LLISTERRI, J. (2003) "Lingüística y tecnologías del lenguaje", *Lynx. Panorámica de Estudios Lingüísticos* (Departament de Teoria dels Llenguatges, Universitat de València) 2: 9-71.
http://liceu.uab.es/~joaquim/publicacions/TecnoLing_Lynx02.pdf

Lingüística y tecnologías del lenguaje

Joaquim Llisterri
Universitat Autònoma de Barcelona

1. Lingüística y tecnologías del lenguaje

Empieza a ser ya relativamente habitual en nuestra vida cotidiana utilizar servicios como los que ofrecen los portales de voz, escuchar el correo electrónico leído a través del teléfono móvil, dictar textos que se escriben automáticamente en la pantalla del ordenador, consultar páginas web en otras lenguas mediante los traductores automáticos que ofrece la red o, desde hace ya bastante tiempo, redactar documentos con la ayuda de los correctores ortográficos y gramaticales integrados en los procesadores de textos. Sin embargo, seguramente pocas personas que emplean estos sistemas son conscientes de que en su desarrollo no únicamente intervienen informáticos e ingenieros, sino que también participan muy a menudo los lingüistas.

Mientras que el gran público ve al lingüista, en el mejor de los casos, como un erudito que pasa buena parte del tiempo aprendiendo lenguas o corrigiendo los "errores" que con pertinaz regularidad se empeñan en cometer los hablantes, la realidad es que encontramos hoy en día profesionales de la lingüística que dedican sus esfuerzos a la creación de los servicios y aplicaciones que acabamos de citar. Especialidades como la Lingüística Computacional, que se enseña en las facultades de Filología, han encontrado su lugar, junto a la informática o la ingeniería de telecomunicaciones, en equipos multidisciplinares que orientan su trabajo al desarrollo de las tecnologías del lenguaje.

Las tecnologías del lenguaje (TL o LT, *Language Technologies*), también conocidas como tecnologías lingüísticas o tecnologías para el lenguaje humano (TLH o HLT, *Human Language Technologies*), son todas aquellas que se integran en aplicaciones informáticas para permitir el tratamiento de textos escritos –como en el caso de la traducción automática o la corrección ortográfica–, o el procesamiento del habla –requerido para el dictado automático o la lectura en voz alta de un mensaje de correo electrónico-. Se trata, en conjunto, de tecnologías que hacen posible la creación de herramientas pensadas para ayudarnos a utilizar los ordenadores sin renunciar por ello a nuestro uso habitual del lenguaje como medio de interacción y de intercambio de información (Cole *et al.*, 1997; *HLT Central*; *Language Technology World*; Llisterri y Martí, 2002; Martí, 2001; Uszkoreit, 2002).

En este contexto se utiliza también el término ingeniería lingüística (IL o LE, *Language Engineering*) para referirse a la aplicación de las técnicas informáticas al desarrollo de aplicaciones que incluyen componentes relacionados con el tratamiento del lenguaje y del habla (*Ingeniería lingüística*; Martí y Llisterri, 2001; Pierrel, 2000). En cambio, con el uso de la expresión "industrias de la lengua" se pretende reflejar el potencial económico y comercial del ámbito que nos ocupa. Existe igualmente la denominación "lingüística informática", que suele hacer referencia al uso de herramientas informáticas en la investigación lingüística. Finalmente, "lingüística computacional" (LC o CL, *Computational Linguistics*) podría entenderse como la disciplina que abarca tanto el procesamiento del lenguaje como el del habla desde una perspectiva general o desde un punto de vista teórico (Gómez, 2000a, b; Grishman, 1986; Jurafsky y Martin, 2000; Sidorov, 2001; Uszkoreit, 2000), aunque en ocasiones se encuentra esta denominación empleada como sinónimo de "procesamiento del lenguaje natural"[1].

En las tecnologías lingüísticas suele distinguirse entre las que se centran en la lengua escrita y las que tienen por objeto el habla. Las primeras se engloban en el campo conocido como procesamiento del lenguaje natural –aunque también podrían definirse como tecnologías del texto escrito-, mientras que las segundas se denominan tecnologías del habla. El desarrollo de estas tecnologías y sus aplicaciones requieren disponer de los llamados recursos lingüísticos, entre los que se cuentan los corpus, los diccionarios y las gramáticas.

El presente trabajo pretende describir la labor del lingüista en los ámbitos mencionados, tanto en lo que se refiere al desarrollo de las tecnologías y recursos básicos como en lo que concierne la creación de aplicaciones que directamente puedan integrarse en programas informáticos de uso local, en la red o en entornos que requieran la interacción entre personas y ordenadores. En ningún caso tiene intención de presentar exhaustivamente las tecnologías del lenguaje, ni tampoco describir con detalle las diversas áreas que las constituyen; por tal motivo, se ha intentado recoger una bibliografía básica para cada tema, de modo que el lector pueda profundizar en las cuestiones que más le interesen. Cabe advertir también que se han primado los aspectos lingüísticos más genéricos sobre los tecnológicos, por lo que, en muchos casos, no se refleja la complejidad real de las diversas técnicas y aplicaciones que se discuten. Finalmente, aunque existe una actividad muy notable en la creación de tecnologías y recursos para el catalán, el gallego y el vasco, tanto los ejemplos

[1] Una discusión más detallada de la terminología y de las relaciones entre estas disciplinas puede encontrarse en Moure y Llisterri (1996).

como las referencias bibliográficas se centran, esencialmente por razones de espacio, en los trabajos sobre el español llevados a cabo en España[2].

Para facilitar la exposición, se mantiene la división tradicional entre tecnologías del habla (apartado 2), tecnologías del texto (apartado 3) y recursos lingüísticos (apartado 4). Sin embargo, las tecnologías dependen, para su funcionamiento eficaz, de la existencia de recursos, por lo que no pueden separarse las unas de los otros. Debe tenerse también en cuenta que las fronteras entre el trabajo que se lleva a cabo en el campo de las tecnologías del habla y en el de las del texto se difuminan cada vez más, especialmente si se consideran aplicaciones como la traducción automática del habla (3.2.3) o la recuperación de información en archivos sonoros (3.2.4), así como la progresiva incorporación de las herramientas propias del procesamiento del texto escrito a la síntesis, al reconocimiento y al diálogo.

2. Tecnologías del habla

Como se ha señalado, las tecnologías del habla (*Speech Technologies*) tienen por objeto el tratamiento informático de la lengua oral y permiten que un ordenador ofrezca información hablada –síntesis del habla-, reconozca los enunciados emitidos por un locutor –reconocimiento automático del habla– o combine ambas tecnologías para entablar una interacción con el fin de recabar información o realizar transacciones –sistemas de diálogo– en una o varias lenguas (Cortázar *et al*., 2002; Huang *et al*., 2001; Lleida, 2000; Llisterri, 2001a, c; O'Shaughnessy, 1987; Rodríguez *et al*., 2001).

El desarrollo de las tecnologías del habla ha estado tradicionalmente ligado al mundo de la ingeniería de telecomunicaciones y, posteriormente, al de la informática. Esto se explica por su origen histórico en el campo de la telefonía y, posteriormente, por su estrecha relación con el tratamiento digital de señales. Sin embargo, llegó un momento en que algunos centros de investigación

[2] En lo que se refiere al catalán, remitimos al lector a las páginas del TALP-Tecnologies i Aplicacions del Llenguatge i de la Parla de la Universidad Politécnica de Cataluña, del CliC–Centre de Llenguatge i Computació de la Universitat de la Universitat de Barcelona, del IULA–Institut Universitari de Lingüística Aplicada de la Universitat Pompeu Fabra, de la Secció de Teoria del Senyal de la Universitat Ramon Llull y a las del IEC-Institut d'Estudis Catalans. Para el gallego, pueden consultarse las del Grupo de Tratamiento de la Señal de la Universidade de Vigo, del SLI-Seminario de Lingüística Informática de la misma universidad, del grupo COLE–Compiladores y Lenguajes de la Universidade de La Coruña, del ILGA–Instituto da Lingua Galega y del Centro Ramón Piñeiro. Los trabajos sobre el vasco pueden verse en las páginas de los grupos IXA, AhoLab y del Grupo de Reconocimiento de Formas y Tecnología del Habla de la Universidad del País Vasco, y en las del grupo DELI- Grupo de Lingüistas, Informáticos e Ingenieros de Deusto de la Universidad de Deusto.

comprendieron que el tratamiento automático del habla puede beneficiarse de los conocimientos propios de diversas ramas de la lingüística y, por tal motivo, iniciaron una colaboración regular con equipos de lingüistas o los incorporaron a sus propias plantillas.

En este apartado presentaremos aquellos ámbitos de las tecnologías del habla que mantienen una mayor vinculación con la lingüística: la síntesis del habla (2.1) –centrándonos especialmente en la conversión de texto en habla (2.1.2.)-, el reconocimiento del habla (2.2) y los sistemas de diálogo (2.3.). Como se expone más adelante, la fonética, tanto en su vertiente descriptiva como experimental, es quizás la disciplina lingüística que más directamente entronca con las tecnologías del habla (Greenberg, 2001; Llisterri, 2002; Llisterri *et al*., 1999), aunque en ciertas ocasiones sea necesario recurrir a conocimientos fonológicos, morfológicos, sintácticos, semánticos e incluso pragmáticos.

2.1. Síntesis del habla

El objetivo de la síntesis del habla es la generación automática de mensajes orales, partiendo de un texto escrito, en la denominada conversión de texto en habla, o de otros tipos de representación simbólica (Cole, 1997a). Por ello, la síntesis puede considerarse en cierto modo un modelo de la producción humana del habla, con independencia de que para las diferentes aplicaciones que en la actualidad se encuentran en el mercado se utilicen estrategias muy distintas a las que activan los hablantes a la hora de convertir un pensamiento en un enunciado. En este apartado nos centraremos en el proceso de conversión de un texto en su equivalente sonoro (2.1.1) y en la evaluación de los resultados de la conversión (2.1.2).

2.1.1. Conversión de texto en habla

Un sistema de conversión de texto en habla (CTH o TTS, *Text-to-Speech Synthesis*) transforma automáticamente cualquier texto escrito y disponible en formato electrónico en su correspondiente realización sonora (Dutoit, 1997, 1999; Llisterri, 2001b; Olive, 1998). La estructura de un conversor es habitualmente modular (Figura 1), de manera que cada módulo se ocupa de un aspecto de conversión de la cadena inicial de caracteres –es decir, el texto- hasta llegar a la señal sonora, equivalente a su lectura en voz alta[3].

[3] Pueden encontrarse demostraciones de sistemas de conversión de texto en habla en español realizados en universidades españolas en las páginas del Grupo de Tecnología del Habla de la Universidad Politécnica de Madrid, del ECA-SIM-Grupo de Computación Avanzada y Entornos de Comunicación Multimodal de la Universidad de Valladolid, del Grupo de Tecnologías de las Comunicaciones de la Universidad de Zaragoza, del TALP y del Grupo de Tratamiento de la Señal de la Universidad de Vigo.

