nada infrecuente de aparecer escrita sin ninguna coma tras la prótasis-, de si se ha realizado una pregunta como *¿Dónde ponemos la maleta?* (Figura 2)*, ¿Qué piensas bajar ahora?* (Figura 3) o *¿En qué caso dices que podemos poner una repisa ahí arriba?* (Figura 4). En el primer caso *bajo* es preposición, en el segundo verbo y en el tercero adjetivo. En las curvas melódicas reproducidas en las figuras 2, 3 y 4 puede observarse cómo varían tanto el acento de la palabra como la entonación de toda la oración (Llisterri *et al.*, 2003a).

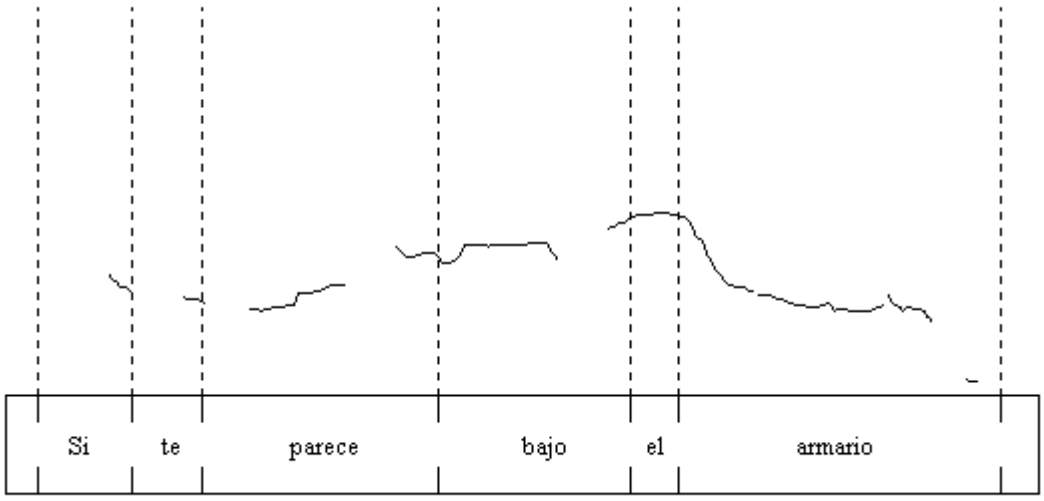

Figura 2: Curva melódica del enunciado *Si te parece bajo el armario*, en la que *bajo* es una preposición

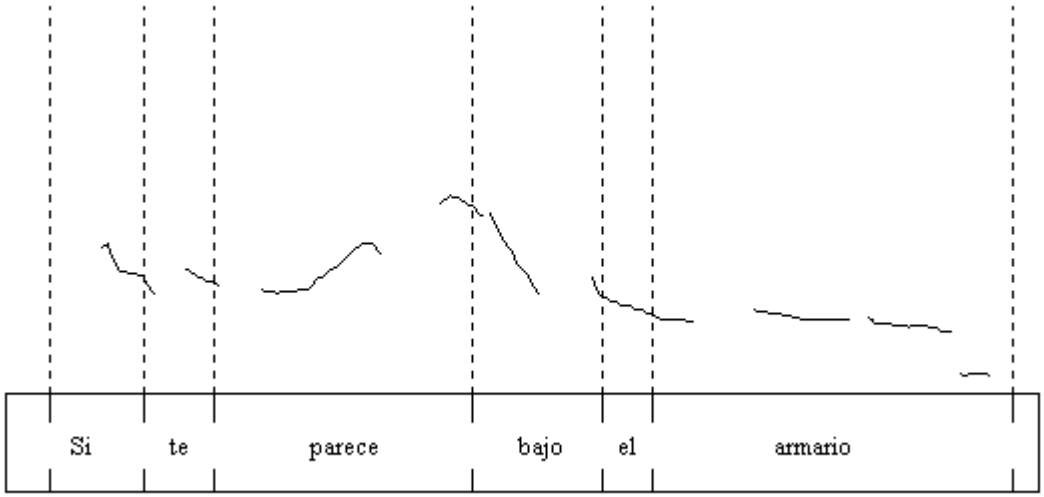

Figura 3: Curva melódica del enunciado *Si te parece bajo el armario*, en la que *bajo* es un verbo

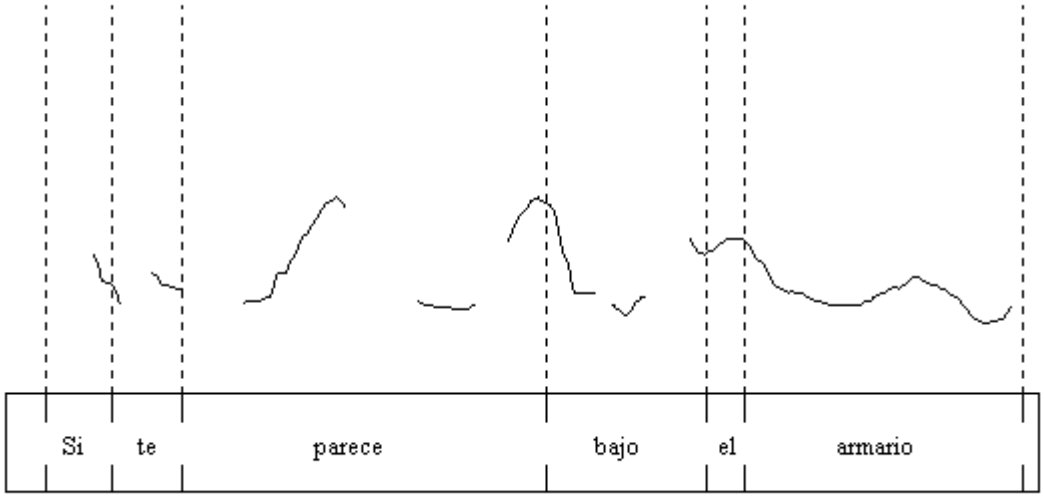

Figura 4: Curva melódica del enunciado *Si te parece bajo el armario*, en la que *bajo* es un adjetivo

En inglés, los adjetivos compuestos, que tienen dos acentos, como *good looking* ['gud 'lukiŋ], pierden el primero cuando funcionan como modificadores del nombre (*A good looking boy looked at me*) y el segundo cuando funcionan como predicativos (*She finds him a good looking boy*) (Jones, 1918: 259; Finch y Ortiz, 1982: 99). Asimismo, los verbos auxiliares suelen ser átonos, excepto -entre otros casos- cuando el verbo principal no aparece: *I may have said so* [ai'meiəv'sedsou], *Yes, I*

have [ˈjɛs aiˈhæv] (Jones, 1918: 271); obsérvese que, además, la pronunciación de la forma átona es distinta a la de la tónica.

En casos de ambigüedad estructural, es decir, en secuencias en las que la disparidad de interpretaciones no viene motivada por la categoría de las palabras sino por las distintas relaciones sintácticas que se establecen entre ellas, es preciso también recurrir al entorno textual. Por ejemplo, *deshilachados* puede ser modificador del nombre o predicativo en *Laura llevaba los vaqueros deshilachados* y es preciso deshacer la ambigüedad para asignar la prosodia pertinente.

También las características léxico-semánticas de los adverbios y, como consecuencia, la relación que establecen con el resto de la oración influyen en la melodía del enunciado. Las oraciones *Sinceramente, no sé qué hay que hacer* y *Desafortunadamente, no sé qué hay que hacer* no presentan el mismo patrón melódico debido al cambio de adverbio. Asimismo, en inglés, no es posible saber la pronunciación del participio *used* sin tener en cuenta su significado, pues se pronuncia [ˈjust] cuando significa *acostumbrado* y [ˈjuzd] cuando significa *utilizado* (Jones, 1918: 187).

Por otro lado, para asignar la prosodia adecuada a los elementos que aportan información nueva y contrastiva (de la Mota, 1995, 1997; Gussenhoven, 2002; Hirschberg, 2002) es necesario realizar además un análisis de la estructura informativa del texto. A pesar de que una posibilidad consiste en marcar tipográficamente tales elementos, en los libros de estilo se desaconseja este procedimiento (Reyes, 1998: 147). Compárese, por ejemplo, la diferente prosodia de una secuencia como *Marcos perdió la llave con* entonación no marcada con la de la misma secuencia como réplica a *Seguro que fue Laura quien perdió la llave*. En inglés, según Jones (1918: 265), en los enunciados que contienen un elemento que ya ha aparecido anteriormente, éste generalmente se desacentúa. Asimismo, el análisis pragmático permite generar de forma adecuada la entonación de enunciados como *thank you,* cuya melodía depende del grado de cortesía que quiera mostrar el hablante.

Dado que la implementación de la información mencionada en este apartado es compleja y costosa, en ocasiones se reduce el procesado lingüístico al mínimo. Como consecuencia, la naturalidad se ve afectada notablemente y, en ocasiones, también la inteligibilidad del texto oralizado.

2.4. Transcripción fonética automática

La transcripción fonética automática tiene como objetivo transformar un texto, escrito en ortografía convencional, en una representación en caracteres fonéticos que refleje la pronunciación de un hablante al leerlo. El primer paso es definir el inventario de alófonos que se va a usar, el cual está condicionado por el modelo de pronunciación elegido y por el tipo de sistema de síntesis que se emplee.

Generalmente, el modelo de pronunciación suele ajustarse al de una variedad estándar, aunque también se podría transcribir, si el tipo de aplicación lo requiere, una pronunciación marcada por rasgos dialectales o sociales (Pachès *et al*., 2000). Así mismo, deberá decidirse la pronunciación de los extranjerismos y la consecuente

inclusión de sonidos pertenecientes a otros sistemas, por ejemplo la [ʃ] de *show* en un conversor en español.

No obstante, sea cual sea la variante de pronunciación que se quiera reflejar, la complejidad de la transcripción depende del sistema de síntesis elegido: en una síntesis por concatenación de difonemas -segmento comprendido entre las partes estables de dos sonidos contiguos- o de unidades mayores no es necesaria una transcripción fonética estrecha, ya que las unidades almacenadas en el diccionario de síntesis recogen el resultado de la coarticulación entre sonidos adyacentes, como es el caso de los alófonos nasales (véase el apartado 2.6); en los sistemas de síntesis por reglas o por concatenación de unidades menores que el difonema, la transcripción deberá ser más estrecha para poder dar cuenta de los fenómenos de coarticulación.

Para pasar de una representación ortográfica a una representación en caracteres fonéticos un sistema automático necesita disponer de la misma información que permite a un hablante leer un texto: interpretación fónica regular e irregular de los grafemas, conocimiento de la estructura silábica de la lengua, de la posición del acento en la palabra y en la frase, y de los procesos fonológicos (Enríquez, 1991; Ríos, 1993, 1994, 1999).

Así, un sistema de transcripción fonética automática deberá tener presente que hay grafemas que representan un solo fonema (en español el grafema <a> siempre representa el fonema /a/); que hay grafemas que pueden representar a más de un fonema, según el contexto en que aparezcan (<g> puede representar los fonema /g/ y /x/); que hay grafemas que pueden representar un conjunto de fonemas (<x> representa el conjunto /ks/); que hay fonemas sin representación fónica (<h>, salvo cuando forma parte del dígrafo <ch>), y grafemas que tienen una interpretación fónica anómala (<x> representa el fonema /x/ en los mejicanismos).

El sistema también deberá tener información sobre los límites silábicos y sobre los fonemas que forman agrupaciones tautosilábicas en ataque o en coda. Por ejemplo, en español -como se ha señalado- el grafema <x> tiene como interpretación fónica regular el grupo fonemático /ks/, que es homosilábico -en coda- cuando está seguido de consonante (*experimento*) y heterosilábico en posición intervocálica (*examen*), pero que nunca aparece en ataque (*xilófono* debe transcribirse como [si'lofono]). Además, el sistema deberá contemplar los fenómenos irregulares que se den en la lengua, como sucede en *sublunar* ([suβ̪.lu.'nar]), donde existe un límite morfológico que impide la formación de un grupo tautosilábico /bl/, como el de [su.'β̪li.me].

Por otro lado, la división silábica es necesaria para asignar el acento léxico. En esta parte de la transcripción se deberá considerar la existencia de palabras átonas (por ejemplo, los pronombres clíticos y los artículos determinados del español), tónicas con un solo acento y tónicas con acento primario y secundario (por ejemplo, los adverbios en -*mente*).

En cuanto a los procesos fonológicos, es esencial tenerlos en cuenta para determinar los alófonos que se han de transcribir. Además, el sistema deberá contemplar los contextos

fónicos en los que operan y poseer las estrategias necesarias para resolver las excepciones de su aplicación. Por ejemplo, en catalán existe un proceso regular de reducción vocálica que no se aplica -entre otros casos- en palabras cultas como *cráter*, *ego* y *esperanto* (Recasens, 1993).

Esencialmente, en la transcripción fonética automática se siguen dos tipos de procedimientos: por reglas o por diccionario. La elección de cualquiera de ellos depende de la mayor o menor regularidad en la interpretación fónica de los grafemas y del acento prosódico a partir de la acentuación gráfica. En aquellas lenguas en las que la pronunciación no puede deducirse regularmente de la ortografía, es conveniente usar un diccionario de unidades previamente transcritas -palabras y/o morfemas-, auxiliado por un conjunto de reglas de concatenación (Allen *et al.*, 1987; Laporte, 1988). En las lenguas con una correspondencia regular entre la representación ortográfica y la pronunciación es más económico seguir una transcripción mediante reglas, complementadas con un diccionario de excepciones que incluiría tanto las pertenecientes al propio sistema (por ejemplo, la silabación de *sublunar*) como las provenientes de otros (los extranjerismos).