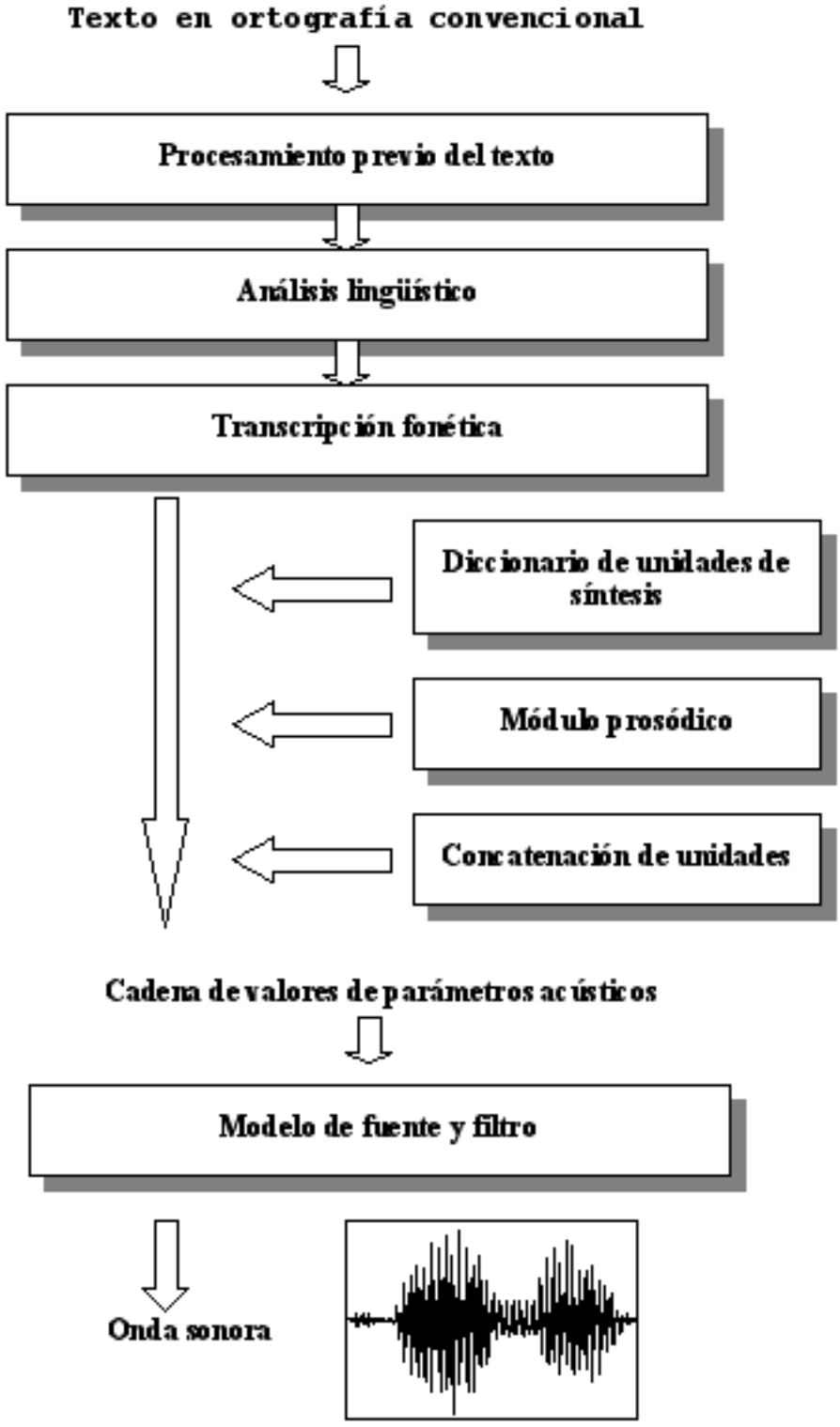

Figura 1: Principales módulos de un conversor de texto en habla

El primer módulo de un conversor tiene esencialmente como misión transformar en texto deletreado abreviaturas, siglas, acrónimos, números, ordinales, fechas, horas, medidas, números romanos o símbolos especiales

como los de las monedas, porcentajes, los relativos a la propiedad intelectual o los que se utilizan en direcciones de Internet.

En general, se elaboran tablas que asocian a cada uno de estos elementos con su representación ortográfica completa, de modo que RAE se expandiría como "rae", ONG como "o ene ge" y CSIC como "ce sic". En algunos casos es preciso un tratamiento un poco más complejo: por ejemplo, "200" debe deletrearse como "doscientos" o "doscientas" en función del género del nombre que aparece a continuación y, de un modo análogo, el símbolo del euro se expande en singular o en plural según la cantidad que anteceda.

Una vez el texto de entrada se ha convertido ya en una cadena de caracteres, es preciso transformar la forma ortográfica en una representación más cercana a la forma sonora, tarea que aborda el módulo de transcripción fonética automática (Enríquez, 1991; Pachès *et al*., 2000; Ríos, 1999).

En conjunto, suelen utilizarse dos tipos de estrategias para la transcripción: un diccionario ayudado por un analizador en lenguas con una correspondencia muy irregular entre grafía y sonido –por ejemplo, el inglés-, o bien un conjunto de reglas complementadas por un diccionario de excepciones en lenguas con una correspondencia regular entre sonido y grafía, como suele ser el caso en español.

Las reglas de transcripción constituyen una sistematización de la correspondencia que se establece en la lengua entre grafemas y alófonos, formalizando una información que se encuentra en los manuales de pronunciación y en los trabajos sobre ortología. Es importante destacar que en esta fase de la conversión de texto en habla se toman decisiones que se sitúan inequívocamente en el ámbito de la lingüística. Por una parte, se define el inventario de alófonos utilizados en el conversor, lo que implica considerar las descripciones fonéticas y fonológicas de la lengua para evaluar la pertinencia de incorporar un mayor o menor grado de variación alofónica. Por otra, la aplicación de las reglas de transcripción determina el nivel de adecuación fonética del conversor a las normas habituales de pronunciación y fija también cómo se reflejan las variantes geográficas o las de estilo.

Es necesario también encontrar soluciones para la pronunciación de los nombres extranjeros, adoptando una versión más o menos cercana a la nativa según las tradiciones de cada lengua, con el problema añadido de que los sonidos que no formen parte de la lengua deberán incorporarse al inventario de alófonos para la síntesis; así, un sistema de síntesis del catalán deberá incluir la fricativa velar y la fricativa interdental sordas, inexistentes en esta lengua, para pronunciar nombres castellanos como "Juan López", que pueden aparecer fácilmente en un texto.

Una vez el texto está transcrito fonéticamente, es preciso realizar un análisis lingüístico que complementa la tarea de otros módulos del conversor. Para tal fin se utilizan herramientas desarrolladas en el campo del procesamiento del lenguaje natural (véase el apartado 3.1.2.), como los analizadores morfológicos y los analizadores sintácticos (Monzón *et al*., 1993). La información sobre las partes de la oración es relevante para evitar, por ejemplo, la aparición de una pausa entre un adjetivo y un nombre o la colocación de acento en un artículo; por su parte, el análisis en constituyentes sintácticos es casi imprescindible en la determinación de las unidades melódicas para dotar al texto leído de una prosodia adecuada. Aspectos más complejos como la detección del foco de una oración podrían también tratarse como parte del análisis lingüístico, dada su incidencia en la entonación.

Una de las áreas a la que más esfuerzos se dedican en la actualidad en el ámbito de la conversión de texto en habla es, sin lugar a dudas, la prosodia, pues de ella depende en gran medida la naturalidad de la lectura (Monaghan, 2002; van Santen, 1997). El módulo prosódico de un conversor consta de un conjunto de reglas que especifican la duración –y en algunos casos la intensidad- de los segmentos, el contorno melódico del enunciado, las modificaciones acústicas producidas por el acento y la colocación y la duración de las pausas.

La definición de reglas que sistematicen la información prosódica suele llevarse a cabo partiendo del análisis de un corpus de habla natural (véase el apartado 4.1.1), en cuyo diseño no puede dejar de intervenir un experto familiarizado con la descripción suprasegmental de la lengua sobre la que se trabaja. En la actualidad, es técnicamente posible adquirir los datos necesarios de un modo relativamente automático partiendo de un corpus suficientemente grande y convenientemente etiquetado, pero aun así, parece lógico pensar que el propio proceso de etiquetado debe responder a criterios lingüísticos si se desea ir más allá de un modelo que cumpla la única función de proporcionar datos estadísticos "ciegos" a un sistema concreto.

En lo que se refiere a la duración, el análisis acústico del habla natural ha puesto de manifiesto la existencia de diversos factores que condicionan los valores temporales de cada uno de los segmentos en un enunciado; entre los más relevantes cabe citar el acento, la longitud de la palabra en la que se encuentra el segmento, la consonante o la vocal que le sigue, la existencia de pausa después del segmento, la posición del mismo en el enunciado y, naturalmente, la velocidad de elocución del hablante.

Por este motivo, un conversor de texto en habla que se proponga alcanzar un elevado grado de naturalidad incorpora un modelo de duración segmental que considera la duración intrínseca de cada segmento del habla y define mediante reglas las modificaciones contextuales a las que está sometido por la

influencia de los factores citados anteriormente (Macarrón *et al.,* 1991; Santos *et al*., 1988). Las reglas de duración así entendidas implican, al igual que en los casos anteriores, una formalización y una sistematización del conocimiento adquirido a partir del análisis del habla natural mediante los métodos y los instrumentos que proporciona la fonética.

La intensidad de cada segmento varía también en la producción del habla. Este aspecto no suele tenerse en cuenta en muchos de los sistemas de conversión de texto en habla, pero podrían considerarse estrategias análogas a las usadas para la asignación de duración segmental, basadas igualmente de los datos derivados del estudio de corpus.

Para una lectura inteligible, un sistema de conversión de texto en habla debe insertar las pausas marcadas mediante signos de puntuación en los textos, asignándoles una duración diferente en función de cada tipo de signo –distinguiendo, por ejemplo, una coma de un punto y aparte-, pero precisa tener en cuenta también, para una lectura natural, la inserción de pausas no marcadas ortográficamente. Si pensamos, además, en aplicaciones como la consulta del correo electrónico a través del teléfono, es fácil darse cuenta de que no siempre se contará con textos adecuadamente puntuados, con lo que el análisis morfológico y sintáctico adquiere una importancia primordial a la hora de determinar el lugar en el que debe intercalarse una pausa para una buena comprensión del texto.

Por último, el módulo prosódico de un conversor se ocupa de determinar el patrón melódico –la entonación- que corresponde a cada fragmento del texto de entrada (Escudero y Cardeñoso, 2001; Fernández y Rodríguez, 2000; Garrido *et al*., 2000; Llisterri *et al*., 2003; López y Hernández, 1995). En el habla natural, es sabido que la melodía, entre otras funciones, señala la modalidad oracional, estructura el enunciado en unidades entonativas estrechamente relacionadas con el significado y aporta información sobre aspectos pragmáticos. La melodía constituye también un indicador del estado emocional del hablante, de su estatus sociocultural y de su procedencia geográfica. Debido a estos y a otros factores, modelar acertadamente los movimientos melódicos es esencial para conseguir una conversión de texto en habla de calidad.

Tal como señala Beaugendre (1996), en síntesis se utilizan tres estrategias diferentes para la generación de movimientos melódicos: sistemas de reglas que definen la forma de la curva melódica a partir de un conjunto de símbolos, patrones melódicos previamente almacenados y, finalmente, curvas melódicas obtenidas automáticamente a partir de un corpus mediante técnicas estadísticas, como son los Modelos Ocultos de Markov o las redes neuronales. Obviamente, el procedimiento que incorpora un mayor grado de conocimiento fonético es el primero de los citados, ya que relaciona la realización acústica de la curva

melódica con su representación a un nivel abstracto que podría considerarse fonológico. Para ello se requiere contar con un sistema de transcripción o anotación de la entonación que, en muchos casos, presupone una teoría prosódica orientada por principios lingüísticos (Quazza y Garrido, 1998).