Las reglas que suelen utilizarse en los sistemas de transcripción tienen un formato de uso habitual en la lingüística: son reglas de reescritura de aplicación contextual, como la que transcribiría el grafema *c* del dígrafo *ch*: c → [t͡ʃ] / ___ h. Las reglas pueden llevar a cabo diversas operaciones: transformar un signo en otro (n → [m] / ___ v/b/m), insertar un elemento (x → [ks]) o elidirlo (u → ø / g __ e/i).

Según el aspecto de la transcripción que traten, se pueden proponer reglas de distinto tipo: conversión de grafema a fonema o alófono, silabación y acentuación. No obstante, en algunos sistemas, las reglas de silabación y acentuación se aplican en el módulo de preprocesamiento (Castejón *et al*., 1994).

2.5. Asignación de elementos prosódicos: duración, intensidad, pausas y entonación

El módulo prosódico es esencial para generar enunciados más naturales e inteligibles. Los elementos que suelen modelarse son la duración (2.5.1) y la intensidad (2.5.2.) de los sonidos que componen los enunciados, las pausas (2.5.3), y la evolución de la frecuencia del fundamental (F_0) a lo largo del enunciado (2.5.4).

2.5.1. Duración

Uno de los fines del módulo prosódico es dotar a cada uno de los elementos segmentales que componen el habla sintetizada de una determinada duración. Para ello debe tenerse en cuenta la duración de los segmentos en función de factores intrínsecos y extrínsecos (Navarro Tomás, 1918a; Lehiste, 1995).

Cada uno de los sonidos del habla tiene una duración inherente condicionada por el esfuerzo necesario para la realización del gesto que se ha de hacer para articularlo; además, la duración de un mismo sonido es distinta por la influencia del acento, del punto, modo y sonoridad de los segmentos adyacentes, de la estructura silábica, de la posición del segmento en la sílaba y en el enunciado, y de la velocidad de elocución.

Los estudios realizados para el español indican que, en el caso de las vocales, existe una correlación entre la abertura y la duración de las vocales (Navarro Tomás, 1916): la elevación de la lengua es un factor que condiciona la duración intrínseca de estos sonidos. Entre los factores extrínsecos que mayor incidencia tienen en este parámetro se encuentran el acento, la posición de la vocal en el interior del grupo fónico -prepausal o no prepausal-, y la estructura de la sílaba -cerrada o abierta- (Navarro Tomás, 1917; Borzone y Signorini, 1983).

En el caso de las consonantes, tanto Navarro Tomás (1918b) como Borzone y Signorini (1983) las clasifican, por su duración intrínseca -de mayor a menor-, del siguiente modo: africadas, fricativas, oclusivas sordas, vibrante múltiple, nasales, laterales, aproximantes y vibrante simple. En cuanto a los factores extrínsecos, el acento, la posición dentro de la sílaba y la estructura de ésta, y la posición dentro del grupo fónico son algunos de los que mayor incidencia tienen.

Teniendo en cuenta los factores citados, se han desarrollado distintos procedimientos para modelar de la duración en la síntesis. Aunque existen también métodos basados en reglas (Klatt, 1979), la mayoría recurren a métodos estadísticos, y parten de unidades como la sílaba (Campbell, 1992), el difonema (O'Shaughnessy *et al.*, 1988) o el segmento (Klatt, 1979) para determinar el patrón temporal de un enunciado. Entre las principales técnicas estadísticas usadas para modelar la duración se encuentran las de regresión paramétrica -como los modelos aditivos y multiplicativos (Möbius y van Santen, 1996)-, las de regresión no paramétrica -como los árboles de clasificación y regresión (CART) y las redes neuronales (Riedi, 1998)-.

2.5.2. Intensidad

En el habla natural no todos los sonidos de un enunciado se realizan con la misma intensidad. Las vocales de las sílabas tónicas, además de tener, generalmente, una mayor duración y valores más altos de F_0 que las sílabas átonas, también se caracterizan por poseer una mayor amplitud (Canellada y Madsen, 1987). De igual modo, los elementos realzados fonológicamente -debido a que aportan información nueva contrastiva. presentan un incremento de la intensidad (véase la figura 5). En tales casos, aumenta también la duración de la palabra y se inhibe el habitual desplazamiento del valor máximo de F_0, que aparece con un valor más alto (de la Mota, 1995).

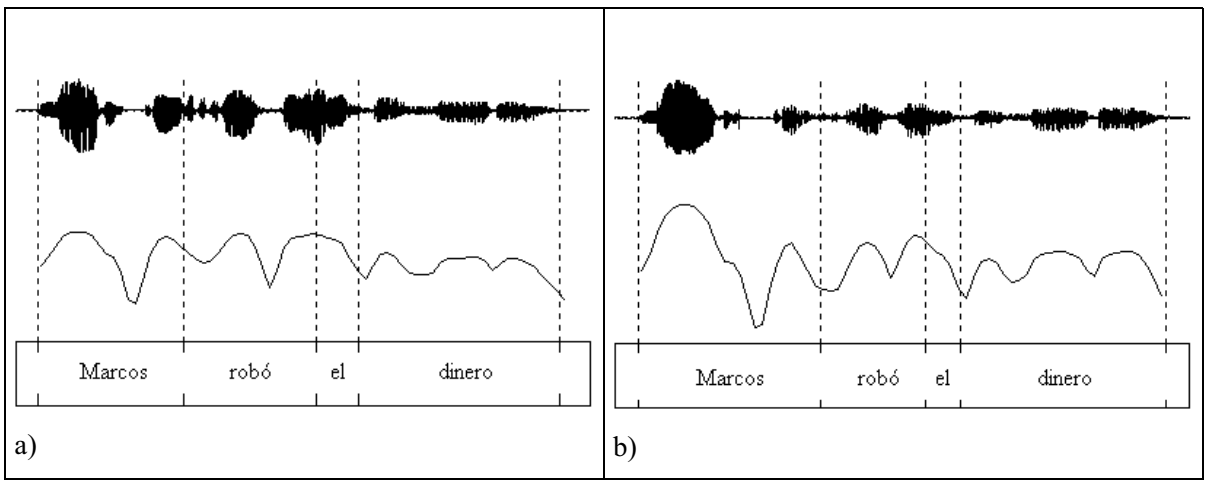

Figura 5: Oscilogramas y curvas de intensidad correspondientes al enunciado "Marcos robó el dinero" sin realce fonológico en el sujeto (a) y con él (b).

La mayoría de los conversores de texto en habla, sin embargo, prescinden de modelar la energía y, en el mejor de los casos, únicamente normalizan la intensidad de las unidades concatenadas para que no se perciba una distorsión en la señal.

No obstante, en los últimos años algunos sistemas de conversión de texto en habla han incorporado un módulo de asignación de intensidad con el fin de reproducir la voz que corresponde a un cierto estado de ánimo (Montero *et al.*, 1999), o para realzar determinada información en un enunciado (Bartkova *et al*., 1993).

2.5.3. Pausas

Aunque en la oralización automática de un texto no pueden darse pausas de origen fisiológico, puesto que no existe fonación, su presencia en lugares lingüísticamente relevantes contribuye a emular la actuación natural de un locutor humano y garantiza la inteligibilidad. Las pausas dividen el enunciado en grupos fónicos y para predecir su localización no sólo debe tenerse en cuenta la puntuación, sino también la organización del texto en constituyentes sintácticos y prosódicos, la duración de las secuencias y la velocidad de elocución. En algunos límites sintácticos suelen realizarse pausas, como entre los miembros de una enumeración o entre un sujeto antepuesto y su predicado (Navarro Tomás, 1945), mientras que existen partes del discurso que constituyen unidades gramaticales, tonales y de sentido muy cohesionadas, como la combinación de artículo y sustantivo o los tiempos compuestos de los verbos (Quilis, 1981).

La tarea del módulo pausador es determinar dónde se insertan las pausas -ortográficas y no ortográficas- y asignarles la duración adecuada. En cuanto a las pausas marcadas ortográficamente, la primera tarea es distinguir los signos de puntuación empleados para estructurar el discurso de los signos que aparecen en expresiones numéricas, siglas o abreviaturas:

> *La hispanoargentina Repsol YPF que preside Alfonso Cortina ganó 1.952 millones, el 90,4%, si bien en estos resultados se incluyen las importantes plusvalías de la venta del 24% de Gas Natural. (El Periódico, 9-3-03)*

En español el número de sílabas de un grupo fónico suele oscilar entre cinco y diez (Navarro Tomás, 1945), así que cuando el número de sílabas existentes entre dos signos de puntuación es superior quizás sea conveniente manejar información lingüística para generar otras pausas. Dicha información puede ir desde una simple lista de palabras que acostumbren a inducir la presencia de una pausa hasta un análisis sintáctico complejo. La oralización del texto siguiente, por ejemplo, requiere la producción de pausas, pero no hay ningún signo de puntuación que lo indique.

> *Los cuatro agentes que mintieron sobre la conducta de dos jóvenes detenidos durante las manifestaciones antiglobalización del 15 de marzo del 2002 deshonran al cuerpo policial y sus jefes tienen la obligación moral y profesional de expedientarlos y de ser los primeros en llevarlos ante el juez. (El Periódico, 9-3-03)*

Una vez decidida la localización de las pausas, debe asignárseles una duración adecuada. Por un lado, la longitud de las pausas suele estar en proporción inversa a la relación que exista entre los fragmentos de discurso que separan: a mayor vinculación, menor duración (Navarro Tomás, 1945). Como consecuencia, la gradación que se observa en la extensión de las pausas marcadas ortográficamente es la siguiente: puntos suspensivos > punto > punto y coma > coma (Moreno y Martínez, 1988; Puigví *et al*., 1994). Por otro lado, dada la cantidad de pausas opcionales que existen en la lectura natural, parece existir una relación inversa entre el número de pausas y la duración de las mismas (López y Martínez, 1988).

2.5.4. Entonación

Las variaciones melódicas permiten reconstruir la estructura gramatical y pragmática del texto y proporcionan información sociocultural, geolingüística y psicológica acerca del locutor. En un sistema automático, un buen modelado de la entonación, y de la prosodia en general, proporciona una sólida inteligibilidad, no ya segmental sino textual, y simula la voz de un hablante humano capaz de comunicar de forma expresiva. La manifestación acústica de las fluctuaciones melódicas que deben reproducirse para dar cuenta de fenómenos como las inflexiones tonales, las junturas, los acentos o el escalonamiento descendente es la variación temporal de la frecuencia fundamental (F_0). Como indican Sproat *et al*. (1999: 22), esta tarea plantea dos problemas: por un lado, hay que predecir la situación en el texto de los fenómenos relevantes melódicamente -que pueden ser representados de forma simbólica- y, por otro, debe encontrarse el modo de asignar los valores de F_0 correspondientes a cada fenómeno de manera que sea posible la síntesis.

Desde que en 1966 Mattingly concibió el primer algoritmo para determinar un contorno melódico -que fue implementado en el sintetizador por reglas de Holmes y sus colaboradores (Klatt, 1987)-, se han propuesto métodos diversos para generar los movimientos de la frecuencia fundamental (Llisterri *et al*., 2003b). En general, puede decirse que se siguen tres estrategias distintas: almacenar previamente patrones melódicos, obtener curvas melódicas a partir de sistemas estadísticos -como Modelos Ocultos de Markov o redes neuronales- y predecir la forma de la curva melódica de forma simbólica mediante conjuntos de reglas (Beaugendre, 1996). Lo habitual, no obstante, es combinar métodos en parte basados en datos (*data-driven*) y en parte en reglas (*rule-based*). Independientemente del uso que cada modelo haga de la información prosódica, ésta se obtiene, en todos los casos, a partir de bases de datos de habla natural (Batliner *et al*., 2001). La mejora que se consigue en un conversor tras la implementación del módulo encargado de asignar la melodía constituye un incentivo para proseguir en la investigación básica, ya que quedan todavía muchas cuestiones abiertas.

2.6. Constitución de un diccionario de unidades de síntesis

La síntesis por concatenación de unidades es una de las estrategias más usadas en los conversores de texto en habla actuales. En este tipo de sistemas es necesaria la constitución de un diccionario de unidades de síntesis que permita la generación de cualquier enunciado de una lengua determinada mediante un número limitado de

segmentos fónicos. A pesar de que las unidades que tradicionalmente se han utilizado en la síntesis por concatenación son los difonemas, también existen otras como: el fonema, el trifonema -unidad que comprende dos difonemas adyacentes-, la sílaba o la palabra. La selección de una u otra unidad depende básicamente de dos criterios: el tamaño necesario para el almacenamiento de la base de datos, y la distorsión que se genera al concatenar dichas unidades. En la actualidad se emplea una técnica que permite seleccionar de un corpus de unidades de tamaño variable (*unit selection*) aquellas que son óptimas para la concatenación (Hunt y Black, 1996).

Para la constitución del diccionario de unidades de síntesis se ha de partir del inventario de fonemas y alófonos de la lengua. El número de unidades que constituye el diccionario depende de las características acústicas consideradas para cada alófono y de sus variantes contextuales. Por ejemplo, como se puede observar en la tabla 2, en español existen ocho alófonos nasales distintos [m], [ɱ], [n̪+], [n̪], [n], [nʲ], [ɲ] y [nˠ], y sólo tres son distintivos /n/, /m/ y /ɲ/. Para decidir cuáles deben formar parte del inventario debe tenerse en cuenta el modelo de pronunciación y la economía del inventario, además de los criterios mencionados.

Alófonos nasales	**Ejemplos**
[m]	a<u>m</u>bos
[ɱ]	e<u>n</u>fermo
[n̪+]	a<u>n</u>zuelo
[n̪]	ca<u>n</u>to
[n]	e<u>n</u>lace
[nʲ]	a<u>n</u>cho
[ɲ]	a<u>ñ</u>o
[nˠ]	ma<u>n</u>go

Tabla 2: Alófonos nasales en español.

Además de las unidades propias de la lengua, también pueden incorporarse otras unidades fónicas ajenas a la lengua que aparecen en la pronunciación de términos extranjeros (*jeep*, *pizza*, *show*).

Una vez definidas las unidades del diccionario, es necesario diseñar un corpus que contenga dichas unidades y grabarlo. En la elección del locutor se consideran los siguientes factores:

1. El sexo del hablante: la utilización de una voz masculina o femenina según el tipo de aplicación del sintetizador hace que el usuario se sienta más cómodo a la hora de recibir determinado tipo de información.

2. Las características dialectales del hablante y las posibles interferencias lingüísticas con otras lenguas: habitualmente, se intenta generar un modelo de habla que siga la variante de pronunciación estándar, salvo en aquellos casos en los que se pretenda reproducir el habla propia de un determinado dialecto. Por ejemplo, si una aplicación desarrollada en español se utiliza en la Península, se intentará que el locutor no tenga características tales como la velarización de la lateral en posición postnuclear (propia del español hablado en Cataluña), la velarización de la nasal en posición prepausal (característica del español hablado en Galicia) o la aspiración de la fricativa alveolar en posición postnuclear (rasgo que se encuentra en Andalucía). En cambio, si la aplicación se emplea también en Hispanoamérica, deberán tenerse en cuenta características dialectales del español de América (Villarrubia *et al*., 2002). Por otro lado, la selección de un locutor bilingüe libre de interferencias puede ser beneficiosa para obtener una misma voz en dos lenguas diferentes, como sucede, por ejemplo, en el caso del castellano y el catalán en el sistema Actor comercializado por Loquendo.
3. La capacidad del hablante para articular unidades que no son propias de la lengua: como ya se ha mencionado, el conversor debe estar preparado para pronunciar términos extranjeros, por esta razón, el locutor tiene que saber articular alófonos que no formen parte de su sistema fonológico.

Aparte de los requisitos previamente mencionados, el locutor debe ser capaz de realizar las variaciones prosódicas necesarias para la grabación del corpus de unidades.

A fin de obtener unas condiciones óptimas, la grabación debe realizarse en una cámara anecoica e insonorizada y desarrollarse en varias sesiones, de modo que se evite el cansancio del locutor y el posible cambio de las características articulatorias y acústicas de la voz debidas a la fatiga.

Finalmente, es imprescindible la presencia del lingüista en el momento de la grabación para asegurar una pronunciación correcta y controlar el ritmo, las pausas y la entonación de la lectura. De lo contrario, es posible que se cometan errores que obliguen a repetir la grabación una vez finalizado el proceso, que algunas unidades sean inservibles o que una realización prosódica inadecuada tenga consecuencias negativas en la calidad final de la síntesis.

2.7. Conversión en valores de parámetros acústicos

La última fase en la conversión de texto en habla es la generación de la onda sonora. En los sistemas de síntesis por reglas los parámetros de control del sintetizador pueden ser de tipo articulatorio o acústico. En el primer caso -los de tipo articulatorio- las reglas simulan el mecanismo de producción del habla mediante un conjunto de parámetros que se relacionan con los movimientos del tracto vocal durante la articulación. En el segundo caso -los de tipo acústico-, las reglas se extraen a partir de los valores de frecuencia, duración e intensidad obtenidos del análisis de un corpus de habla natural, separando las características de la fuente y las del filtro.

En los sintetizadores por concatenación primero se seleccionan las unidades óptimas extraídas de un corpus grabado (véase el apartado 2.6). Posteriormente, si es necesario, se modifican los valores de los parámetros acústicos para evitar la distorsión que se

produzca al concatenar las unidades y poder obtener una mayor calidad en el sistema. En esta fase es importante no sólo considerar las características segmentales, sino también los rasgos prosódicos que pueden diferenciar, por ejemplo, un enunciado declarativo de uno interrogativo, o una sílaba átona de una sílaba tónica, además de los cambios en la intensidad y en la calidad de voz.

3. El reconocimiento del habla

Un sistema de reconocimiento realiza la tarea inversa a la de un conversor, ya que convierte la señal sonora del habla en una representación simbólica que, en muchas ocasiones, será un texto escrito (Bernstein y Franco, 1996; Zue *et al*., 1997; Deroo, 1999; Kurzweil, 1998; Lamel y Gauvain, 2003). La aplicación más conocida del reconocimiento se encuentra en los productos de dictado automático, que permiten redactar un documento empleando la voz en lugar de utilizar el teclado de un ordenador. También es esencial la existencia de un reconocedor para la creación de servicios telefónicos, tal como se pone de manifiesto en el apartado 4.

3.1. La estructura de un sistema de reconocimiento del habla

Los reconocedores pueden considerarse, genéricamente, como unos sistemas que, en una primera etapa, aprenden de un extenso corpus de habla y, en el momento de enfrentarse a un nuevo enunciado, lo comparan con los datos que previamente han extraído de este corpus. Por ello, el desarrollo de un sistema de reconocimiento automático de habla se inicia en la fase que se conoce como entrenamiento, en la que se proporciona al sistema toda la información necesaria para efectuar después el reconocimiento. Dicha información se incorpora a los módulos que se muestran en la figura 6 y que constituyen, a grandes rasgos, los elementos básicos de un sistema de reconocimiento de habla.

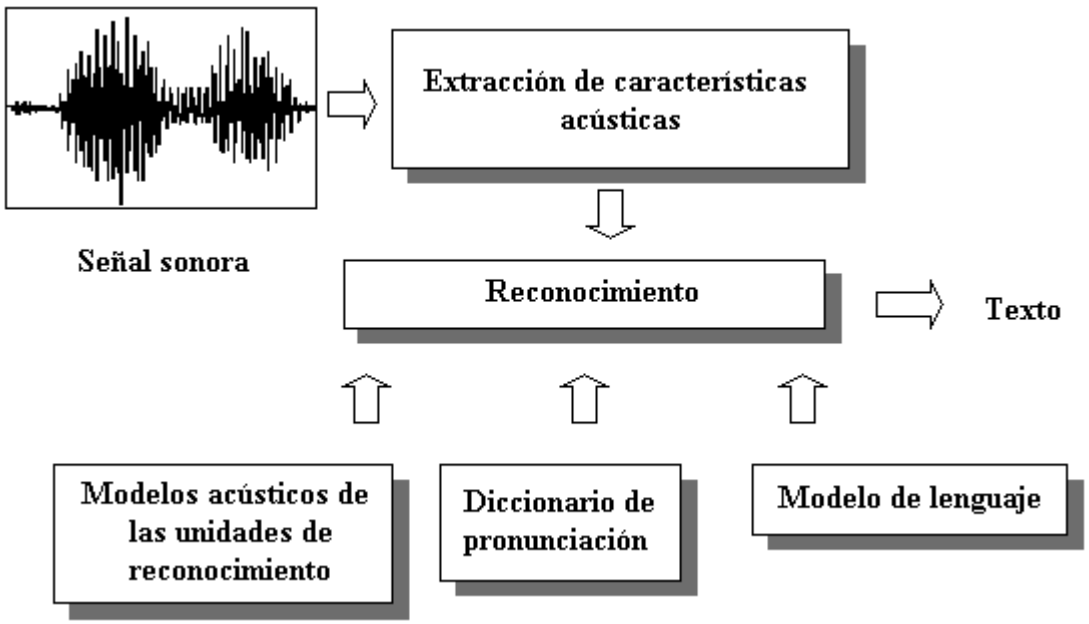

Figura 6: Principales módulos de un sistema de reconocimiento de habla

En primer lugar, la señal sonora se analiza para extraer los parámetros acústicos establecidos en el momento de diseñar el sistema (Nadeu, 2001) y, después, se compara con los modelos acústicos de las unidades de reconocimiento que se han almacenado previamente; la decisión final suele tomarse con la ayuda de las reglas gramaticales que constituyen el modelo de lenguaje, en las que se define la probabilidad de las secuencias de palabras que pueden encontrarse en una aplicación o en un dominio.

3.2. El corpus de entrenamiento

Como se ha indicado, una de las primeras actividades a la hora de desarrollar un sistema de reconocimiento es diseñar y recoger lo que se conoce como corpus de entrenamiento (o de aprendizaje), a partir del cual se adquirirá la información necesaria para crear modelos de cada una de las unidades de reconocimiento, análogas, en ocasiones, a las utilizadas en la síntesis (véase el apartado 2.6). Mientras que para obtener las unidades de reconocimiento es preciso un corpus oral, para crear la gramática o modelo de lenguaje del reconocedor se parte de un corpus textual, mediante el que se establecen las probabilidades de aparición de una palabra en una determinada posición.

Al igual que en síntesis se determinan las unidades necesarias para la generación de enunciados, en reconocimiento se definen aquellas que el sistema utilizará para convertir la señal acústica en un texto. Sin embargo, existe una diferencia esencial entre ambas tecnologías, ya que si en la conversión de texto en habla las unidades se extraen de la grabación de un único locutor, un reconocedor debe estar preparado para tratar no sólo la variación en las realizaciones fonéticas de una misma persona (variación intralocutor), sino también las realizaciones fonéticas de un gran número de usuarios (variación interlocutor) si, por ejemplo, quiere aplicarse para automatizar un servicio de información telefónica.

Por tal motivo, el corpus de entrenamiento de un reconocedor debe contener la mayor variedad posible de realizaciones para que puedan crearse los modelos -"plantillas" o representaciones internas que posee el sistema- de cada una de las unidades, reflejando, entre otros elementos, la variación individual de las voces, los distintos acentos debidos a factores geográficos o sociolingüísticos y las diferencias en la velocidad de elocución que puedan darse entre hablantes. La variación fonética y su correcto modelado es, pues, uno de los primeros problemas lingüísticos que deben abordarse en el diseño de un sistema de reconocimiento (Strik y Cucchiarini, 1999).

Una de las tareas del lingüista consiste, por tanto, en definir las zonas geográficas en las que es necesario recoger muestras, estableciendo, con la ayuda de criterios demográficos, el porcentaje de hablantes que representarán a cada una de las áreas seleccionadas. Un corpus de entrenamiento tiene que incluir necesariamente hablantes de las principales variantes dialectales de la lengua para que el futuro usuario del reconocedor no encuentre dificultades debidas a su acento como consecuencia de la falta de muestras del mismo en el corpus. En el caso del español, por ejemplo, es necesario contemplar no únicamente las diversas variedades peninsulares, sino también tomar en consideración las diferencias de pronunciación existentes en el español de América (Caballero y Moreno, 2001; Villarrubia *et al*., 2002).

Por otra parte, el corpus de entrenamiento debe ser exhaustivo en lo que se refiere a la aparición o cobertura de las unidades de reconocimiento. Tal como se ha expuesto en el caso de la síntesis, se define, en primer lugar, el inventario de alófonos de la lengua y, en segundo, las posibles combinaciones entre ellos formando, por ejemplo, difonemas (véase el apartado 2.6). Deben igualmente tenerse en cuenta factores relacionados con la frecuencia de aparición, ya que cada unidad ha de estar representada un mínimo de veces en el corpus para que la creación de los modelos o plantillas a las que nos referíamos sea fiable. En este sentido, la labor del fonetista se centra en la determinación del inventario de alófonos y en el estudio de sus posibles combinaciones teniendo en cuenta las reglas fonotácticas de la lengua.

3.3. Segmentación y transcripción del corpus de entrenamiento

El aprendizaje por parte del sistema de reconocimiento requiere que el corpus se encuentre segmentado y etiquetado, es decir, que a la señal sonora grabada se añadan marcas que indiquen el principio y el final de cada sonido y de cada unidad de reconocimiento, así como la correspondiente transcripción fonética, sincronizada con la representación ortográfica.

Es necesario para tal fin establecer criterios coherentes de segmentación, de etiquetado y de transcripción, partiendo del conocimiento de las propiedades acústicas y de la variabilidad fonética de los sonidos de la lengua en cuestión. Aunque la segmentación y el etiquetado suelen realizarse de manera automática, es necesaria una revisión manual a cargo de un fonetista experto para corregir los errores que inevitablemente se producen debido a la variación fonética existente entre los diferentes locutores.

3.4. Diccionarios de pronunciación

Un reconocedor puede incorporar un diccionario en el que se encuentran transcritas fonéticamente las palabras que aceptará el sistema, incluyendo además las variantes de pronunciación que se han localizado en el corpus de aprendizaje. La tarea del lingüista consiste, en este caso, en la determinación de la forma "canónica" de cada palabra y en la definición de reglas fonéticas que relacionen esta forma con las posibles pronunciaciones alternativas, tanto las que aparecen en el corpus como las que sean previsibles en función del conocimiento de la variación fonética de la lengua.