En algunos conversores la asignación de la curva melódica se basa en un análisis de la estructura entonativa de los enunciados (*prosodic parsing*), para lo que es necesario contar con un modelo prosódico de base lingüística que defina los diversos tipos de unidades entonativas, distinguiendo entre las de tipo local –el grupo acentual, por ejemplo– y las más globales –los grupos melódicos– (Garrido, 2001).

En las últimas etapas de la conversión de texto en habla, una vez se dispone de la transcripción fonética y de la correspondiente información prosódica asociada, se realiza la selección de las unidades acústicas –o unidades de síntesis– que darán forma sonora al mensaje. Estas unidades se encuentran en los denominados diccionarios de unidades de síntesis, en cuya confección intervienen también los expertos en fonética.

El diseño de un diccionario de unidades de síntesis se inicia con la definición del inventario completo de unidades –alófonos y fonemas– de la lengua sobre la que se trabaja, tarea de la que también dependen, como hemos visto, las reglas de transcripción automática. Esto implica, por ejemplo, en el caso del español, decidir si se incluyen los alófonos abiertos de /e/ y /o/, los alófonos nasalizados de las vocales, los alófonos velar y palatal de /a/, o si en el diccionario se almacenan por separado muestras de vocales acentuadas y de vocales no acentuadas. En el consonantismo, por ejemplo, es preciso considerar si se incluyen los alófonos interdentales y dentales de /n/ y /l/, la aproximante alveolar o sonidos como la fricativa palatal sonora que no son propios del español pero que se pueden encontrar en la pronunciación de nombres extranjeros. Las decisiones se toman considerando factores fonéticos como las diferencias acústicas entre alófonos, la aparición regular del alófono condicionada por el contexto, o su variación libre en la lengua; intervienen también consideraciones más generales como la economía del inventario y el modelo de pronunciación deseado.

En segundo lugar, es preciso definir el tipo de unidad que va a utilizarse. En la conversión de texto en habla son habituales los difonemas (denominados, también, con más propiedad, dialófonos), que consisten en una combinación entre la mitad del primer sonido que lo forma y la mitad del segundo. Con ello se persigue que, a la hora de concatenar las unidades para sintetizar un determinado mensaje, la unión entre una unidad y otra se produzca por las partes en las que existe una menor variación acústica –es decir, en el "centro" de un sonido– y no por aquellas en las que se encuentra la transición de un sonido a

otro. Para sintetizar la palabra "mesa" mediante difonemas se recurriría a juntar [me] con [es] y [es] con [sa], de modo que la unión de realizaría entre una mitad y otra mitad de [e] y entre una mitad y otra mitad de [s], zonas en las que el tracto vocal se mantiene en una posición relativamente estable en comparación con el momento de cambio de [m] a [e] o de [s] a [a].

La selección del locutor a partir de cuya voz se constituirá el diccionario de unidades de síntesis es también una labor que requiere conocimientos fonéticos, ya que es conveniente valorar, en función del ámbito de uso de la aplicación, los rasgos de pronunciación del locutor que pueden incidir en el resultado de la síntesis. Suele prestarse atención a la presencia de características dialectales o sociolectales marcadas, de idiosincrasias en la producción de un determinado sonido o patrón entonativo, a la existencia de interferencias con otras lenguas –esto es especialmente relevante en el caso de hablantes bilingües– y a la capacidad de adaptación a la tarea que, como se verá más adelante, requiere en ciertos casos un elevado grado de control de la articulación y de los elementos prosódicos.

Una cuarta etapa en la que la presencia del lingüista es relevante es el momento de la grabación de las unidades de síntesis. Estas unidades se acostumbran a insertar en palabras o en frases para su grabación. Es por ello esencial que un experto en fonética supervise la adecuada pronunciación de cada elemento segmental, además del ritmo, las pausas y la entonación de la lectura. De lo contrario, es probable que se cometan errores que obliguen a repetir la grabación una vez finalizado el proceso, que algunos difonemas sean inservibles o que una realización prosódica inadecuada tenga consecuencias negativas en la calidad final de la síntesis.

Una vez realizada la grabación del corpus de síntesis, las unidades se segmentan, se etiquetan como se muestra en la figura 10 (apartado 4.1.1), y se almacenan en el diccionario, trabajo que implica un buen conocimiento de la fonética acústica para establecer las fronteras entre los segmentos. Aunque este proceso puede automatizarse, en muchas ocasiones es necesaria una revisión manual a cargo de un experto, ya que una mala segmentación de las unidades repercute en el resultado de la conversión, creando la impresión de discontinuidades excesivamente bruscas en el mensaje.

Por razones de economía, las unidades de síntesis se guardan de forma parametrizada, acudiendo para ello a los modelos acústicos de producción del habla que proporciona la fonética. Así, cuando el conversor utiliza un sintetizador por formantes, se almacena para cada unidad la información correspondiente a los valores frecuenciales, temporales y de intensidad de los formantes de los sonidos que la componen, tomando como base el modelo de la fuente y el filtro. Precisamente la función del sintetizador es, una vez

seleccionadas las unidades que configurarán un mensaje, convertir los valores de los parámetros acústicos en su correspondiente manifestación sonora. En esta fase debe también considerarse el modo en que se establecen las transiciones entre unidades, modelando de la manera más fiel posible la coarticulación –es decir, la influencia mutua entre sonidos adyacentes– que se produce en el habla natural y que se ha estudiado con detalle tanto desde la perspectiva de la fonética acústica como de la articulatoria.

Aunque necesariamente hemos simplificado algunos aspectos, es patente que la conversión de texto en habla constituye una tecnología en la que la información lingüística, y muy especialmente la fonética, tiene un papel determinante y una clara incidencia en la calidad de los resultados. Por tal motivo es habitual que los equipos más avanzados que trabajan en este ámbito incorporen especialistas en fonética, que intervienen para dotar de contenido lingüístico a los módulos que lo requieren.

2.1.2. Evaluación de la síntesis

El creciente uso de la conversión de texto en habla en diversos servicios y aplicaciones conlleva la necesidad de contar con evaluaciones que faciliten la tarea de seleccionar el sistema más adecuado para una determinada función y que permitan comparar la calidad de diversas versiones de un mismo conversor o de conversores diferentes.

Las denominadas evaluaciones subjetivas –diferentes de aquellas que realiza el propio desarrollador estudiando la salida de cada uno de los módulos del conversor y determinando los errores que se producen– se basan en la respuesta de un grupo significativo de personas a diversas pruebas mediante las que se pretende cuantificar la inteligibilidad y la naturalidad del sistema de conversión. Para tal fin se preparan una serie de pruebas, algunas de las cuales tienen un contenido netamente fonético y lingüístico (Aguilar *et al*., 1994; Pols, 1996; Pols y Jekosch, 1997).

Uno de los aspectos más básicos que puede evaluarse en un conversor es la inteligibilidad de los elementos segmentales. Las pruebas consisten, por lo general, en la presentación de una serie de estímulos sintetizados para poner de manifiesto los errores que se producen cuando los oyentes deben identificar los sonidos que los forman. Una de las pruebas más habituales es el llamado Test de Rimas, formado por conjuntos de palabras monosilábicas que difieren en un único segmento, sea en posición inicial o final; en la hoja de respuesta se presenta toda la serie –por ejemplo, "ved, ven, ves, ver" o "can, tan, pan, dan"- y los oyentes deben identificar cuál de las palabras es la que han escuchado en su versión sintetizada. Este método entronca directamente con la noción de par mínimo, de modo que el diseño de la prueba se fundamenta en el conocimiento de las oposiciones fonológicas de la lengua.

En otro tipo de pruebas en las que se valora la inteligibilidad de palabras en frases con o sin sentido se utilizan conjuntos de enunciados una de cuyas características principales es que son fonéticamente equilibrados. Tal es el caso de las Frases Psicoacústicas de Harvard o las Frases Semánticamente Anómalas de Haskins, en las que la frecuencia de aparición de cada uno de los fonemas o alófonos corresponde a la frecuencia de aparición propia en la lengua. La adaptación de estas pruebas, inicialmente concebidas para el inglés, a otra lengua, supone disponer de datos fiables en lo que se refiere a la frecuencia de aparición de elementos segmentales, obtenidos preferentemente a partir de un corpus amplio y representativo de la lengua hablada.

Es también importante en muchas ocasiones evaluar la comprensión del habla sintetizada. Para ello se recurre a menudo a la audición de un texto, tras la cual se realizan una serie de preguntas de elección múltiple sobre su contenido, igual que sucede en las pruebas de comprensión auditiva utilizadas en la enseñanza de lenguas extranjeras. La selección de los textos debe tener en cuenta, entre otros, factores semánticos, discursivos y los relacionados con la tipología textual, de modo que la intervención de un experto con buenos conocimientos en estas materias se hace necesaria para garantizar la fiabilidad de los resultados.

Observamos, pues, como la evaluación del resultado de la síntesis se fundamenta, en parte, en conocimientos sobre la estructura de la lengua, de modo que el lingüista encuentra también un lugar en esta etapa, imprescindible en el momento en que se plantea una aplicación real.

2.2. Reconocimiento del habla

En el reconocimiento automático del habla (RAH o ASR, *Automatic Speech Recognition*) se realiza, en cierto modo, la tarea inversa a la que se lleva a cabo en la conversión, puesto que lo que se pretende con esta tecnología es transformar una señal sonora –el habla– es su correspondiente representación simbólica que, en general, será un texto escrito (Cole y Zue, 1997; Deroo, 1999; Kurzweil, 1998; Tapias, 1999, 2002)[4]. Tal es el objetivo, por ejemplo, de los programas comerciales de dictado automático orientados a los usuarios que desean escribir sus textos sin recurrir al teclado del ordenador.

Los reconocedores pueden considerarse, en esencia, como unos sistemas que, en una primera etapa, aprenden automáticamente 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 tal razón, una de las

[4] En las páginas del Grupo de Procesado del Habla de la Universidad Politécnica de Cataluña se encuentran demostraciones de sistemas de reconocimiento de habla en español.

primeras actividades a la hora de desarrollar un sistema de reconocimiento es diseñar y recoger lo que se conoce como corpus de aprendizaje (o de entrenamiento), a partir del cual el sistema adquirirá la información necesaria para crear modelos de cada una de las unidades de reconocimiento, análogas, en ocasiones, a las de la síntesis descritas en el apartado 2.1.1. El corpus de entrenamiento se utiliza también para la obtención de la gramática del reconocedor, entendida como un modelo que recoge las probabilidades de aparición de palabras en un determinado punto.

Al igual que en síntesis se establecen las unidades que se emplearán 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 las realizaciones fonéticas de un gran número de usuarios si, por ejemplo, quiere aplicarse para automatizar un servicio de asistencia telefónica a clientes o emplearse en un portal de voz que ofrezca información general.

El corpus de entrenamiento de un reconocedor debe, por tanto, contener la mayor variedad posible de hablantes para que puedan crearse los modelos –"plantillas" o representaciones internas que posee el sistema- de cada una de las unidades, reflejando la variación individual de las voces, los distintos acentos debidos a factores geográficos o sociolingüísticos y, entre otros elementos, las diferencias en la velocidad de elocución (Strik y Cucchiarini, 1999). La dialectología constituye en esta fase una disciplina casi imprescindible, ya que contribuye a definir las zonas geográficas en las que es necesario recoger muestras, mientras que la demografía establece 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 (Caballero y Moreno, 2001).