3.5. La información fonética en el reconocimiento del habla

Si bien los reconocedores de habla operan esencialmente con técnicas estadísticas basadas en la comparación de patrones, parece existir un cierto consenso sobre la utilidad de incorporar información fonética -segmental y suprasegmental- al proceso de reconocimiento. Por este motivo, se han formulado, desde los primeros trabajos de Zue (1983, 1985) múltiples estrategias para extraer del enunciado que el sistema debe reconocer determinados datos de tipo fonético que puedan ser útiles para los posteriores módulos que intervienen en el proceso del reconocimiento.

La información fonética segmental se ha planteado como un complemento o una alternativa a los parámetros puramente acústicos que, en el primer módulo de un reconocedor, se emplean para analizar la señal (Figura 2). Los rasgos fonéticos de tipo articulatorio (Frankel y King, 2001) o acústico (Koreman y Andreeva, 2000), han sido

los que han recibido más atención, tanto considerados independientemente como en combinación (Kirchhoff *et al*., 2002). Es interesante señalar que, en los momentos de mayor auge de los sistemas expertos, se intentó el reconocimiento imitando las estrategias de lectura de espectrogramas adoptadas por fonetistas, buscando así formalizar la relación entre los indicios acústicos presentes en la onda sonora y sus correlatos fonéticos (Lamel, 1993). Otras aproximaciones se han orientado hacia los rasgos distintivos (Koreman *et al*., 1999) o hacia el tratamiento de problemas específicos como puede ser el de la coarticulación.

Un segundo problema relacionado con la información fonética necesaria en el reconocimiento es la variación tanto interlocutor como intralocutor que se mencionaba en el apartado 3.2. Se han realizado diversas propuestas para incorporar y tratar adecuadamente esta variación (revisadas en Strik y Cucchiarini, 1999), partiendo, por ejemplo, del análisis detallado del corpus de entrenamiento e intentando sistematizar la variación contextual de los alófonos (Fosler-Lusier *et al*., 1999).

En lo que se refiere a los elementos suprasegmentales, la investigación se ha centrado principalmente en dos niveles: el léxico, incorporando al reconocimiento información sobre los correlatos acústicos del acento en la palabra, y el oracional, estudiando los correlatos prosódicos de las fronteras entre constituyentes (Llisterri *et al*., 2003b).

El acento léxico ha sido uno de los aspectos que más atención ha recibido en el campo del modelado prosódico para el reconocimiento del habla (Rubio y Milone, 2002; Wang, 2001). Las primeras aproximaciones se orientaron a detectar automáticamente el patrón acentual de una palabra aislada con dos objetivos: reducir el número de palabras acústicamente similares en sistemas con un vocabulario muy amplio, y diferenciar pares mínimos que únicamente se distinguen por el acento. El reconocimiento del habla espontánea presenta, en cambio, el problema de que no todas las sílabas léxicamente acentuadas en palabras aisladas lo son en el discurso continuo; por otra parte, al acento léxico se superpone el acento de frase, con lo que la dificultad para reconocerlo es aún mayor. Algunas propuestas se han basado en el uso de modelos acústicos distintos para unidades acentuadas y no acentuadas, de modo que se pueda contemplar conjuntamente el efecto del acento en la duración, la intensidad y la frecuencia fundamental. En conjunto, los resultados obtenidos al incorporar información sobre el acento léxico al reconocimiento parecen indicar que su principal utilidad reside, al menos en el caso de las lenguas de acento libre, en la reducción de errores causados por el módulo de análisis acústico, ya que la determinación de la sílaba tónica contribuye a disminuir el número de palabras posibles entre las que el reconocedor debe elegir en un determinado punto del enunciado (Wang, 2001).

Por lo que respecta a la detección automática de indicios prosódicos en el ámbito oracional, trabajos como los de Pagel (1999) o Batliner *et al*. (2001) ponen de relieve la importancia de las fronteras prosódicas en el reconocimiento del habla espontánea y destacan la necesidad de incorporar un analizador sintáctico (véase el apartado 2.3) para evitar las dificultades que surgen en el nivel puramente acústico.

A pesar del interés que despierta el tema y del amplio número de trabajos realizados, los sistemas de reconocimiento no hacen aún un uso completamente eficaz de la información prosódica. Como indica Cassidy (2001) el principal problema radica en que

"por el momento, no sabemos cómo extraer información de la señal acústica de un modo fiable"; parece pues que es conveniente una mayor aproximación entre los expertos en fonética y quienes se dedican al desarrollo de sistemas de reconocimiento.

4. Los sistemas de diálogo

Actualmente existen muchas aplicaciones en las que una persona puede interactuar con un ordenador utilizando un medio escrito, la transmisión oral, o mecanismos sensibles al tacto. En este tipo de aplicaciones -aunque se puedan consultar agendas personales, oír mensajes de correo electrónico o realizar determinadas gestiones bancarias (Marx y Schmendat, 1996; Sadek *et al.*, 1996; Hirschman *et al.*, 1993; Danieli y Gerbino, 1995; Walker *et al.*, 1997; Tapias, 2000)- no se da propiamente una interacción: el usuario solicita la información que requiere y la aplicación se limita a emitir un mensaje de respuesta sin que se produzca un auténtico intercambio. Sin embargo, en los sistemas de diálogo existe una relación directa entre el usuario y la máquina, estableciéndose una interacción oral semejante a la de una situación comunicativa real, motivo por el cual algunos autores emplean la expresión "interfaces conversacionales" en este contexto (Zue, 1997).

4.1. La estructura de un sistema de diálogo

Los sistemas de diálogo tienen como finalidad que un usuario, a menudo a través del teléfono o de un micrófono y sin un intermediario humano, pueda acceder automáticamente a una determinada información, realizar una transacción o conseguir la ejecución de determinadas órdenes (Giachin, 1997; Gibbon *et al*., 2000; Minker y Bennacef, 2001). Dado que el medio utilizado es la lengua oral, se requiere la integración de diferentes componentes, como se puede observar en la figura 7:

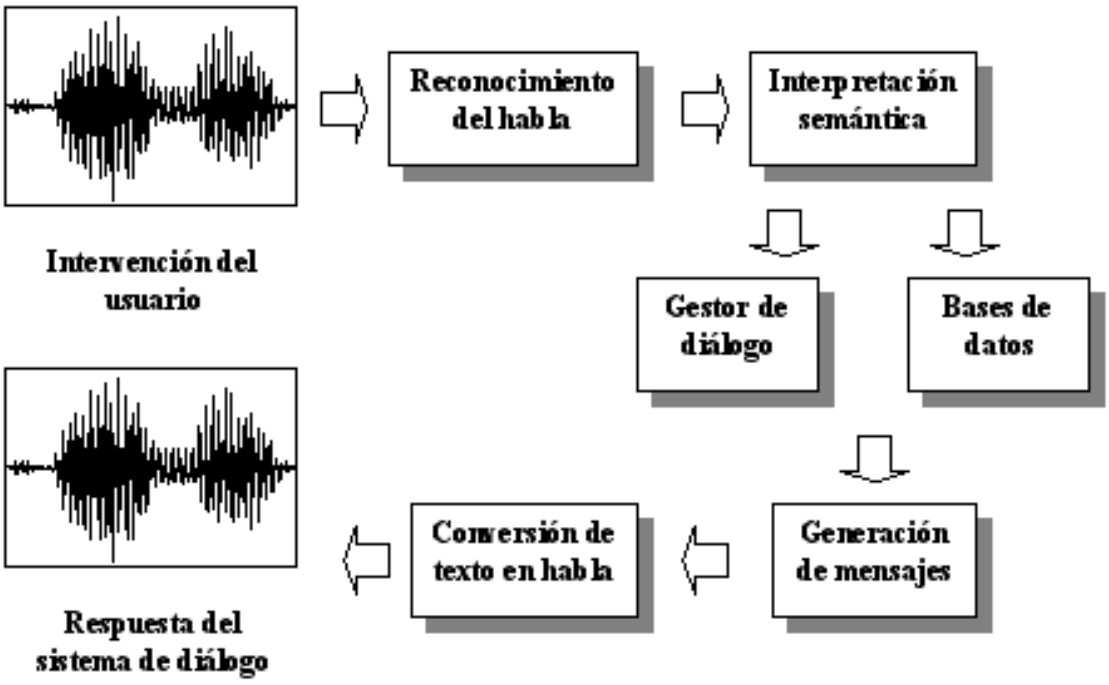

Figura 7: Principales módulos de un sistema de diálogo

Cada uno de los módulos que forman parte de un sistema de diálogo tiene una función específica:

- El reconocimiento de habla. Este módulo debe "reconocer" el mensaje que recibe del usuario para poderlo interpretar semánticamente (véase el apartado 3).
- La interpretación semántica. El sistema analiza la intervención del usuario y la interpreta. El descodificador o módulo de interpretación semántica debe ser capaz de resolver el significado de un determinado tipo de expresiones relacionadas con la aplicación que va a tener ese sistema; por ejemplo, para que un sistema de acceso automático a información de ferrocarriles sea eficaz deben interpretarse periodos temporales como *a media mañana*, *a última hora de la tarde*, *antes de la hora de la comida*.
- El gestor de diálogo. Este módulo controla las interacciones entre el sistema y el usuario mediante unas estrategias de diálogo diseñadas para extraer toda la información necesaria antes de pasar a otras etapas.
- Consulta a la base de datos. Una vez que el sistema ha identificado lo que desea el usuario, accede, si es preciso, a la base de datos y genera un mensaje de respuesta adecuado.
- El módulo de conversión de texto en habla. Es el encargado de proporcionar un mensaje oral al usuario, de manera que éste acabe obteniendo una respuesta adecuada a sus necesidades (véase el apartado 2).

La recolección de diferentes muestras orales constituye una herramienta imprescindible en las tecnologías de habla; prueba de esto es la creación de corpus teniendo en cuenta diferentes situaciones comunicativas: habla espontánea, diálogos, monólogos, etc. (Llisterri, 1999). Sin embargo, para el desarrollo de un sistema de diálogo pocas veces se pueden aprovechar recursos ya existentes, puesto que se necesita, por un lado, un corpus relacionado estrechamente con la aplicación (corpus "persona-persona") y, por otro, un corpus que simule el funcionamiento del sistema (corpus "Mago de Oz"). Como veremos a continuación, ambos corpus tienen finalidades diferentes en la creación de un sistema de diálogo.

4.2. Estudio de corpus de interacciones naturales entre personas

El corpus persona-persona consta de un conjunto de diálogos en los que dos personas interaccionan en una situación de habla semejante a aquella en la que se va a aplicar el sistema. El análisis de este tipo de corpus sirve para llevar a cabo las siguientes tareas:

1. Establecer qué tipos de consulta son más frecuentes para restringir la clase de información que proporcionará el sistema.

Las grabaciones del corpus persona-persona suelen llevarse a cabo en franjas horarias diferentes y en días distintos para obtener una muestra representativa de los tipos de consulta que se realizan durante diferentes periodos. La delimitación de la tarea suele realizarse a partir del porcentaje de aparición en el corpus de los diversos tipos de consulta, ya que resulta más eficaz crear un buen sistema de diálogo que pueda proporcionar información sobre las consultas más frecuentes (por ejemplo, los horarios

en un servicio de información de trenes) que no intentar desarrollar un sistema que intente satisfacer consultas muy variadas pero difíciles de prever. En los casos en los que se restringe la tarea, el sistema debe informar al usuario del tipo de datos o servicios que le puede proporcionar.

2. Elaborar escenarios para evaluar el funcionamiento de un prototipo del sistema.

Un escenario simula una situación que se daría en el funcionamiento real de un sistema de diálogo, y se diseña para proporcionar a las personas que participarán en las pruebas para evaluar el prototipo del sistema una indicación o unas pautas sobre cuál ha de ser su papel. Los escenarios se elaboran teniendo en cuenta la relación entre los participantes en el acto comunicativo: el usuario que solicita una determinada información y el operador humano que se la facilita. Suelen definirse tres tipos de escenarios: los cerrados, los semicerrados y los abiertos.

En los escenarios cerrados se presenta una situación totalmente dirigida. El usuario debe pedir una determinada información a partir de las instrucciones que se le proponen en el escenario. La situación que se muestra a continuación es un ejemplo de ello:

> *Usted vive en Paseo de Gracia de Barcelona y tiene que ir al médico a la Mutua de Tarrasa a las nueve. Debe pedir información sobre los horarios de los trenes que van hacia Tarrasa para estar allí media hora antes.*

Como podemos observar en el ejemplo, la persona que interviene en las pruebas como potencial usuario debe ceñirse a la situación propuesta en el escenario: la estación de origen, la estación de destino y la hora de llegada.

En los escenarios semicerrados el usuario conoce parte de la información que necesita para realizar determinado tipo de consulta. En este tipo de escenarios se tiene mayor libertad para solicitar la información:

> *Usted vive cerca de la estación Reina Elisenda y tiene una cita con su gestor a una hora determinada. La gestoría se ha trasladado a Vía Augusta, número 215, y de las tres estaciones que pasan por esta calle no sabe cuál de ellas queda más cerca.*

En este caso concreto, el usuario sólo conoce la estación de origen, pero para solicitar la información debe determinar la estación de destino y la hora de llegada.

Por último, en los escenarios abiertos el usuario desconoce la mayoría de los detalles de la consulta que ha de realizar. El siguiente escenario es un ejemplo de este tipo:

> *Vienen unos amigos a visitarle el fin de semana. Quiere llevarlos de excursión a Montserrat y para saber toda la información que necesita llama al servicio de información de ferrocarriles.*

3. Diseñar las estrategias para gestionar el diálogo.

El corpus persona-persona permite analizar la actuación de los dos participantes en la tarea. Así, si el objetivo del sistema es atender demandas de información, se estudian dos comportamientos: el de quien pide la información y el de quien se la facilita. Para

gestionar el diálogo es muy importante conocer los datos que necesitan los operadores antes de responder a la consulta. Por ejemplo, si se observa el comportamiento de un operador humano de un servicio de información de ferrocarriles cuando proporciona información de horarios, se pone de manifiesto que solicita los siguientes datos: la estación de origen, la estación de destino, el día que el usuario desea viajar y la franja horaria.

Por tanto, el corpus persona-persona sirve como punto de partida para diseñar las estrategias de diálogo y para saber cómo orientar al usuario en el momento de emplear el sistema. También se utiliza para generar las posibles respuestas ante una intervención concreta.

Para facilitar el análisis del corpus persona-persona y obtener toda la información mencionada -tipos de consulta, extracción de escenarios, estrategias de diálogo- se lleva a cabo un proceso previo de transcripción y anotación del corpus. La anotación será más o menos exhaustiva en función del modelo que se quiera conseguir para desarrollar el sistema de diálogo y de la complejidad de la aplicación.

4.3. Estudio de corpus de interacciones ficticias entre personas y sistemas informáticos

Las interacciones ficticias sirven para situar a las personas que participan en la evaluación de un prototipo como potenciales usuarios en un acto comunicativo semejante al de la aplicación real del sistema de diálogo. Estas interacciones se recogen en un corpus mediante el protocolo denominado “Mago de Oz”, en el que un operador humano -el “mago”- sustituye al sistema, de modo que los usuarios suponen que el intercambio se establece con una aplicación ya desarrollada (Dahbälck *et al.*, 1998). Naturalmente, la persona que actúa de mago necesita una etapa previa de entrenamiento y un conocimiento muy detallado de los escenarios que se hayan elaborado. En este sentido, los escenarios abiertos son los más difíciles de abordar, ya que el usuario puede preguntar cualquier detalle sobre la información que quiere solicitar. Para facilitar la tarea del mago es aconsejable incluir los módulos del sistema que ya están preparados. Por ejemplo, si el módulo de reconocimiento o el de interpretación semántica ya están en funcionamiento, el mago no necesitará simular errores de reconocimiento y sabrá qué estrategia debe seguir según el comportamiento del usuario (Bonafonte *et al.*, 2000).

El método del Mago de Oz se suele emplear en la creación y evaluación de prototipos con los siguientes objetivos:

1. Probar la eficacia del sistema. La evaluación de las interacciones entre el usuario y el prototipo es el punto de partida para modificar, si es necesario, las estrategias de diálogo, mejorando así el funcionamiento del sistema.
2. Analizar el comportamiento de las personas en la interacción con una máquina para extraer la información lingüística que necesitará el sistema real cuando haya de interpretar los mensajes del usuario. En principio, la riqueza del léxico y de las estructuras lingüísticas que se observan en las conversaciones naturales no es la misma que cuando la comunicación se lleva a cabo con un ordenador. Por tal motivo, se debe establecer una comparación entre el corpus persona-persona

y el corpus persona-sistema informático teniendo en cuenta, entre otros factores y en función de la aplicación, la duración total de la interacción, el tiempo de espera, el número de turnos utilizados desde que el usuario solicita la información hasta que se le facilita y el número de llamadas que se desvían porque el mago no es capaz de contestarlas.

3. Entrenar el módulo de reconocimiento, ya que en la mayoría de las aplicaciones se requiere el reconocimiento del habla espontánea. El reconocedor puede mejorar si se incorpora información sobre ciertos fenómenos lingüísticos propios de este estilo de habla: repeticiones, inserciones que truncan el discurso, relajaciones en la articulación de los sonidos, alargamientos de vocales, ruidos del usuario, vocalizaciones, etc.

En la obtención de este tipo de diálogo se deben considerar la edad y el sexo del usuario para descubrir las dificultades que pueda tener, en función de esas variables, cuando interactúa con el sistema.

Finalmente, se suele incluir un cuestionario para evaluar el prototipo una vez efectuado el intercambio comunicativo. En estos cuestionarios el usuario debe valorar, entre otras cuestiones, si le ha gustado el sistema, si lo ha considerado ágil y si le ha proporcionado la información solicitada. A partir de las respuestas sobre el grado de satisfacción del usuario y de la intervención grabada del mismo, se analizan una serie de rasgos -el tiempo que se ha utilizado en el escenario, la adecuación de la respuesta, el número de turnos empleados- que permiten evaluar cuantitativa y cualitativamente el sistema (Colás, 2001).

4.4. Diseño de estrategias de diálogo

En un sistema automático de diálogo, las estrategias se diseñan en función de los objetivos que deben conseguirse. En primer lugar, es preciso considerar la aplicación concreta en la que se empleará, puesto que en unos casos el usuario deseará obtener información, mientras que en otros querrá que se ejecute una orden. En segundo lugar, es necesario planificar el tipo de diálogo que se espera que el usuario mantenga con el sistema. No obstante, aunque en la actualidad las estrategias de diálogo empleadas responden, por lo general, a los requisitos de una determinada aplicación, un tratamiento más profundo de la pragmática de las distintas situaciones de habla permitiría concebir sistemas más flexibles (Ludwig, 2001).

En cuanto a las tareas, aunque los límites no siempre son nítidos, pueden perseguirse cuatro objetivos distintos (Minker y Bennacef, 2001: 95):

1. Aprendizaje. Puede darse tanto por parte del usuario -como en el caso de la enseñanza asistida por ordenador o en el entrenamiento de controladores aéreos- como por parte del sistema, que debe adquirir conocimientos a partir de la interacción con el usuario.
2. Información. El objetivo es atender las demandas de información por parte del usuario, por ejemplo, las relativas a horarios, destinos, itinerarios, etc., de los medios de transporte públicos.

3. Orden. El usuario precisa manejar objetos en un entorno específico; tal es el caso del manejo de robots o de herramientas para el diseño gráfico.
4. Asistencia. La interacción proporciona al usuario orientaciones que le ayudan a tomar decisiones, como sucede en la diagnosis médica.

El desarrollo de cada tarea se planifica y se descompone en subtareas de forma jerárquica. Cada nuevo intercambio entre el sistema y el usuario contribuye a seleccionar un itinerario en la cadena de estrategias de diálogo previstas. Con cada intervención, se descarta una serie de posibles réplicas por parte del sistema y se selecciona la respuesta más adecuada de entre las previstas. El intercambio con el usuario irá determinando de forma sucesiva las respuestas escogidas, hasta que se produzca el cierre de la comunicación.

Como consecuencia, en un sistema de diálogo es preciso planificar tanto las acciones que deben seguirse para obtener, corregir o confirmar la información que ha de conducir a la realización de las tareas como aquellas estrategias de diálogo que prevean e identifiquen las intenciones del usuario. Es conveniente emplear formas lingüísticas que permitan distinguir claramente un cometido del otro. Cada intervención debe ser clara y sencilla, y el sistema debe respetar un determinado lapso de tiempo para cada una de ellas. En este sentido, una instrucción como *Por favor, diga el nombre de la empresa y hable después de la señal* puede confundir al usuario porque no está secuenciada correctamente. *Por favor, hable después de la señal. Diga el nombre de la empresa* es una propuesta más adecuada.

El diseño de estrategias requiere tener presentes los siguientes objetivos: (a) conseguir que el usuario tenga la certeza de que la aplicación funciona y procesa correctamente, (b) asegurar que le sea posible rectificar y enmendar errores, ya sean propios o de la aplicación, (c) evitar los malentendidos originados por problemas de reconocimiento o descodificación no corregidos por él, y (d) poder recibir en un lenguaje claro y directo las respuestas (Lamel *et al*., 1998: 210).

A continuación se presentan los principales tipos de interacción que se pueden dar en un sistema de diálogo y las estrategias correspondientes. Las interacciones que permiten la puesta en funcionamiento de cualquier aplicación están destinadas a iniciar y cerrar la comunicación, y a completar el funcionamiento del reconocedor y del decodificador (véase el apartado 4.1) mediante demandas de confirmación y rectificación de errores. Para cada aplicación concreta se desarrollan, además, estrategias específicas.

4.4.1. Estrategias de acceso

El sistema debe dar la bienvenida de forma concisa, y de manera que el usuario pueda comprender de forma inequívoca que está a punto de iniciar un diálogo con una máquina. Además, suele indicar el servicio que proporciona, como sucede en los siguientes ejemplos:

Le systeme MASK vous écoute

(MASK, `http://www.limsi.fr/Recherche/TLP/mask.html`)

> *When you call, you will be connected with Jupiter (the connection will take about 5 sec.), and the system will greet you with something like* "*Welcome to Jupiter - the automated weather service from MIT. How may I help you?*"
>
> (JUPITER, http://www.sls.lcs.mit.edu/sls/whatwedo/applications/jupiter.html)

> *Benvingut al servei de Ferrocarrils de la Generalitat, bon dia.*
> *Benvingut al servei de Ferrocarrils de la Generalitat, bona tarda.*
> *Benvingut al servei de Ferrocarrils de la Generalitat, bona nit.*
>
> (Machuca *et al*., 2000)

Tras el saludo, es posible que el usuario realice ya una consulta, así que, si después de realizar una búsqueda ordenada, el sistema detecta una palabra clave de las que tiene almacenadas para la aplicación, se pondrá en funcionamiento una estrategia de diálogo. En caso contrario, se iniciará una demanda de información con alguna pregunta (*¿Qué tipo de consulta desea realizar?*) hasta obtener una palabra clave que permita lanzar alguna estrategia de diálogo.

4.4.2. Estrategias de salida

Antes de terminar completamente la interacción el sistema debe ofrecer al usuario la posibilidad de continuar el diálogo solicitando un nuevo servicio (*¿Quiere realizar otra consulta?)*. Si la respuesta es positiva, la comunicación se reanuda, ya que se pone en funcionamiento la cadena de diálogos correspondiente, pero si es negativa, se finaliza la comunicación con alguna fórmula cortés de cierre (*Gracias por utilizar este servicio*).

4.4.3 Estrategias de confirmación

Las estrategias de confirmación tienen como objetivo comprobar lo que el usuario desea, no necesariamente lo que ha dicho. Según el número de datos que se pretenden confirmar (San-Segundo *et al*., 2001), las estrategias pueden centrarse en un único dato (*¿Desea ir a Barcelona?)* o en varios *(¿Desea ir de Madrid a Barcelona?*)**.** Según la facilidad de corrección (Lavelle, 1999; San-Segundo *et al*., 2001), pueden distinguirse los siguientes procedimientos:

1. Confirmación explícita: el sistema utiliza una pregunta directa para comprobar que ha procesado correctamente (*Entiendo que desea saber el tiempo que hace en Barcelona. ¿Es correcto?*).
2. Confirmación totalmente implícita: se informa parcialmente del resultado de la descodificación, es decir, de lo que ha entendido el sistema, pero no se permite realizar la corrección (*Desea llegar a las seis de la tarde a Valencia. ¿Cuál es la ciudad de origen?*).
3. Confirmación parcialmente implícita: se informa del resultado de la descodificación y se permite realizar la corrección (*Entiendo que desea ir a Barcelona, diga "corregir" si no es correcto, o diga la hora de llegada*.).

4. Confirmación ausente: el sistema no informa del resultado de la descodificación. Generalmente se utiliza cuando el grado de confianza en el reconocimiento es alto (por ejemplo, en preguntas que se responden con un *sí* o un *no*).
5. Repetición de la pregunta: el sistema estima que el grado de confianza en lo reconocido es muy bajo, así que lo desestima y pregunta nuevamente (*Perdone, no le he entendido muy bien. Repita la ciudad de origen*).

Cuando se sigue el método de la confirmación implícita debe tenerse en cuenta que el sistema puede haber realizado suposiciones equivocadas, así que se ha de prever un mecanismo que permita recuperar y corregir los errores. Esto puede conseguirse, por ejemplo, dotando al sistema de la capacidad de reaccionar adecuadamente ante órdenes como "corregir" o "volver a empezar" e informando al usuario de esta posibilidad. Dado que la producción de errores altera el funcionamiento, es conveniente diseñar estrategias que permitan evitar el cierre brusco de la comunicación y reconducir apropiadamente el diálogo. Una interfaz bien diseñada puede minimizar los efectos producidos por los errores de reconocimiento, mientras que una aplicación con un buen reconocedor puede fracasar si la interfaz no está bien planificada. Los errores pueden producirse no sólo por un fallo en el reconocedor acústico o en la descodificación, sino como consecuencia de la superposición de voces o porque el usuario proporciona una nueva información mientras que se está procesando la anterior. El sistema debe detectar el tipo de error producido y procurar subsanarlo. Para ello se emplean mensajes del tipo *Mientras buscaba la respuesta me ha pedido más información. ¿Podría repetir la pregunta, por favor?* Si no es capaz de corregirlo después de varios intentos, es aconsejable desviar la consulta hacia un operador humano.