Por otra parte, el corpus de entrenamiento debe ser también exhaustivo en lo que a la aparición o cobertura de las unidades de reconocimiento se refiere. Como ocurre en 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 2.1.1). Intervienen también aquí factores relacionados con la frecuencia de aparición, ya que cada unidad debe estar representada un mínimo de veces en el corpus para que la creación de los modelos o plantillas a las que nos referíamos anteriormente sea fiable. Todo ello exige un trabajo fonético de análisis de la lengua en términos estadísticos, así

como también una labor de descripción previa del conjunto de los alófonos aceptados y de su distribución contextual.

Una vez recogido el corpus, éste se somete a un proceso de segmentación y etiquetado automáticos para establecer las fronteras entre los segmentos y las unidades y para sincronizar la representación ortográfica con la señal sonora, tal como se muestra en la figura 10. Nuevamente el conocimiento de la fonética acústica es esencial para establecer criterios homogéneos de segmentación, de etiquetado y de transcripción, así como para revisar manualmente los resultados del tratamiento automático.

Finalmente, un reconocedor incluye también, en muchas ocasiones, un diccionario en el que se encuentran transcritas fonéticamente las palabras que puede aceptar el sistema (véase el apartado 4.2). Es una labor propia del lingüista definir la transcripción canónica de cada palabra, y establecer mediante reglas la relación entre ésta y las variantes que se hayan encontrado en el corpus de entrenamiento o las que puedan preverse en función de la variación fonética documentada en la lengua en la que se desarrolle el sistema.

Toda la información recogida en la fase de entrenamiento se incorpora a los módulos que se muestran en la figura 2 para el reconocimiento de nuevos enunciados. La señal sonora se analiza, en una primera etapa, para extraer los parámetros acústicos que se han considerado relevantes en el momento del diseño (Nadeu, 2001), y después se compara, en el módulo de reconocimiento, con los modelos acústicos de las unidades de reconocimiento que se han almacenado previamente en el sistema; la decisión final suele tomarse con la ayuda de las reglas gramaticales que constituyen el modelo de lenguaje, en las que se definen, a grandes rasgos, la probabilidad de las secuencias de palabras que pueden encontrarse en el contexto de una determinada aplicación.

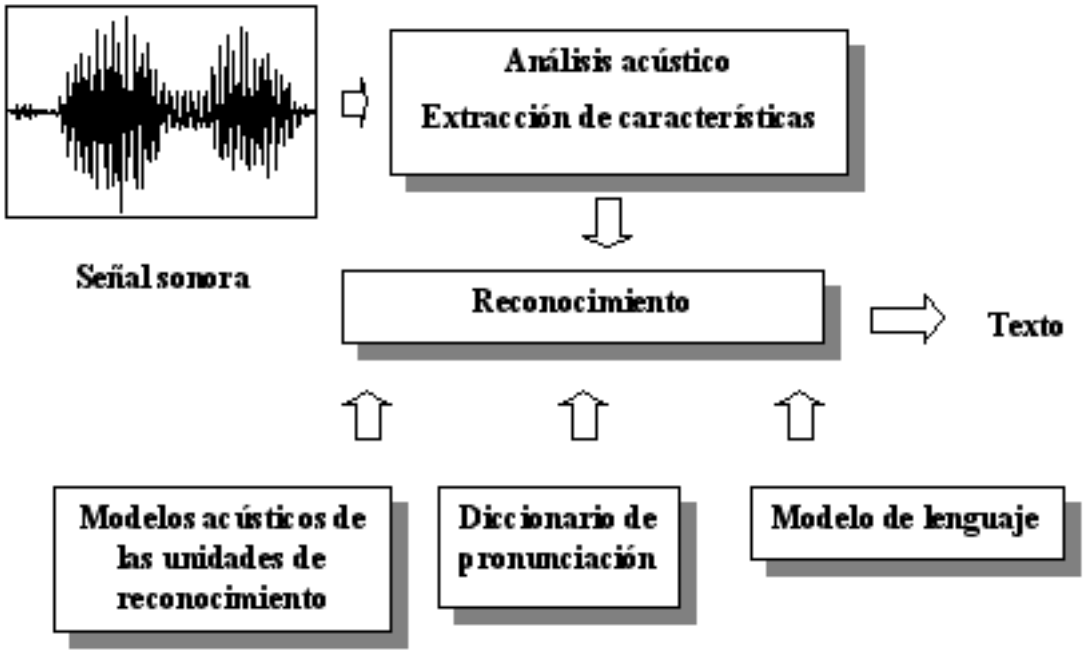

Figura 2: Principales módulos de un sistema de reconocimiento de habla

En conclusión, a pesar de que el papel del conocimiento lingüístico no sea tal vez tan determinante como en la conversión de texto en habla, parece evidente que disciplinas como la fonética y la dialectología son muy relevantes para el reconocimiento automático del habla, en la medida en que aportan información básica y criterios prácticos para el desarrollo de los sistemas.

2.3. Sistemas de diálogo

Los sistemas de diálogo (SLS, *Spoken Language Systems*) tienen como objetivo facilitar la interacción mediante el habla entre una persona y un sistema informático (Bonafonte *et al*., 2000; Gibbon *et al*., 2000; Minker y Bennacef, 2001; Tapias, 2002; Zue, 1999)[5]. En consecuencia, se utilizan en servicios telefónicos automáticos de información y de atención al cliente, o en ámbitos como la banca y el comercio electrónicos. Los sistema de diálogo constituyen también una de las tecnologías básicas que sustentan los denominados portales de voz, a través de los que es posible acceder a una amplia gama de servicios -información meteorológica, cartelera, museos, restaurantes, farmacias de guardia, compañías de taxi, etc.- equivalentes a los que se encuentran en los portales convencionales en la web (Fernández *et al*., 2000).

[5] Se encuentran ejemplos de sistemas de diálogo en español en las páginas del GSTC-Grupo de Investigación en Señales, Telemática y Comunicaciones de la Universidad de Granada y en las del proyecto BASURDE (Sistema de Diálogo para Habla Espontánea en un Dominio Semántico Restringido).

Un sistema de diálogo consta de un conjunto de módulos (Figura 3) que realizan todas las tareas necesarias para facilitar una información o llevar a cabo una transacción. El primero es un reconocedor automático del habla, que procesa las preguntas del usuario y convierte la señal sonora en una representación simbólica accesible al sistema informático. A continuación se lleva a cabo la interpretación semántica del enunciado, a partir de la cual se consulta, si es necesario, una base de datos con la información relevante para proporcionar la respuesta a la petición realizada. Un tercer módulo genera un enunciado completo que contiene los resultados de la consulta o que, en su caso, solicita al usuario que confirme un dato o proporcione una información adicional. Finalmente, un conversor de texto en habla se encarga de transformar los resultados del módulo de generación en su equivalente sonoro para que pueda ser escuchado por el usuario al otro lado del teléfono. Las tareas de estos módulos están, en cierto modo, coordinadas por lo que se conoce como un "gestor del diálogo", que establece, por ejemplo, los turnos de palabra, verifica la coherencia entre la pregunta y la respuesta e interpreta las intervenciones del usuario que hacen referencia a información previa, poniendo en práctica las estrategias diseñadas por los investigadores para que la interacción entre la persona y el sistema automático se lleve a cabo de la forma más natural posible.

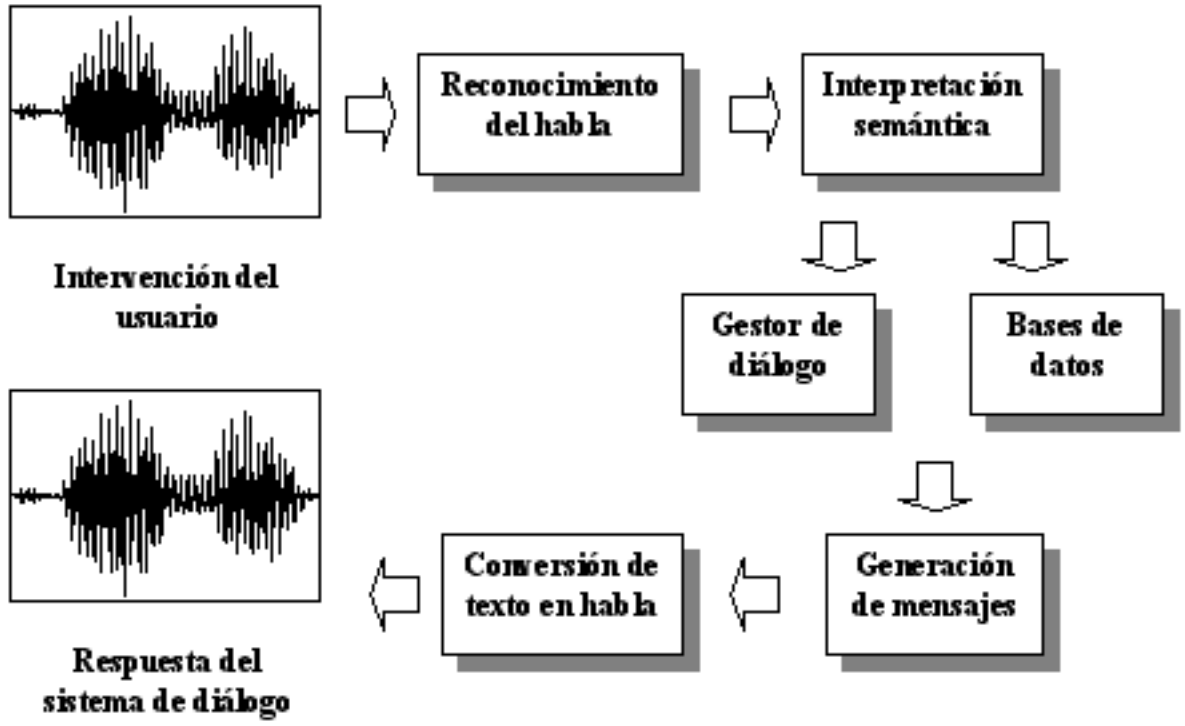

Figura 3: Principales módulos de un sistema de diálogo

El desarrollo de un sistema de diálogo se inicia estudiando con detalle la tarea que se desea automatizar. Para ello se parte de corpus que recogen interacciones auténticas entre personas, como, por ejemplo, grabaciones de

llamadas a un servicio de información convencional. A partir de estos datos se determina el tipo de consultas que se realizan más habitualmente y se definen una serie de escenarios, para cada uno de los cuales se especifican las intervenciones que debe realizar el sistema de diálogo en función de las peticiones del usuario.

Sin embargo, puesto que las personas no actúan del mismo modo cuando se dirigen a un interlocutor humano que cuando se enfrentan a un sistema automático, se recurre, en una segunda fase del diseño, al procedimiento conocido como el "Mago de Oz". En este caso, el usuario que realiza una llamada escucha al otro lado del hilo telefónico una voz sintetizada que le proporciona las respuestas a sus consultas; en la práctica, estas respuestas las decide, en función de los escenarios previamente establecidos, un investigador humano que sigue la conversación y que envía a un conversor de texto en habla los mensajes más adecuados a cada situación. La grabación de las conversaciones proporciona unos datos que responden en mayor medida que los intercambios persona–persona a la situación real de uso de un sistema de diálogo, constituyéndose así un corpus a partir del cual se refina el diseño del sistema.