Los sistemas de diálogo cerrados se basan en el uso de pares del tipo pregunta-respuesta, de manera que impiden la iniciativa del usuario, pero en los sistemas cooperativos existe verdadera interacción, ya que se aceptan interrupciones y negociaciones, especialmente en el caso de los sistemas adaptativos, diseñados para responder en función del nivel de destreza del usuario (Veldhuijzen van Zanten, 1999). La comunicación es más rápida y efectiva si se proporciona el grado adecuado de detalle explicativo sobre el procedimiento de respuesta; por ejemplo, un usuario avezado no precisa instrucciones del tipo *Hable después de escuchar la señal*, mientras que para otro puede ser imprescindible. Un mismo usuario, además, adquiere experiencia a medida que se desarrolla el diálogo, algo que puede comprobarse examinando la evolución en el número de errores o de demandas de confirmación. Si el sistema es lo suficientemente versátil, puede adecuarse en sus intervenciones a las del usuario, ya sea relajando el modo de realizar las preguntas, ya sea ampliando el nivel de detalle de las explicaciones (*Hable después de oír la señal. Diga la franja horaria en la que desea viajar; por la mañana, a primera hora de la mañana, al mediodía, por la tarde, a primera hora de la tarde o por la noche*). Para perfeccionar este tipo de interacciones se puede emplear, en una fase previa a la presentación del producto, la técnica del Mago de Oz (véase el apartado 4.3). Mediante la simulación se consigue tanto comprobar el diseño de las estrategias de diálogo previamente establecidas como estudiar el comportamiento de los usuarios en una situación similar a la real.

4.5. Los problemas de los sistemas de diálogo

Tal como se ha expuesto, en un sistema de diálogo se suceden fases de reconocimiento de habla, de procesamiento lingüístico y de conversión de texto en habla. Los posibles errores proceden, por tanto, de las dificultades que puedan surgir en estas fases.

Durante la fase del reconocimiento de habla el sistema puede incurrir en errores de distinto origen:

1. Las características del locutor: lengua materna, edad y sexo, ya que, evidentemente, estas características inciden en la variedad lingüística del usuario. Un hablante no nativo, por ejemplo, puede encontrar dificultades en la comprensión de los mensajes y expresarse de forma que resulte difícil para un reconocedor entrenado con hablantes nativos.
2. La variedad lingüística del usuario (diatópica, diafásica, diastrática o idiolectal) puede ser diferente a la variedad con la que se ha entrenado el sistema. Un sistema preparado para identificar la palabra *tomàquet* ("tomate") del catalán podría tener problemas ante variantes como *tomaca, tomata, tomàtiga.*
3. Las variedades en la pronunciación de préstamos. Por ejemplo, un servicio automático de encargos de comida debería reconocer las diferentes maneras de pronunciar *pizza* en español.
4. El ruido ambiental y el canal de transmisión (red fija, móvil, Internet), así como la codificación que se emplea (Golderos, 2001), ya que la modificación de las propiedades acústicas de la señal dificulta el reconocimiento del habla. Así, por ejemplo, no es lo mismo intentar establecer un diálogo con una máquina desde un despacho sentado delante de un micrófono conectado al ordenador que desde el coche un día de tráfico intenso, con condiciones atmosféricas desfavorables, y desde un teléfono *manos libres*.

Después del reconocimiento del habla, durante la fase del procesamiento lingüístico, muchos de los problemas dependen del grado de perfección logrado por el tipo de sistema de diálogo que se utilice: dependiente o independiente del hablante, basados en el habla continua o discreta, en la asignación de la gestión de los turnos de palabra, etc. (Guinn y Montoya, 1998; Waibel, 2001).

Así, por ejemplo, un sistema de diálogo que pretenda reconocer habla espontánea se enfrentará a problemas en la interpretación de elipsis y de pronombres (*Sí, sí lo quiero, el seguro*), correcciones por parte del hablante (*No, no quiero los horarios quiero…*), uso de neologismos y acepciones nuevas (por ejemplo, un usuario que solicite información sobre lugares en los que pueda *chatear*).

Si, además, se permite que sea el usuario quien gestione los turnos de palabra, el sistema tendrá que tener en cuenta la relación entre la expresión lingüística y su valor ilocutivo. Compárense, por ejemplo, las distintas formas que pueden utilizarse para realizar el acto de solicitud o petición de información: *¿Cuál es…?, ¿Podría decirme…?, ¿Puede decirme…?, Quiero saber…, Quisiera saber…, Querría saber...* La identificación de las estrategias que los usuarios emplean para conservar el turno de

palabra en una conversación ayudaría, además, a evitar solapamientos con las respuestas del sistema (*¿Adónde quiere ir? A Madrid, eeeeh…).*

La adecuación de las respuestas depende en buena medida de factores pragmáticos tales como la estructura y el tipo de aporte informativo. En un diálogo, el conocimiento que el usuario considera que comparte con su interlocutor procede, en primer lugar, de aspectos como la concepción del mundo, la percepción humana del espacio y el tiempo, o la tradición histórica y cultural de la comunidad en la que tiene lugar la comunicación; en segundo lugar, del entorno situacional concreto del momento del intercambio y, en tercer lugar, de todas las expresiones lingüísticas aparecidas anteriormente en el discurso. Así, por ejemplo, si a un robot de limpieza se le pide que coloque un cenicero *más arriba* existe un margen fuera del cual es improbable que deba colocarlo, y si un usuario ha expresado que detesta *viajar encogido*, el sistema no deberá preguntarle si prefiere un asiento junto al pasillo o junto a la ventana. Tener en cuenta aspectos como la inferencia pragmática ayuda a detectar posibles inconsistencias en el diálogo y a resolver problemas de ambigüedad.

En la actualidad se trabaja en cuestiones como la interpretación de expresiones vagas (*el final del pasillo*) o metonímicas (como *Ve a la puerta*, en la que se da valor locativo a un objeto), que podrían provocar errores de comprensión (Bos *et al*., 2003). También se pretende dotar a los sistemas de mayor naturalidad con la incorporación de construcciones sintácticas poco empleadas hasta ahora porque su uso depende del análisis pragmático del texto. Por ejemplo, las respuestas condicionales (Kruijff-Korbayová *et al*., 2002), además de poder constituir respuestas a preguntas formuladas mediante oraciones interrogativas totales, pueden emplearse para contestar una pregunta efectuada con una interrogativa parcial (*¿A qué hora llegaré? A las cuatro, si toma el rápido*) o para realizar confirmaciones:

> *-Hay un tren que sale de Barcelona el lunes.*
> *-Entonces llegaré el martes.*
> *-(a) No si sale/a menos que salga/a no ser que salga por la mañana.*
> *(b) Sí, si sale/a menos que salga/a no ser que salga por la noche.*

Finalmente, en la etapa de conversión de texto en habla, suponiendo que el sistema haya interpretado adecuadamente el mensaje del usuario, los errores se pueden deber a la baja calidad del habla sintetizada, tanto en los aspectos segmentales como en la asignación de la prosodia.

4.6. Adecuación y corrección lingüística del sistema

Un sistema de diálogo de calidad, además de no incurrir en errores de reconocimiento y descodificación, debería estar diseñado para realizar una producción cuidada, con enunciados gramaticalmente correctos y pragmáticamente adecuados.

En cuanto al léxico (Alcoba, 1999), no es aceptable, por ejemplo, que en un sistema de diálogo se incurra en impropiedades semánticas, como la que aparece en el enunciado *La climatología será buena en las próximas horas en Barcelona*: *climatología* significa "tratado del clima" o "conjunto de las condiciones propias de un determinado clima" y no "tiempo" (DRAE, 2001).

Tampoco deben utilizarse préstamos innecesarios: *La presentación será en el foyer (vestíbulo) del teatro a las 18 h*., ni comodines léxicos: *Para hacer (presentar) la reclamación debe aportar las siguientes cosas (la siguiente documentación*), y hay que evitar las redundancias: *Este sistema le permite cambiar divisas extranjeras.* Asimismo, no es aceptable mezclar variedades diatópicas (*A través de este sistema puede alquilar coches. Tenemos carros de tres y de cinco puertas*) o utilizar una variedad diastrática o diafásica marcada (*¿Desea estudiar en una Uni (universidad) pública o privada?).*

En la construcción de las oraciones deben respetarse las propiedades sintácticas -*Para recurrir (contra) la sentencia*- y morfológicas -*Los accésits (accésit) han sido otorgados a...*- de las palabras. Además, siempre es conveniente utilizar las construcciones más naturales de la lengua: en español, un enunciado como *La solicitud debe presentarse antes de las 15 h*. es preferible a *La solicitud debe ser presentada antes de las 15 h*, donde aparece una pasiva perifrástica.

La pronunciación debe ser cuidada y atenerse a la norma de la variedad escogida. Así, por ejemplo, no sería adecuado oralizar en español estándar peninsular el enunciado *Si usted ha nacido...* como [sɪ̯u^{h}ˈtɛanaˈsio], aunque es una pronunciación perfectamente adecuada a la norma andaluza.

Para conseguir la adecuación pragmática del diálogo, el sistema también deberá respetar las normas de cortesía de la comunidad a la que se dirige (Brown y Levinson, 1987) y adoptar las formas lingüísticas pertinentes (trato de *tú* o de *usted*, por ejemplo). Deben evitarse las faltas de respeto: *Le he dicho que me diga el destino (Por favor, vuelva a decirme el destino*) y el uso de palabras que constituyen un tabú para una determinada comunidad; por ejemplo, un enunciado como *Puede coger el tren de las tres* sería muy ofensivo para los hablantes de algunas variedades de español.

Además de todo lo esbozado aquí, la calidad de un sistema de diálogo también depende del grado de satisfacción de los usuarios. En general, se prefieren los sistemas en los que el habla se asemeje a la natural y en los que la interacción no implique demasiado esfuerzo; por ejemplo, si un usuario considera que la lengua empleada es artificial y desagradable o le exige un gran esfuerzo (el aprendizaje de palabras claves, el deletreo, la inserción de pausas que no son habituales en el habla espontánea, el excesivo número de turnos de palabra, etc.) acabará abandonando este medio y se inclinará por una estrategia alternativa, por ejemplo, esperar que un operador humano atienda el teléfono.

5. Conclusiones

En las páginas anteriores se ha intentado poner de manifiesto el papel de la lingüística en el desarrollo de las tecnologías del habla. La visión que se desprende es que nos encontramos ante un campo caracterizado por un elevado grado de interdisciplinariedad, que requiere la estrecha colaboración de especialistas procedentes de diversos ámbitos. El lingüista puede aportar sus conocimientos sobre cada uno los niveles de descripción de la lengua -del fonético al pragmático-, tanto en el proceso de desarrollo como en el de evaluación de las aplicaciones de la síntesis, del reconocimiento y del diálogo. Al mismo tiempo, la participación en estas tareas potencia el interés por nuevos estudios, pues en muchos casos no se dispone de la información que sería necesaria -especialmente en el caso de lenguas como el castellano o el catalán- o se cuenta con una

cantidad insuficiente de datos. Así se establece una relación circular entre la investigación básica y la aplicada, que redunda en beneficio de ambas, y se abren también nuevos horizontes profesionales al lingüista que decide orientar su actividad al mundo de las tecnologías.

6. Referencias[1]

Abney, S. (1992), "Prosodic Structure, Performance Structure and Phrase Structure", *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, Harriman, NY, February, 1992, San Mateo, CA, Morgan Kaufmann, pp. 425-428.
http://www.vinartus.net/spa/92d.pdf

Actor, Loquendo TTS, Loquendo Vocal Technologies and Services, Torino.
http://www.loquendo.com/es/products/TTS.htm

Aguilar, L. - Garrido, J.M. - Llisterri, J. (1997), "Incorporación de conocimientos fonéticos a las tecnologías del habla", en Serra, E. - Gallardo, B. - Veyrat, M. - Jorques, D. - Alcina, A. (Eds.), *Panorama de la investigació lingüística a l'Estat Espanyol, Actes del I Congrés de Lingüística General, Volum III, Comunicacions: Fonètica i Fonologia. Semàntica i Pragmática*, Valencia, Universitat de Valencia, pp. 5-13.
http://liceu.uab.es/~joaquim/publicacions/valencia_94.html

Alcoba, S. (1999), "El léxico: condiciones de uso", en Alcoba, S. (Ed.), *La oralización*, Barcelona, Ariel, pp. 63-107.

Allen, J. - Hunnicutt, M.S. - Klatt, D.H. (with R.C. Armstrong and D. Pisoni) (1987), *From Text to Speech: The MITalk System*, Cambridge, Cambridge University Press.

Bartkova, K. - Haffner, P. - Larreur, D. (1993), "Intensity Prediction for Speech Synthesis in French", en House, D.- Touati, P. (Eds.), *Proceedings of an ESCA Workshop on Prosody,* September 27-29, 1993, Lund, Sweden, Lund University Department of Linguistics and Phonetics, Working Papers 41, pp. 280-283.

Batliner, A. - Möbius, B. - Möhler, G. - Schweitzer, A. - Nöth, E. (2001), "Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground", *Eurospeech'01. Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, 3-7 September, 2001, Vol. 4, pp. 2285-2288.
http://www.smartkom.org/prosodic_models.pdf

Beaugendre, F. (1996), "Modèles de l'intonation pour la synthèse de la parole", en Méloni, H. (Coord.), *Fondements et Perspectives en Traitement Automatique de la Parole*, Paris, Éditions AUPELF-UREF, pp. 97-198.
http://www.bibliotheque.refer.org/parole/beaugend/beaugend.htm

[1] La validez de las direcciones en Internet citadas en la bibliografía se ha verificado en marzo de 2003.

Bernstein, J. - Franco, H. (1996), "Speech recognition by computer", en Lass, N.J (Ed.), *Principles of Experimental Phonetics*, St Louis, Mosby, pp. 408-434.

Bonafonte, A. - Aibar, P.- Castell, N. - Lleida, E. - Mariño, J.B. - Sanchís, E. - Torres M.I. (2000), "Desarrollo de un sistema de diálogo oral en dominios restringidos", *I Jornadas en Tecnologías del Habla*, Sevilla, 6-10 de noviembre de 2000.
http://gps-tsc.upc.es/veu/basurde/download/Bon00a_sevilla.pdf

Borzone, A.M. - Signorini, A. (1983), "Segmental duration and rythm in Spanish", *Journal of Phonetics*, 11, pp. 117-128.

Bos, J.- Klein, E.- Oka, T. (2003), "Meaningful Conversation with a Mobile Robot", *Proceedings of EACL03, Tenth Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, April 12-17, 2003.
http://www.cogsci.ed.ac.uk/~jbos/pubs/bko-eacl03.pdf

Brown, P.- Levinson, S. (1987), *Politeness. Some Universals in Language Use*, Cambridge, Cambridge University Press.

Caballero, M. - Moreno, A. (2001), "Reconocimiento automático del habla multidialectal", *Actas del XVI Simposium de la Unión Científica Internacional de Radio*, *URSI'2001,* Madrid.
http://gps-tsc.upc.es/veu/teham/PresentacionURSI.pdf

Campbell, W.N. (1992), "Syllable-based segmental duration", en Bailly, G. - Benoît, C. - Sawallis, T.R. (Eds.), *Talking Machines: Theories, Models and Designs*, Amsterdam, North-Holland, pp. 211-224.

Canellada, M. J. - Madsen, J. K. (1987), *Pronunciación del Español. Lengua hablada y literaria*, Madrid, Castalia.

Castejón, F.- Escalada, G.- Monzón, L.- Rodríguez, M.A.- Sanz, P. (1994), "Un conversor texto-voz para el español", *Comunicaciones de Telefónica I+D*, 5, 2, pp. 114-131.
http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol52/artic8/8.html

Colás, J. (2001), *Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en español*, Estudios de Lingüística Española, 12.
http://elies.rediris.es/elies12/index.html

Dahbälck, N. - Jönsson, A. - Ahrenberg, L. (1998), "Wizard of Oz studies-Why and How", en Maybury, M. - W. Wahlster (Eds.), *Readings in Intelligent User Interfaces*, San Mateo, CA, Morgan Kaufman, pp. 610-119.

Danieli, M. - Gerbino, E. (1995), "Metrics for evaluating dialogue strategies in a spoken language system", *Proceedings of the1995 AAAI Spring Symposium on Empirical Methods in Discourse Interpretation and Generation*, Stanford University, California, March 27-19, pp. 34-39.
http://arxiv.org/PS_cache/cmp-lg/pdf/9612/9612003.pdf

Deroo, O. (1999), *A Short Introduction to Speech Recognition*, TCTS Lab, Faculté Polytechnique de Mons.
http://tcts.fpms.ac.be/asr/introduction.html

DRAE (2001): Real Academia Española (2001), *Diccionario de la lengua española*, vigésima segunda edición, Madrid, Espasa Calpe.

Dutoit, T. (1997), *An Introduction to Text-to-Speech Synthesis*, Dordrecht, Kluwer Academic Publishers.

Dutoit, T.- Stylianou, Y. (2003), "Text-to-speech synthesis", en Mitkov, R. (Ed.), *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press.

Enríquez, E.V. (1991), "El problema de las ambigüedades fonéticas y su tratamiento automático", *Boletín de la Real Academia Española*, Tomo LXXI, Cuaderno CCLII, pp. 157-183.

Finch, D. F. - Ortiz, H. (1982)*, A Course in English Phonetics for Spanish Speakers.,* London, Heinemann.

Fosler-Lussier, E. - Greenberg, S. - Morgan, N. (1999), "Incorporating contextual phonetics into automatic speech recognition", *Proceedings of the 14th International Congress of Phonetic Sciences*. San Francisco, 1-7 August 1999.
http://www.icsi.berkeley.edu/~fosler/papers/ICPhS99-invited.pdf

Frankel, J.- King, S. (2001, "Articulatory speech recognition", *Eurospeech'01, Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg, Denmark, 3-7 September, 2001, pp. 599-602.
http://archive.ling.ed.ac.uk/documents/disk0/00/00/01/53/taal00000153-01/paper.pdf

Giachin, E. (1997), "Spoken Language Dialogue", en Cole, R.A.- Mariani, J.- Uszkoreit, H.- Zaenen, A.- Zue, V. (Eds.), *Survey of the State of the Art in Human Language Technology*, Cambridge, Cambridge University Press, pp. 241-244.
http://cslu.cse.ogi.edu/HLTsurvey/ch6node6.html

Gibbon, D.- Mertins, I.- Moore, R. (Eds.) (2000), *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation*, Dordrecht, Kluwer Academic Publishers.

Golderos, F. (2001), "Tecnologías del habla en español: convergencia con Internet", *I Congreso Internacional de la Lengua Española. El español en la Sociedad de la Información*, Valladolid, 16-19 de octubre de 2001.
http://cvc.cervantes.es/obref/congresos/valladolid/ponencias/el_espanol_en_la_sociedad/4_internet_en_espanol/golderos_f.htm

Gómez, J. (2000), "Perspectivas de la lingüística computacional", *Novática. Revista de la Asociación de Técnicos de Informática*, 145, pp. 85-87.
http://www.ati.es/novatica/2000/145/javgom-145.pdf

Guilder, L. van (1995), *Automated Part of Speech Tagging: a brief overview*, Handout for LING361, Fall 1995, Georgetown University.
http://www.georgetown.edu/faculty/ballc/ling361/tagging_overview.html

Guinn, C.I. - Montoya, J.R. (1998), "Natural language Processing in Virtual Reality Training Environments", *Modern Simulation and Training*, June, pp. 44-55.
http://www.cs.duke.edu/~cig/papers/iitsec_guinn.htm

Gussenhoven, C. (2002), "Intonation and Interpretation: Phonetics and Phonology", *Proceedings of Speech Prosody 2000, An International Conference*, Aix-en-Provence, France, 11-13 April 2002.
http://www.lpl.univ-aix.fr/sp2002/pdf/gussenhoven.pdf

Hirschberg, J. (2002), "The Pragmatics of Intonational Meaning", *Proceedings of Speech Prosody 2000, An International Conference*, Aix-en-Provence, France, 11-13 April 2002.
http://www.lpl.univ-aix.fr/sp2002/pdf/hirschberg.pdf

Hirschman, L. - Bates, M. - Dahl, D. - Fisher, W. - Garofolo, J. - Pallett, D. - Hunicke-Smith, K. - Price, P. - Rudnicky, A. - Tzoukermann, E. (1993), "Multi-site data collection and evaluation in spoken language understanding", *Proceedings of the ARPA Human Language Technology Workshop*, March, 1993, San Mateo, CA, Morgan Kaufmann, pp.19-24.
http://citeseer.nj.nec.com/hirschman93multisite.html

Hunt, A. - Black, A. (1996), "Unit selection in a concatenative speech synthesis system using a large speech database", *Proceedings of ICASSP 96*, Atlanta, Georgia, 1996, Vol 1, pp. 373-376.
http://www.cstr.ed.ac.uk/publications/papers/1996/Hunt_1996_a.ps

Ingeniería Lingüística. Cómo aprovechar la fuerza del lenguaje, Luxemburg, Anite Systems.
http://www.hltcentral.org/usr_docs/Harness/harness-es.htm

Jones, D. (1918), *An outline of English Phonetic*s, Cambridge, Cambridge University Press, 1972.

JUPITER, Spoken Language Systems Group, MIT Laboratory for Computer Science, Massachusetts Institute of Technology.
http://www.sls.lcs.mit.edu/sls/whatwedo/applications/jupiter.html

Kirchhoff, K.- Fink, G.A.- Sagerer, G. (2002), "Combining acoustic and articulatory feature information for robust speech recognition", *Speech Communication,* 37, 3-4, pp. 303-320.

Klatt, D. H. (1979), "Synthesis by rule of segmental durations in English sentences", en Lindblom, B. - Öhman, S. (Eds.), *Frontiers of Speech Communication Research*, New York, Academic Press, pp. 287-299.

Klatt, D. H. (1987), "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America* 82, 3, pp. 737-793; en Atal, B.S. - Miller, L.J. - Kent, R.D. (Eds.) (1991), *Papers in Speech Communication: Speech Processing*, New York, Acoustical Society of America, pp. 57-114.
http://www.mindspring.com/~ssshp/ssshp_cd/dk_737a.htm

Koreman, J.- Andreeva, B. (2000), "Can we use the linguistic information in the signal?", *Phonus* (Institute of Phonetics, University of the Saarland) 5, pp. 47-58.
http://www.coli.uni-sb.de/Phonetics/Research/PHONUS_research_reports/Phonus5/Koreman_PHONUS5.pdf

Koreman, J.- Andreeva, B.- Strik, H. (1999), "Acoustic parameters versus phonetic features in ASR", *Proceedings of the 14th International Congress of Phonetic Sciences*, 1-7 August 1999, pp. 719-722.
http://lands.let.kun.nl/literature/strik.1999.2.ps

Kruijff-Korbayová, I.- Karagjosova, E. –Larsson. S. (2002) "Enhancing collaboration with conditional responsesin information-seeking dialogues", *Proceedings of EDILOG 2002*, *6th Workshop on the Semantics and Pragmatics of Dialogue*, The University of Edinburgh, September 4-6, 2002. pp. 93-100.
http://www.ltg.ed.ac.uk/edilog/papers/093.pdf

Kurzweil, R. (1998), "When Will HAL Understand What We Are Saying? Computer Speech Recognition and Understanding", en Stork, D.G. (Ed.), *Hal's Legacy: 2001's Computer as Dream and Reality,* Cambridge, Mass., The MIT Press.
http://mitpress.mit.edu/e-books/Hal/chap7/seven1.html

Lamel, L.F. - Gauvain, J.L. (2003), "Speech recognition", en Mitkov, R. (Ed.), *The Oxford Handbook of Computational Linguistics*, Oxford, Oxford University Press.

Lamel, L.F. - Rosset, S. - Gauvain, J.L. - Bennacef, S. - Garnier-Rizet, M. - Prouts, B. (1998), "The LIMSI ARISE system", *IVITTA'98, Proceedings of the 4th IEEE Workshop on Interactive Voice Technology for Telecommunications Applications*, Torino, Italy, September 1998, pp. 209-214.
ftp://tlp.limsi.fr/public/ivtta98.ps.Z

Lamel, L.F. (1993), "A knowledge-based system for stop consonant identification based on speech spectrogram reading", *Computer Speech and Language*, 7,2, pp. 169-191.

Laporte, E. (1988), *Méthodes algoritmiques et lexicales de phonetisation de textes*, Thèse doctorale, Centre d'Études et de Recherches en Informatique Linguistique, Université Paris 7, Paris.

Lavelle, A.C. - Calmes, H. - Pérennou, G. (1999), "Confirmation strategies to improve correction rates in a telephonic inquiry dialogue system", *Eurospeech'99. 6th European Conference on Speech Communication and Technology,* Budapest, September 5-9, 1999, Vol. 3, pp. 1399-1402.

Lehiste, I. (1995), "Suprasegmentals features of speech", en Lass, N. (Ed.), *Principles of Experimental Phonetics*, Mosby, Sant Louis, pp. 226-245.

Liberman, M. - Church, K. (1992), "Text analysis and word pronunciation in text-to speech synthesis", en Furui, S. - Sondhi, M.M. (Eds.), *Advances in Speech Signal Processing,* New York, Marcel Dekker, pp. 791-831.

López, M. - Martínez, G. (1988), *La función de las pausas en la lengua oral*, Manuscrito no publicado, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona.

López-Cózar, R. - Rubio, A.J. - García, P. - Díaz-Verdejo, J. E. (2000), "Test de estrategias de diálogo en un sistema conversacional", *I Jornadas en Tecnologías del Habla*, Sevilla, 6-10 de noviembre de 2000.

Ludwig, B. (2001) "Dialogue Understanding in Dynamic Domains", *Bi-dialog 2001, Proceedings of the 5th Workshop on Formal Semantics and Pragmatics of Dialogue*, University of Bielefeld, Bielefeld, June, 14 - 16, 2001, pp. 287-297.
http://www.uni-bielefeld.de/BIDIALOG/proc.pdf.gz

Llisterri, J. - Aguilar, L. - Garrido, J.M. - Machuca, M.J. - Marín, R. – de la Mota, C. - Ríos, A. (1999), "Fonética y tecnologías del habla", en Blecua, J.M. - Clavería, G. - Sánchez, C. - Torruella, J. (Eds.), *Filología e informática. Nuevas tecnologías en los estudios filológicos*, Barcelona, Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio, pp. 449-479.
http://liceu.uab.es/~joaquim/publicacions/Fonetica_Tecnol Habla.pdf

Llisterri, J. - Carbó, C. - Machuca, M. J. – de la Mota, C. - Riera, M. - Ríos, A. (2003a), "La conversión de texto en habla: aspectos lingüísticos", en Martí, M. A. - Llisterri, J. (Eds.), *Tratamiento del lenguaje natural II*, Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria. (en prensa).
http://liceu.uab.es/publicacions/Linguistica_CTH_FDS02.pd f

Llisterri, J. - Machuca, M.J. – de la Mota, C. - Riera, M. - Ríos, A. (2003b), "Entonación y tecnologías del habla", en Prieto, P. (Ed.), *Teorías de la entonación*, Barcelona, Ariel (en prensa).
http://liceu.uab.es/~joaquim/publicacions/Ariel_Aplicacio nes.pdf

Llisterri, J. - Martí, M.A. (2002, "Las tecnologías lingüísticas en la Sociedad de la Información", en Martí, M.A.- Llisterri, J. (Eds.), *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, Barcelona, Edicions Universitat de Barcelona - Fundación Duques de Soria, pp. 13-28.