Los problemas lingüísticos implicados en el desarrollo de un sistema de diálogo son, como puede deducirse, de naturaleza diversa (Llisterri, 2002; Llisterri *et al*., 2003). En primer lugar encontramos los derivados del reconocimiento del habla: un reconocedor, por ejemplo, puede presentar problemas a la hora de distinguir entre "Palencia" y "Valencia", lo que hace necesario que en el módulo de gestión del diálogo se prevean estrategias de confirmación de la información dudosa, preguntando, por ejemplo "¿Desea usted viajar a Palencia o a Valencia?" o "¿Quiere usted saber los horarios de trenes a Palencia?". El reconocimiento de los rasgos prosódicos puede ser también muy relevante para el buen desarrollo del diálogo: así, la interpretación de dos enunciados como "No, quiero viajar por la mañana" o "No quiero viajar por la mañana" depende exclusivamente de la correcta detección de la pausa; de un modo análogo, la diferencia entre una pregunta -"¿Puedo viajar el lunes?"– o una aseveración –"Puedo viajar el lunes"– se manifiesta únicamente en la entonación, por lo que el módulo de reconocimiento debe estar preparado para proporcionar al módulo semántico la información relevante.

En otras ocasiones, los problemas responden a la diferente interpretación de una palabra en función de su aparición en el enunciado: "Mañana", por ejemplo, no debe interpretarse del mismo modo en "Quiero ir mañana a..." que en "Quiero salir por la mañana"; otros elementos como "cuándo" en "¿Cuándo hay trenes para...?" pueden ser ambiguos al referirse, pongamos por caso, al día, al momento del día o a una hora precisa. También el módulo de interpretación

semántica debe enfrentarse a la diversidad de procedimientos con los que puede solicitarse una misma información: "Quiero saber a qué hora hay trenes a...", "¿A qué hora hay trenes a...?", "Necesito información sobre los horarios de trenes a...", "¿Cuándo sale un tren a...?", etc., corresponden, en realidad, a la misma petición de información –al mismo acto de habla, en términos lingüísticos– y deben tratarse, por tanto, de un modo equivalente. La recuperación de las anáforas y de los deícticos es otra cuestión que se aborda en el módulo de interpretación semántica. Muy probablemente un usuario no repita en cada uno de sus enunciados la palabra "billete", por ejemplo, sino que lo sustituya por "lo" tras haberla utilizado una vez. Del mismo modo, los deícticos de tiempo y de lugar deben interpretarse de forma correcta partiendo del conocimiento adquirido por el sistema a lo largo del diálogo.

La generación de respuestas plantea igualmente problemas de tipo lingüístico, análogos a los que se tratan en la generación del lenguaje (véase el apartado 3.2.1), un ámbito propio del procesamiento del lenguaje que encuentra su aplicación en el diseño de las llamadas interfaces en lenguaje natural (Rodríguez, 2001). Además de ser gramaticalmente correctas, las respuestas de un sistema de diálogo deben ser pragmáticamente adecuadas, no sólo en el contenido, sino también en lo que se refiere, por ejemplo, a la cortesía. Un caso especialmente interesante surge en el momento en que el sistema debe confirmar una información que ha obtenido del usuario. En estas situaciones, puede recurrirse a estrategias directas –como "¿Ha solicitado información sobre trenes a Valencia?– o indirectas –del estilo de "Para ir a Valencia...."– lo que implica, en cierto modo, la elección de un tipo u otro de acto de habla[6].

Finalmente, y aunque pueda parecer una obviedad, las respuestas deben ser aceptables desde el punto de vista de la normativa de la lengua. Un sistema de diálogo técnicamente excelente puede ofrecer una imagen pública no demasiado buena de una institución o de una empresa si las respuestas proporcionadas no responden a un mínimo de corrección lingüística, por lo que la revisión a cargo de un experto no es, en absoluto, una cuestión trivial.

Así pues, una vez más comprobamos que una tecnología como la de los sistemas de diálogo requiere la incorporación de conocimientos lingüísticos. Si en los casos anteriores era la información fonética la que podía tener un papel más relevante, la semántica, la pragmática y, muy especialmente, el análisis de la conversación, son disciplinas que no cabe dejar de lado en el diseño y la evaluación de un sistema de diálogo.

[6] Todos los ejemplos de este apartado están tomados de Llisterri *et al*., 2003.

3. Tecnologías del texto

En contraste con las tecnologías del habla, que tienen por objeto el tratamiento de la señal sonora, las que podríamos llamar tecnologías del texto se ocupan de la vertiente escrita de la lengua. Su desarrollo inicial estuvo estrechamente ligado al de la informática y, en particular, a la inteligencia artificial, lo que se comprende fácilmente si recordamos que a mediados de los años 50 el interés primordial en el campo del procesamiento del lenguaje era la traducción automática.

En el ámbito de las tecnologías del texto, podríamos distinguir, aunque únicamente sea para facilitar la exposición, entre las herramientas con las que se procesa la lengua escrita, que se presentan en el apartado 3.1, y las tecnologías empleadas en el desarrollo de aplicaciones, descritas en el apartado 3.2. Al igual que en el caso de las tecnologías del habla, intentaremos poner de relieve las funciones que el lingüista puede desempeñar en la creación de herramientas y de tecnologías, centrando la exposición en el conocimiento lingüístico que cada una de ellas requiere.

3.1. Herramientas para el tratamiento del texto

En lo que se refiere a las herramientas, cabe diferenciar las que llevan a cabo un análisis lingüístico del texto (3.1.2) de las que ayudan a la escritura verificando la ortografía, la gramática o el estilo (3.1.1). Mientras que las primeras se emplean, en general, como parte de una aplicación para el procesamiento del lenguaje, las segundas están directamente disponibles para el usuario final, integradas en los programas más comunes de tratamiento de textos o disponibles en línea para resolver consultas esporádicas.

3.1.1. Herramientas de ayuda a la escritura

Señalábamos al principio que una de las aplicaciones más extendidas de las tecnologías lingüísticas son los programas de corrección ortográfica y gramatical que se encuentran incorporados a la mayoría de los procesadores de textos y que pueden describirse genéricamente como herramientas de ayuda a la escritura (Gómez, 1999, 2001). Es frecuente distinguir, en la corrección automática de textos, tres niveles de complejidad creciente: verificación ortográfica, verificación gramatical y verificación de estilo; como veremos a continuación, todos ellos son susceptibles de incorporar información lingüística para mejorar sus prestaciones[7].

[7] En las páginas de la empresa Daedalus–Data, Decisions, Language SA. pueden verse varias demostraciones de revisión de textos en línea.

Para quien emplea con asiduidad un corrector ortográfico no es difícil percatarse de que ciertos errores en el texto escapan con facilidad al sistema. Por ejemplo, un corrector comercial como el que se usa para escribir este artículo acepta las palabras "segmental" y "supranacional", mientras que señala como un error la aparición de "suprasegmental"; de igual modo, acepta "formante" pero no "formántico". Ello muestra que, como es habitual en estas herramientas, no sólo aparecen como erróneas palabras que contienen equivocaciones ortográficas o de mecanografiado, sino también palabras existentes en la lengua –cierto es que, en este caso, se trata de dos tecnicismos propios de la fonética– pero que no han sido incluidas en el diccionario en el que se basa el corrector. La incorporación de conocimiento morfológico permitiría, en parte, subsanar estos errores, pues el diccionario contendría "supra" marcado como un prefijo que puede añadirse a ciertas clases de palabras e incorporaría también reglas de derivación que permitieran, a partir de la detección del radical, aceptar "formántico".

Otro procedimiento de corrección de errores ortográficos vendría dado por la comparación entre una transcripción fonética de la palabra y las transcripciones similares que se hallaran en el diccionario; para tal fin podrían usarse los resultados de un motor de semejanza fonética como el que se ejemplifica en la figura 4.

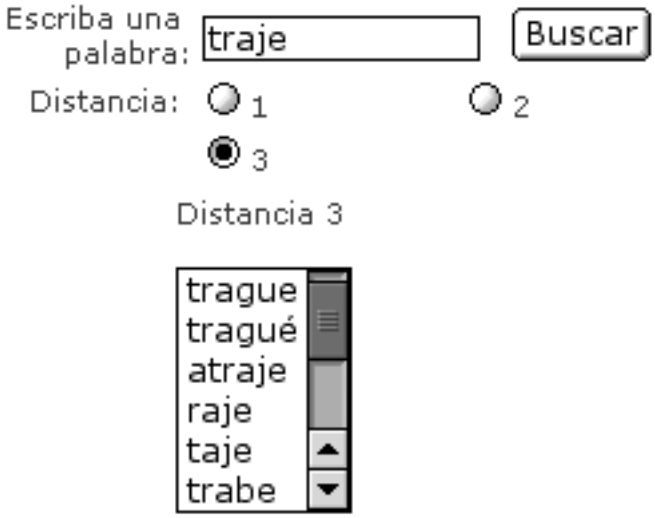

Figura 4: Resultados de la búsqueda en un motor de semejanza fonética de la palabra "traje" (Signum Cia. Ltd.)

Es igualmente una experiencia habitual que un verificador ortográfico no señale errores como la falta de concordancia entre el sujeto y el verbo en casos como "Los niños lee libros". Puesto que la forma "lee" se encuentra en el diccionario, si no se incorpora al proceso de corrección un análisis sintáctico

que divida el enunciado en constituyentes y asigne a cada uno su función sintáctica, el programa no detecta que "lee" debería ser, en realidad, "leen".

Para resolver este último problema existen herramientas como los verificadores gramaticales (Ramírez y Sánchez, 1996; Ramírez *et al*., 1998; Rodríguez *et al*., 1992). De hecho, los sistemas actuales se limitan, por lo general, a comparar secuencias de palabras con unos patrones de errores previamente establecidos. Aunque esta pueda parecer una operación trivial, el establecimiento de unos patrones de validez general para un determinado tipo de error requiere, en ocasiones, un cierto grado de abstracción lingüística complementado con el uso de información morfológica (Gómez, 2001).

Finalmente, los verificadores de estilo comprueban la adecuación del texto a un conjunto de reglas previamente definidas. Una vez el usuario ha establecido el tipo de texto –general, técnico, literario, etc.-, el corrector detecta aquellos elementos que no corresponden a los rasgos aceptados para un estilo determinado. Para llegar a alcanzar buenos resultados se requiere disponer, en primer lugar, de una tipología textual y, en segundo, de una enumeración lo más detallada posible de los rasgos lingüísticos que caracterizan a cada tipo de texto o estilo.

En resumen, la corrección automática de un texto en cualquiera de los tres niveles señalados puede mejorar notablemente si se incorporan herramientas como los analizadores morfológicos y sintácticos (véase el apartado 3.1.2), o incluso con la utilización de información fonética. En el caso de la corrección de estilo se requiere, además, una descripción de las características lingüísticas de cada uno de los estilos o tipos de texto que deban tratarse. Por este motivo, en el equipo que desarrolla un programa de ayuda a la escritura es muy aconsejable la presencia de expertos con conocimiento no únicamente de la normativa de la lengua, sino también de todos los niveles de la descripción lingüística, incluyendo el textual.

3.1.2. Herramientas de análisis lingüístico

Como señalábamos anteriormente, existen otras herramientas que no son directamente accesibles al usuario pero que, en cambio, son imprescindible para el desarrollo de muchas de las tecnologías lingüísticas. Se trata de programas que realizan de un modo automático las operaciones que un especialista en morfología y sintaxis conoce a la perfección: extraer la raíz de una palabra, segmentar la palabra en morfemas, asignarles la categoría gramatical correspondiente, determinar la parte de la oración a la que pertenece la palabra, y descomponer una frase en sus constituyentes indicando la función sintáctica de cada uno de ellos.

Nos referiremos en primer lugar a los lematizadores[8], cuya función es detectar el radical de una palabra –es decir, la forma que en los diccionarios aparece como lema-, separándola de los morfemas derivativos o flexivos que la acompañan (Santana *et al*., 1997, 1999). En la figura 5 puede verse, a modo de ejemplo, el resultado de la lematización y del análisis morfológico automático de la forma "canto".

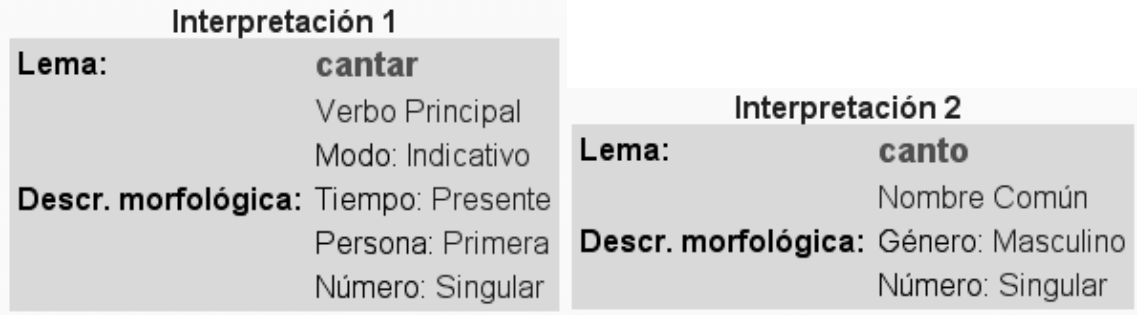

Figura 5: Lematización de la forma "canto" (CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona)

Es importante destacar que, como se discute más adelante (véase el apartado 4.1.2), la ambigüedad constituye un problema muy importante para las diversas herramientas de procesamiento del lenguaje: como se aprecia en el ejemplo de la figura 5, sin disponer de más información, la forma "canto" se analiza automáticamente como nombre o como verbo.

Un lematizador permitiría definir, según indicábamos en 3.1.1, los patrones de corrección gramatical que se aplicarán a un determinado verbo, sin tener que repetirlos para cada una de sus formas conjugadas; facilitaría también la confección de un diccionario para un corrector ortográfico, pues permitiría incorporar al mismo únicamente los lexemas y, con la ayuda de reglas morfológicas, podrían detectarse errores en todas las formas flexionadas de una misma palabra sin necesidad de que todas ellas estuvieran incluidas en el diccionario. En otro orden de cosas, una búsqueda en la web ayudada por un lematizador contribuiría a encontrar una información mucho más rica, pues podrían localizarse páginas que incluyeran no sólo la palabra que se ha introducido en la ventana de búsqueda, sino también las que comparten con ella la misma raíz (véase el apartado 3.2.4).

También para la propia investigación lingüística es útil disponer de un lematizador cuando se trabaja con grandes corpus de textos (descritos en el apartado 4.1.2.): la operación de buscar todas las formas de un verbo en varios

[8] Pueden encontrarse demostraciones del funcionamiento de lematizadores en español en las páginas del CliC, del GEDLC-Grupo de Estructura de Datos y Lingüística Computacional de la Universidad de las Palmas de Gran Canaria y en las de la empresa Signum.

millones de palabras puede ser sumamente larga si debe repetirse para cada persona, tiempo y modo; en cambio, en un corpus lematizado, cada una de las formas que adquiere un verbo en su flexión está asociada al correspondiente infinitivo. De este modo, no sólo se facilita la localización de un verbo en el corpus, sino que también se calculan con facilidad estadísticas de frecuencias de aparición sin tener que recurrir a contar individualmente cada una de las formas conjugadas.

Un analizador morfológico[9] (*Part-of-Speech Tagger* o *POS Tagger*) descompone la palabra en los morfemas que la constituyen y determina la categoría gramatical de cada uno de ellos, así como la categoría léxica de la palabra (Atserias *et al.,* 1998; González *et al*., 1995; van Halteren y Voutilainen, 1999). En la figura 6 se muestra un ejemplo de análisis morfológico automático, acompañado de la correspondiente lematización.

Palabra	Lema	Descr. morfológica
Las	la	**Artículo definido**, género **femenino**, número **plural**.
tecnologías	tecnología	**Nombre común**, género **femenino**, número **plural**.
lingüísticas	lingüístico	**Adjetivo calificativo**, género **femenino**, número **plural**.
mejorarán	mejorar	**Verbo principal**, modo **indicativo**, tiempo **futuro**, persona **tercera**, número **plural**.
las	la	**Artículo definido**, género **femenino**, número **plural**.
economías	economía	**Nombre común**, género **femenino**, número **plural**.

Figura 6: Lematización y análisis morfológico de "Las tecnologías lingüísticas mejorarán las economías" (CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona)

Otra de las operaciones que pueden llevarse a cabo automáticamente es el análisis sintáctico[10] (*Syntactic Parsing*). Tal como se aprecia en la figura 7, con ello se obtiene la estructura de constituyentes de una oración junto con la información sobre su categoría gramatical (Castellón *et al*., 1998; Atserias *et al.,* 1998; Rodríguez, 2002). Este ejemplo muestra también que la ambigüedad de la forma "canto" que encontrábamos anteriormente se resuelve en cuanto se analiza en el contexto de una oración, puesto que de forma automática se ha

[9] Se encuentran demostraciones de analizadores morfológicos en español en las páginas del CliC, del TALP, del gilcUB-Grup d'Investigació en Lingüística Computacional de la Universidad de Barcelona, del GEDLC y del Grupo de Sistemas Inteligentes de la Universidad Politécnica de Madrid.

[10] En las páginas del CliC y del TALP pueden verse también demostraciones de análisis sintáctico en español.

detectado que en su primera aparición es un verbo y en su segunda un nombre común.

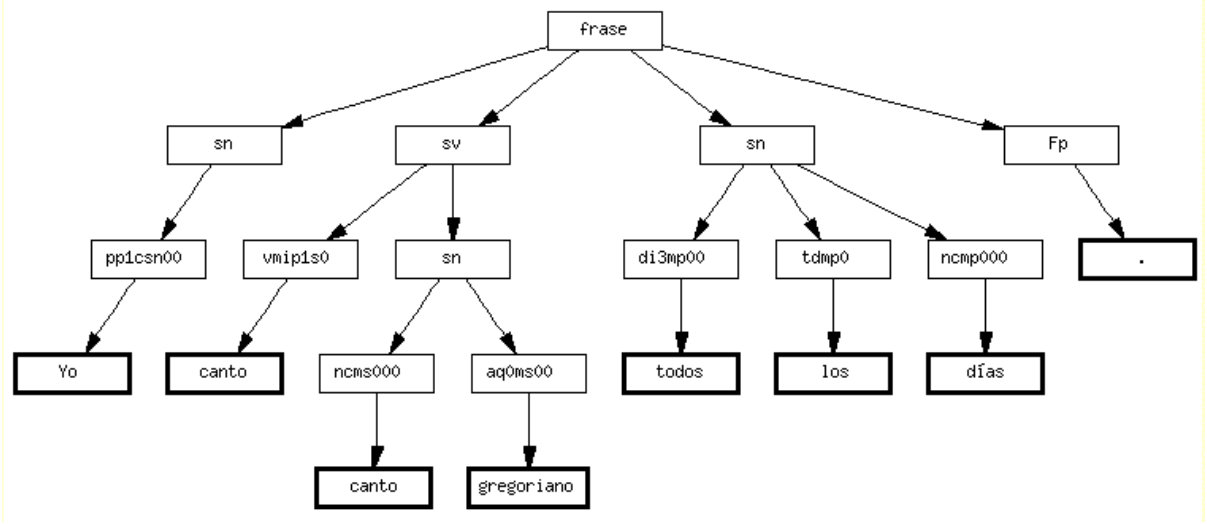

Figura 7: Análisis sintáctico de "Yo canto canto gregoriano todos los días" (CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona)

Con independencia de los algoritmos empleados, las tres herramientas que acabamos de presentar requieren la formalización del conocimiento lingüístico, ya que, de alguna manera, como se planteaba al principio, realizan de forma automática las operaciones de segmentación y análisis que llevaría a cabo un experto en morfología y en sintaxis enfrentado a la misma tarea. La presencia del lingüista es, por tanto, imprescindible en el desarrollo de las herramientas, en una primera fase para definir las "etiquetas", es decir, las categorías morfológicas y léxicas que se asignan en el análisis y, en las etapas sucesivas, para validar y refinar los resultados. Como señalábamos en el caso de la conversión la conversión de texto en habla, la evaluación constituye también aquí una tarea que debería llevar a cabo un especialista, pues del mismo modo que un lingüista difícilmente podrá llegar a valorar los aspectos informáticos de los algoritmos de análisis, tampoco parece lógico que sea el experto en lenguajes de programación quien decida la adecuación lingüística de un análisis morfológico o sintáctico.

Finalmente, cabría hacer referencia a las herramientas de análisis semántico. Como puede suponerse, las dificultades son, en este caso, mucho mayores, en consonancia con los problemas de la representación del significado en el marco de la teoría lingüística. La anotación semántica automática de los textos se lleva a cabo en la actualidad recurriendo a soluciones que podrían calificarse como parciales y muy ligadas a la aplicación. Badia (2001) señala como prácticas realistas la incorporación de rasgos semánticos a los resultados del análisis sintáctico, la integración de la información semántica en el propio análisis sintáctico –el modelo más prometedor, en opinión de este autor– y la

anotación semántica con total independencia de la sintaxis mediante etiquetas estrechamente vinculadas al campo semántico propio de un determinado texto.

3.2. Procesamiento del lenguaje natural

Las herramientas de análisis y las gramáticas (descritas en el apartado 4.3) encuentran su aplicación en los diversos ámbitos que constituyen el procesamiento del lenguaje natural (PLN o NLP, *Natural Language Processing*) (Badia, 2001; Jurafsky y Martin, 2000; Martí y Llisterri, 2002; Moreno *et al*., 1999; Rodríguez, 2000). Centraremos la exposición en algunos ámbitos concretos, con objeto de poner nuevamente de relieve el papel del conocimiento lingüístico en las tecnologías del lenguaje. En los siguientes apartados se abordan, sin entrar en detalles técnicos, la generación (3.2.1) y la comprensión (3.2.2) de textos, la traducción automática (3.2.3), y la recuperación y extracción de información (3.2.4).

3.2.1. Generación del lenguaje

Si en la síntesis del habla el objetivo básico es conseguir la producción de mensajes orales, en la generación del lenguaje (NLG, *Natural Language Generation*) se persigue la creación automática de textos escritos (Bateman y Zock, 2002; Uszkoreit, 1997; Chevreau *et al*., 1999). En este sentido, el módulo de generación de respuestas de un sistema de diálogo (véase el apartado 2.3.) sería una aplicación particular de la técnica que ahora tratamos.