Llisterri, J. (1999) "Corpus orals per a la fonètica i les tecnologies de la parla", *Actes del I Congrés de Fonètica Experimental*, Tarragona, 22, 23 i 24 de febrer de 1999, Universitat Rovira i Virgili - Universitat de Barcelona. pp. 27-38.
http://liceu.uab.es/~joaquim/publicacions/Tarragona_99/Re sum_tarragona_99.html

Llisterri, J. (2001a), "El habla como medio de acceso a la Sociedad de la Información", *La Musa Digital*, 1 (Monográfico: El impacto social de las nuevas tecnologías. La Sociedad de la Información). *La Musa. Pensamiento, Universidad y Red*, 1, pp. 39-44.
http://www.uclm.es/ab/humanidades/lamusa/paginas/monografico/Llisterri.htm

Llisterri, J. (2001b), "La conversión de texto en habla", *Quark. Ciencia, Medicina, Comunicación y Cultura*, 21, pp. 79-89.
http://liceu.uab.es/~joaquim/publicacions/Quark2001/Llisterri_2001.html

Llisterri, J. (2002a), "Las tecnologías del habla: Entre la ingeniería y la lingüística", *Actas del Congreso Internacional La Ciencia ante el Público. Cultura humanística y desarrollo científico y tecnológico*, Universidad de Salamanca, Salamanca, 28-31 de octubre 2002, pp. 51-74.
http://liceu.uab.es/~joaquim/publicacions/TecnolHab_Salamanca_02.pdf

Llisterri, J. (2002b), "Lingüística y tecnologías del lenguaje", *Lynx. Panorámica de Estudios Lingüísticos* (Departament de Teoria dels Llenguatges, Universitat de València) (en prensa).
http://liceu.uab.es/~joaquim/publicacions/TecnoLing_Lynx02.pdf

Machuca, M.J.- Bueno, L.- Calonge, R.- Estruch, M.- Riera, M. (2000), "Corpus de diàleg", *I Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*, Institut d'Estudis Catalans, Barcelona, 4 i 5 d'abril de 2000.
http://liceu.uab.es/publicacions/SFI_UAB_Corpus_Dialeg.pdf;

Marx, M. - Schmandt, C. (1996) "Mailcall: Message presentation and navigation in a nonvisual environment", *Proceedings of ACM CHI96 Conference on Human Factors in Computing Systems,* Vancouver, April, 1996, ACM Press, pp 165-172.
http://www.media.mit.edu/speech/papers/1996/marx_CHI96_mailcall.pdf

MASK, Multimodal-Multimedia Automated Service Kiosk, Spoken Language Processing Group, Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, Centre National de la Recherche Scientifique, Orsay.
http://www.limsi.fr/Recherche/TLP/mask.html

Minker, W.- Bennacef, S. (2001), *Parole et dialogue homme-machine,* Paris, Étidions Eyrolles - Éditions du CNRS.

Möbius B. - van Santen J. (1996), "Modelling Segmental Duration in German Text-to-Speech Synthesis", *ICSLP 96, The Fourth International Conference on Spoken Language Processing*, Philadelphia, PA, October 3 - 6, pp. 2395-2398.
http://www.bell-labs.com/project/tts/bmo-icslp96.ps

Montero, J.M. - Gutiérrez-Arriola, J. - Colás, J. - Macías-Guarasa, J. - Enríquez, E - Pardo, J.M. (1999), "Development of an Emotional Speech Synthesiser in Spanish", *Eurospeech'99, 6th European Conference on Speech Communication and Technology.*, Budapest, September 5-9, 1999, Vol. 5, pp. 2099-2102.
http://lorien.die.upm.es/~macias/doc/pubs/eurosp99/submitted/m058.pdf

Moreno, L. - Martínez, G. (1988), *Estudio de los signos de puntuación y matices de las pausas en los distintos lenguajes y registros*, Manuscrito no publicado, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona.

Mota, C. de la (1995), *La representación gramatical de la información nueva en el discurso,* Tesis Doctoral, Departament de Filologia Espanyola, Universitat Autònoma de Barcelona.

Mota, C. de la (1997), "Prosody of Sentences with Contrastive New Information in Spanish", en Botinis, B. - Kouroupetroglou, G. - Caryannis, G. (Eds.), *Theory, Models and Applications. Proceedings of an ESCA Workshop,* Athens, European Speech Communication Association, pp. 75-78.

Nadeu, C. (2001), "Representación de la voz en el reconocimiento del habla", *Quark. Ciencia, Medicina, Comunicación y Cultura* 21, pp. 63-71.
http://www.imim.es/quark/num21/021063.htm

Navarro Tomás, T. (1916), "Cantidad de las vocales acentuadas", *Revista de Filología Española*, 3, pp. 387-407.

Navarro Tomás, T. (1917), "Cantidad de las vocales inacentuadas", *Revista de Filología Española*, 4, pp. 71-388.

Navarro Tomás, T. (1918b), "Diferencias de duración entre las consonantes españolas", *Revista de Filología Española*, 5, pp. 367-3938.

Navarro Tomás, T. (1918a), *Manual de pronunciación española*, Madrid, Consejo Superior de Investigaciones Científicas, 1990[24].

Navarro Tomás, T. (1945), *Manual de entonación española*, New York, Hispanic Institute. Cuarta edición: Madrid, Guadarrama, 1974.

O'Shaughnessy, D. - Barbeau, L. - Bernardi, D. - Archambault, D. (1988), "Diphone Speech Synthesis", *Speech Communication*, 7, pp. 55-65.

Olive, J.P. (1998), "'The Talking Computer': Text to Speech Synthesis", en Stork, D.G. (Ed.), *Hal's Legacy: 2001's Computer as Dream and Reality*, Cambridge, Mass., The MIT Press.
http://mitpress.mit.edu/e-books/Hal/chap6/six1.html

Pachès, P. - de la Mota, C. - Riera, M. - Perea, M. P. - Febrer, A. - Estruch, M. - Garrido, J. M. - Machuca, M. J. - Ríos, A. - Llisterri, J. - Esquerra, I. - Hernando, J. - Padrell, J. - Nadeu, C. (2000), "Segre: An automatic tool for grapheme-to-allophone transcription in Catalan", en Ó Croinín, D. (Ed.), *Proceedings of the Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic*

Priorities (LREC-2000 Second International Conference on Language Resources and Evaluation), Athens, 30 May 2000, pp. 52-61.
http://liceu.uab.es/~joaquim/publicacions/Paches_et_al_2000.pdf

Pagel, V. (1999), *De l'utilisation d'informations acoustiques suprasegmentales en reconnaissance de la parole continue,* Thèse Doctorale. LORIA, Laboratoire Lorrain de Recherche en Informatique et ses Applications, Université Henri Poincaré, Nancy.
http://vincent.pagel.free.fr/THESE/

Puigví, D. - Jiménez, D. - Fernández, J.M. (1994), "Parametrización de las pausas ortográficas en castellano. Aplicación a un conversor de texto a habla", *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Córdoba, 20-22 de julio de 1994.

Quilis, A. (1981), *Fonética acústica de la lengua española*, Madrid, Gredos.

Recasens, D. (1993), *Fonètica i fonologia*, Barcelona, Enciclopèdia Catalana.

Reyes, G. (1998), *Manual de Redacción. Cómo escribir bien en español*, Madrid, Arco/Libros.

Riedi, M. (1998), *Controlling Segmental Duration in Speech Synthesis Systems*, PhD Thesis, Computer Engineering and Networks Laboratory, ETH Zurich.
http://www.tik.ee.ethz.ch/~spr/publications/Riedi:98.ps

Ríos, A. (1993), "La información lingüística en la transcripción fonética automática del español", *Boletín de la Sociedad Española para el Procesamiento del Languaje Natural*, 13, pp. 381-387.

Ríos, A. (1994), "El contenido fónico en el Sistema de Diccionarios Electrónicos del Español", en Llisterri, J. - Poch, D. (Eds.), *Actas del XII Congreso Nacional de la Asociación Española de Lingüística Aplicada AESLA, Nuevos Horizontes de la Lingüística Aplicada*, Barcelona, 20-22 de abril, Barcelona: Universidad Autónoma de Barcelona, pp. 333-340.

Ríos, A. (1999), *La transcripción fonética automática del diccionario electrónico de formas simples flexivas del español: estudio fonológico en el léxico*, Estudios de Lingüística Española, 4.
http://elies.rediris.es/elies4/

Rubio, A.J. - Milone, D.H. (2002), "Información prosódica y acentual para el reconocimiento automático del habla", en Díaz García, J. (Ed.), *Actas del II Congreso de Fonética Experimental*, Sevilla 5, 6 y 7 de marzo de 2001, Sevilla, Laboratorio de Fonética, Facultad de Filología, Universidad de Sevilla. pp. 56-77.

Sadek, M.D. - Ferrieux, A. - Cosannet, A. - Bretier, P. - Panaget, F. - Simonin, J. (1996) "Effective human-computer cooperative spoken dialogue: The AGS demonstrator", *ICSLP 96, The Fourth International Conference on Spoken Language Processing.*, Philadelphia, October 3 – 6 1996, pp. 169-173.
http://www.asel.udel.edu/icslp/cdrom/vol1/790/a790.pdf

San-Segundo, R. - Montero, J.M. - Ferreiros, J. - Macías-Guarasa, J. - Pardo, J.M. (2001), "Sistema de información ferroviaria por teléfono: propuesta de una metodología de diseño de gestores de diálogo", *Proceedings of the Second International Workshop on Spanish Language Processing and Language Technologies (SLPLT-2),* Jaén, September 2001, pp. 241-245.
http://lorien.die.upm.es/~macias/doc/pubs/slplt01/MetodologiaDialogov3.pdf

Selkirk, E.O. (1984), *Phonology and Syntax: The Relation between Sound and Structure,* Cambridge, Mass., The MIT Press.

Sproat, R. - Ostendorf, M. - Hunt, A. (Eds.) (1999), *The Need for Increased Speech Synthesis Research: Report of the 1998 NSF Workshop for Discussing Research Priorities and Evaluation Strategies in Speech Synthesis*.
http://www.research.att.com/~rws/newindex/report10.pdf

Strik, H. - Cucchiarini, C. (1999), "Modeling pronunciation variation for ASR: A survey of the literature", *Speech Communication*, 29, 2-4, pp. 225-246.
http://lands.let.kun.nl/TSpublic/strik/a64b.html

Tapias, D. (2002), "Interfaces de voz con lenguaje natural", en Martí, M.A.- Llisterri, J. (Eds.), *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita*, Barcelona, Edicions Universitat de Barcelona - Fundación Duques de Soria, pp. 189-207.

Veldhuijzen van Zanten, G. (1999), "User modelling in adaptive dialogue management", *Eurospeech'99, 6th European Conference on Speech Communication and Technology.*, Budapest, September 5-9, 1999, Vol. 3, pp. 1183-1186.
http://www.niii.kun.nl/~veldhvz/papers/Eurospeech99.pdf

Villarrubia, L. - Garrido, J.M. - Relaño, J. - Caminero, J. - Escalada, J.G. - Rodríguez, M.C. - Hernández, L.A. (2002), "Productos de tecnología del habla para Latinoamérica", *Comunicaciones de Telefónica I+D*, 27, pp. 53-72.
http://www.tid.es/presencia/publicaciones/docs_comtid/numero27.pdf

Waibel, A. (2001), "Los sistemas integrales completos del habla, del lenguaje y la interfaz humana", *Quark. Ciencia, Medicina, Comunicación y Cultura*, 21, pp. 95-102.
http://www.imim.es/quark/Num21/021095.htm

Walker, M.A. - Litman, D. - Kamm, C. - Abella, A. (1997), "PARADISE: A framework for evaluating spoken dialog agents", *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid, July 7-12, 1997. pp. 271-280.
http://www.research.att.com/~walker/acl21.pdf

Wang, C. (2001), *Prosodic Modeling for Improved Speech Recognition and Understanding*, PhD Dissertation. Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

http://www.sls.lcs.mit.edu/sls/publications/2001/Wang_phd_thesis.pdf

Zue, V. (1997), "Conversational interfaces: advances and challenges", *Eurospeech'97. 5th European Conference on Speech Communication and Technology.* Rhodes, Greece, 22-25 September 1997. Vol. 1. pp. KN-9 - KN 18.
http://www.sls.lcs.mit.edu/sls/publications/1997/eurospeech97-zuekeynote.pdf

Zue, V. - Cole, R. - Ward, W. (1997), "Speech Recognition", en Cole, R.A. - Mariani, J. - Uszkoreit, H. - Zaenen, A. - Zue, V. (Eds.). *Survey of the State of the Art in Human Language Technology*, Cambridge, Cambridge University Press. pp. 4-10.
http://cslu.cse.ogi.edu/HLTsurvey/ch1node4.html