La generación se realiza a partir de una representación abstracta que debe transformarse en un texto bien formado en todos sus aspectos. Como argumenta Badia (2001), el principal problema de la generación reside en que el contenido de una misma representación puede manifestarse en la lengua de diversos modos. Las dificultades son menores cuando la información de la que se parte es únicamente de tipo morfológico, pero aumentan en el momento de generar una oración o un texto a partir de su representación sintáctica o semántica. Esto se debe a que, al igual que un mismo acto de habla puede realizarse de modos muy diversos (véase el apartado 2.3.), la expresión, por ejemplo, de la impersonalidad, puede también llevarse a cabo a través de distintos procedimientos gramaticales.

Por otro lado, la selección del tipo de texto que se crea a partir de una determinada forma abstracta no depende tan solo de factores puramente gramaticales, sino que también intervienen aspectos pragmáticos, criterios relacionados con la tipología textual, y otros condicionantes de tipo sociolingüístico, como el registro. El procedimiento mediante el que el sistema de generación lleve a cabo la elección de la versión final que proporcionará al usuario tiene, naturalmente, una incidencia importante en la naturalidad del resultado (Badia, 2001).

La información de tipo lingüístico es, pues, esencial tanto en el desarrollo como en la evaluación de un sistema de generación de lenguaje, ya que de ella dependen la determinación del tipo de representación abstracta más adecuada y la estrategia para decidir la forma final que ésta va a adoptar.

3.2.2. Comprensión del lenguaje

Cuando los expertos en lingüística computacional se refieren a comprensión del lenguaje (NLU, *Natural Language Understanding*), ciertamente emplean el término "comprensión" de un modo restringido (Allen, 1988; Zaenen, 1997). Un ejemplo lo observábamos al referirnos a la interpretación de las preguntas del usuario por parte de un sistema de diálogo: el proceso, aun con toda su complejidad, se limitaba a extraer una representación del significado casi exclusivamente enfocada a localizar la información necesaria en una base de datos (véase el apartado 2.3). De un modo análogo, la comprensión en el procesamiento del lenguaje natural debe entenderse como la creación, a partir de un texto escrito, de una representación del contenido necesaria para realizar otras operaciones.

En el campo de las tecnologías del habla se investiga en la actualidad sobre la comprensión de la lengua oral (SLU, *Spoken Language Understanding*). Para lograr este objetivo se requiere la integración de un sistema de reconocimiento automático del habla con un procedimiento de comprensión del lenguaje natural, de modo que el reconocedor puede aportar, por ejemplo, información prosódica que no se recoge en el texto escrito y el sistema de comprensión proporciona la información sintáctica y semántica (Colás, 2001; Price, 1997).

Como es lógico, la comprensión depende fundamentalmente de las herramientas de análisis morfológico, sintáctico y, especialmente, semántico que describíamos en el apartado 3.1.2, así como también del desarrollo de una gramática y de un diccionario. No es preciso insistir en que algunos de los problemas más básicos residen en la formalización del conocimiento lingüístico y en la creación de gramáticas y herramientas de análisis cada vez más sofisticadas.

3.2.3. Traducción automática

Como ya se ha señalado, la traducción automática (TA o MT, *Machine Translation*) fue uno de las primeras aplicaciones que se intentaron abordar en el campo del procesamiento informático del lenguaje (Abaitua, 2002a; Alonso, 2001; Amores, 2000; Cerdà, 1995; Trujillo, 2000). Desde sus inicios en los años 50 hasta los actuales sistemas comerciales, la traducción automática ha sufrido una historia llena de altibajos, con momentos de gran optimismo y con épocas en las que la investigación en este campo no gozaba, en ciertos sectores, de mucho prestigio.

Hoy en día, un usuario de los sistemas gratuitos que se ofrecen en Internet puede llegar a pensar que no se han realizado grandes avances; tal opinión no sería del todo injustificada si se observa, por ejemplo, el texto periodístico traducido automáticamente del inglés al español mediante Systran (Figura 8), un sistema utilizado por la Comisión Europea y que se ha incorporado a portales como Altavista o Google.

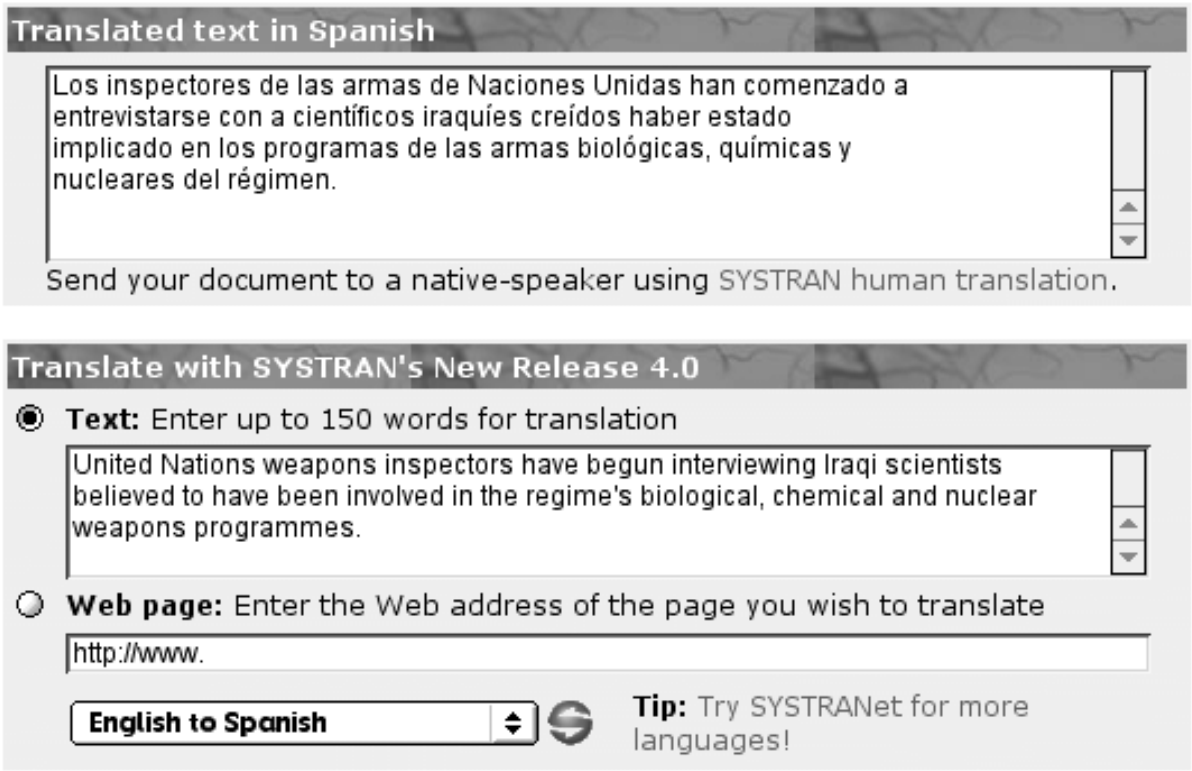

Figura 8: Traducción automática de un texto periodístico mediante el sistema SystranBox (`http://www.systranbox.com/`)

Sin embargo, existen aplicaciones profesionales que permiten obtener buenos resultados con textos especializados en dominios bien delimitados y se han desarrollado, además, sistemas de traducción asistida (TAO o CAT, *Computer Assisted Translation*) que mejoran notablemente la labor del traductor.

Los problemas de la traducción automática son, como pone de relieve Alonso (2001), los propios de la interpretación de enunciados en el lenguaje humano: por un lado, se requiere conocimiento morfológico, sintáctico, léxico y semántico, mientras que, por otro, es imprescindible en ciertos casos lo que se denomina el conocimiento del mundo, información que difícilmente puede formalizarse, por el momento, en un programa informático.

Las estrategias para la traducción automática suelen clasificarse en función de la potencia lingüística del programa, distinguiendo entre la traducción

directa, los sistemas basados en transferencia y los que utilizan la *interlingua*. La traducción directa, que recurre únicamente a léxicos monolingües y bilingües, ofrece, como puede suponerse fácilmente, una calidad muy baja. En cambio, los sistemas que se basan en la transferencia permiten obtener mejores resultados, a costa de una mayor complejidad en el procesamiento.

Como se muestra en la figura 9, en los sistemas que emplean la transferencia, tras la segmentación en frases del texto de entrada en la lengua de origen, se realiza el análisis lingüístico, recurriendo a herramientas de tratamiento morfológico y sintáctico que emplean las reglas de la gramática de análisis y los datos de un léxico monolingüe de la lengua de origen; con ello se crea una representación de la que, en la fase de transferencia, se traduce cada palabra por medio de un léxico bilingüe, teniendo también en cuenta toda la información estructural acumulada durante el análisis. Finalmente, en la fase de generación se convierten los resultados de la transferencia en oraciones gramaticalmente aceptables en la lengua de destino (Alonso, 2001).

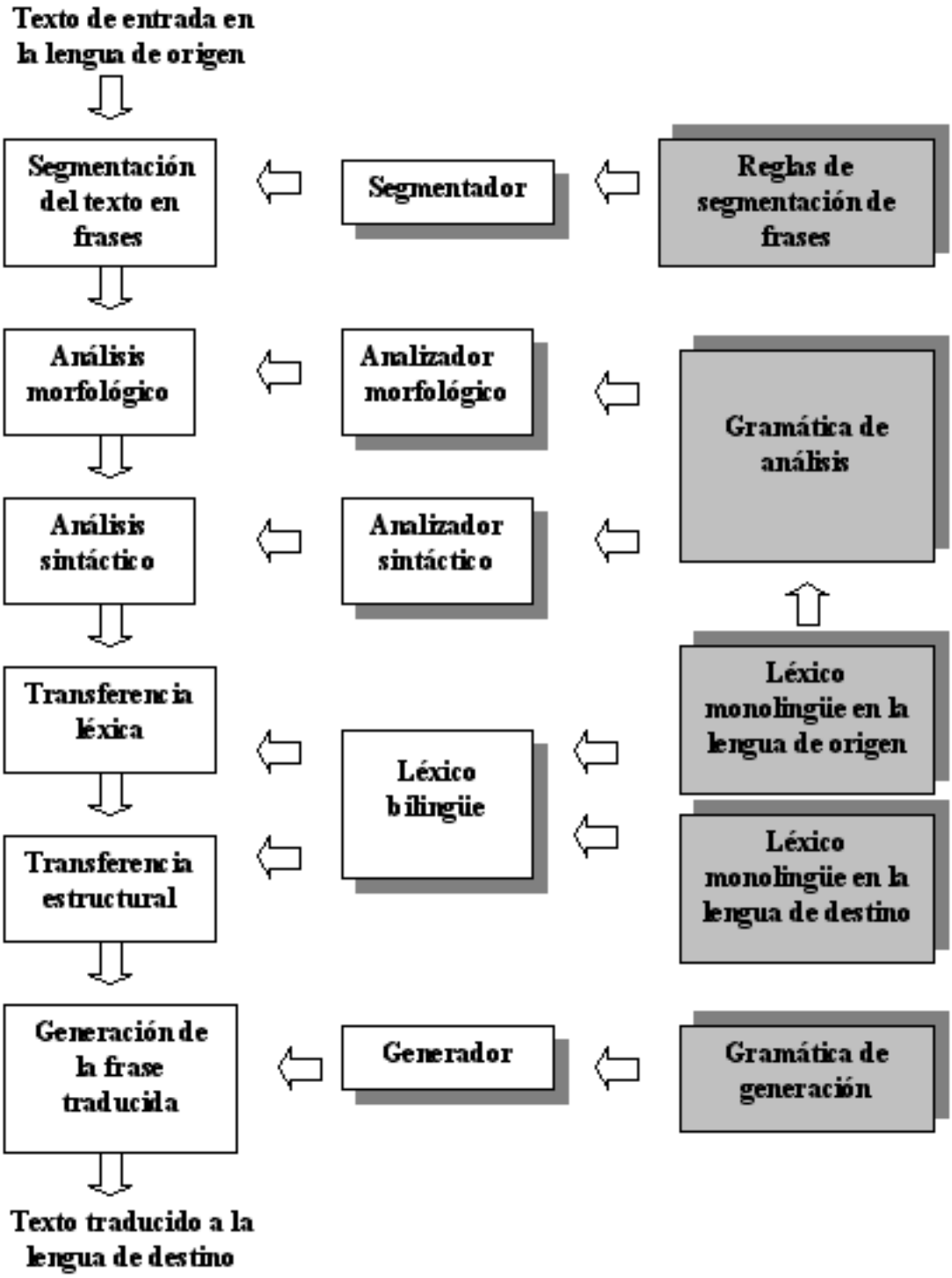

Figura 9: Principales módulos lingüísticos de un sistema de traducción automática basado en la transferencia (Basado en Alonso, 2001)

Los sistemas que utilizan lo que se conoce como *interlingua* emplean para la traducción una representación abstracta del significado, extraída durante la fase de análisis, y que se utiliza como base para la generación. La principal dificultad estriba, por tal motivo, en la representación exhaustiva de los conceptos en términos de rasgos semánticos y de las relaciones que pueden

establecerse entre los mismos. Un traductor automático basado en la *interlingua* puede, por tanto, proporcionar buenos resultados con textos de un ámbito muy restringido, pero presenta aún problemas importantes tanto en el diseño como en la puesta en práctica (Alonso, 2001).

En los últimos años ha surgido también un notable interés por la traducción de la lengua oral (SLT, *Spoken Language Translation*), con el fin de facilitar que dos personas se comuniquen, por ejemplo, en una conversación telefónica, en dos lenguas diferentes sin renunciar cada una a la suya propia (Pastor *et al*., 2001; Wahlster, 2000; Waibel, 2000). Para lograr este objetivo se integra una herramienta de traducción automática de textos en un sistema de diálogo, de modo que el enunciado en la lengua de entrada procesado por el reconocedor de habla se traduce a la lengua de salida y se envía a un conversor de texto en habla.

Uno de las principales obstáculos para esta tarea radica en el propio carácter del habla espontánea, caracterizada por lo que se ha dado en llamar "disfluencias", como son las dudas relacionadas con la planificación del discurso –que se manifiestan en elementos vocales como "eh", "mmm" o en alargamientos vocálicos-, los falsos principios, las repeticiones, y una velocidad de elocución mucho mayor que la que se daría al redactar un texto con un programa de dictado automático. Waibel (1996) ofrece como ejemplo de tales dificultades el resultado de traducir al inglés la transcripción ortográfica literal de una conversación telefónica en español mediante un sistema comercial de traducción automática:

> "...sí sí el viernes diecinueve puedo sí porque sabes me voy de viaje d hoy la verdad así es que este mes es muy viajero me voy el día seis de viaje y estoy hasta el doce así que el día diecinueve me viene muy bien francamente..."
>
> "..yes yes on friday nineteen can yes because know I go me of trip D today the truth such is that this month is very traveler I go me the day six of trip and I am until the twelve as son as the day nineteen comes me very well outspokenly.."

Al margen de los problemas que para un reconocedor presentaría el habla espontánea –y más si se da en un entorno ruidoso como una calle, un aeropuerto, el interior de un vehículo, etc.-, el sistema de traducción debería enfrentarse, como puede observarse, a enunciados gramaticalmente mal formados, al menos desde el punto de vista de la lengua escrita. Tal como sugiere Waibel, la estrategia adecuada en tales situaciones sería intentar recuperar la información semántica y la intención comunicativa del hablante para expresarla de un modo adecuado en la lengua de llegada.

Contrariamente a la traducción automática de textos escritos, la traducción del habla no ha llegado aún al mundo comercial. Existen, sin embargo, prototipos que se centran en dominios restringidos como la información turística, la información sobre vuelos, las reservas de hoteles o el establecimiento de una cita con otra persona[11].

Reproducir automáticamente el proceso de la traducción, en su vertiente escrita y oral es, pues, por lo que acabamos de exponer, un reto tanto para la informática como para la lingüística. El hecho de que los mejores resultados se obtengan en dominios muy específicos constituye un claro indicador de que los principales problemas se encuentran en el tratamiento de los aspectos semánticos y pragmáticos, así como en la formalización, como indicábamos, de ese "conocimiento del mundo" implicado en la comprensión de cualquier texto. La tarea del lingüista en un equipo dedicado a la traducción automática se centra en el contenido de los módulos del sistema que realizan el análisis del texto de entrada y en el desarrollo de los léxicos (véase el apartado 4.2) y de las gramáticas de análisis y de generación (4.3). Por tanto, la morfología, la lexicología, la sintaxis, la semántica y la pragmática encuentran su lugar en la traducción automática de la lengua escrita, así como el análisis de la conversación lo halla en el de la traducción del habla. Al igual que en otras tecnologías, también es esencial la participación de un lingüista en la evaluación de los resultados del sistema, en este caso, en colaboración con traductores profesionales.

3.2.4. Recuperación y extracción de información

La expansión de Internet y el crecimiento de la información acumulada en la web, por una parte, y la digitalización cada vez más habitual de grandes fondos documentales que se almacenan en intranets corporativas, por otra, han creado la necesidad de disponer de un acceso automático a los datos almacenados, puesto que su volumen hace imposible una búsqueda manual. Las técnicas de recuperación y de extracción de información, algunas de las cuales incorporan elementos tomados del procesamiento del lenguaje natural, constituyen una respuesta a este problema y, por tal motivo, son unas de las áreas que más atención reciben en estos momentos en el ámbito que estamos tratando.

La recuperación de información (RI o IR, *Information Retrieval*) consiste en seleccionar, en un conjunto de documentos, aquellos que contienen la

[11] Desde las páginas del Grupo de Reconocimiento de Formas y Tecnologías para el Lenguaje Humano del Instituto Tecnológico de Informática de la Universidad Politécnica de Valencia puede accederse a una demostración de sistemas de traducción automática del habla del español al inglés y del catalán al español.

información que un usuario solicita mediante una consulta (Gonzalo y Verdejo, 2001; Martínez y García, 2002; Spark-Jones, 1999; Verdejo *et al*., 1999). Un ejemplo claro de recuperación de información se encuentra en los buscadores más conocidos de la web, como Altavista o Yahoo, que proporcionan un listado de páginas potencialmente relevantes en función de las palabras utilizadas en la búsqueda[12].

Si bien la recuperación de información ha sido un campo tradicionalmente alejado del procesamiento del lenguaje natural, algunas empresas dedicadas a desarrollar sistemas de búsqueda emplean ya en sus productos comerciales técnicas y herramientas como las mencionadas en los apartados anteriores, y es previsible que su uso se incremente en el futuro (Molano, 2002; Vossen, 2001). Los lematizadores y los analizadores morfológicos hacen posible, como señalábamos en el apartado 3.1.2, encontrar palabras que contengan una determinada raíz y no únicamente las que corresponden a la forma exacta que se ha introducido en la consulta. En otro orden de cosas, el uso de redes semánticas al estilo de WordNet (véase 4.2) facilita la localización de palabras que mantengan una relación semántica con la que se ha empleado en la búsqueda, de modo que, por ejemplo, una consulta con la palabra "coche" podría dar también como resultado documentos que contengan "automóvil"; WordNet es, por otra parte, un recurso que puede contribuir a deshacer la ambigüedad léxica en tareas de recuperación de información, precisando el significado de una palabra ambigua de modo que no aparezcan documentos en los que la palabra no tiene el sentido que inicialmente deseaba el usuario (Ureña, 2002).

Se reconoce también en la actualidad que la introducción de las técnicas de procesamiento del lenguaje natural en el campo de la recuperación de información viene determinada, sobre todo, por la necesidad de tratar documentos en más de una lengua. En la recuperación de información multilingüe (CLIR, *Cross-Language Information Retrieval*) se pretende que el usuario llegue a encontrar los documentos que sean relevantes con independencia de la lengua en la que estén escritos y de la lengua en la que haya realizado su consulta.

Los recursos desarrollados para la traducción automática, como los léxicos y los corpus bilingües, y herramientas como los analizadores morfológicos y sintácticos tienen una función muy relevante en este contexto, ya que no parece

[12] En las páginas de los proyectos ITEM (Recuperación de Información Textual en un Entorno Multilingüe con Técnicas de Lenguaje Natural) y HERMES (Hemerotecas Multilingües: Recuperación Multilingüe y Extracción Semántica) se encuentran demostraciones de sistemas de recuperación de información en español y en entornos multilingües.

factible abordar una tarea de recuperación de información multilingüe sin tener en cuenta los aspectos léxicos, sintácticos y semánticos que en ella están implicados (Gonzalo y Verdejo, 2001).

La digitalización de los archivos de los medios de comunicación orales ha planteado también el problema de recuperar la información recogida en grabaciones que no es factible, por razones económicas y de tiempo, transcribir ortográficamente. En el marco de las tecnologías del habla se ha configurado un área de investigación que, respondiendo a esta necesidad, tiene como objeto conseguir el acceso automático a documentos sonoros (SDR, *Spoken Document Retrieval*), combinando el reconocimiento del habla con las técnicas de recuperación de información textual (Renals y Robinson, 2000).

Mucho más compleja que la recuperación es la extracción de información (IE, *Information Extraction*). La finalidad de la búsqueda, en este caso, no es únicamente seleccionar los documentos relevantes, sino encontrar unos datos determinados en el contenido de un conjunto de documentos y ofrecérselos al usuario de la forma más organizada posible.

Los problemas lingüísticos que se presentan a la hora de realizar automáticamente esta operación son de naturaleza muy diversa. Uno de ellos es el reconocimiento de los nombres propios, ya que éstos se encuentran de diversas formas en los textos -por ejemplo "Lorca", "García Lorca" o "Federico García Lorca"– y el mismo nombre puede referirse a entidades distintas –Lorca es el nombre de un poeta y, a la vez, el de una ciudad-. La correferencia constituye también un obstáculo importante, pues la misma persona puede aparecer en un documento con su apellido, su nombre y apellido o su cargo, y los tres deben identificarse como referentes al mismo individuo. La anáfora es también objeto de atención, dado su papel en la interpretación del contenido de los textos. La extracción de información se lleva a cabo partiendo de un análisis morfológico, léxico y sintáctico de los documentos, y se basa en nociones como entidades, relaciones, o acontecimientos en el marco de un dominio determinado. Con los datos obtenidos se rellenan las denominadas "plantillas", que contienen los campos sobre los que se ha buscado información, proporcionando así el resultado final de todo el proceso al usuario (Gonzalo y Verdejo, 2001).

La recuperación y extracción de información, tanto hablada como escrita, constituyen dos campos en los que las tecnologías lingüísticas se puede aplicar a la solución de los problemas que derivan de la inmensa cantidad de información disponible en formato digital. La incorporación de técnicas de procesamiento del lenguaje y del habla, con el consiguiente tratamiento de las cuestiones lingüísticas básicas que se plantean, hace pensar que las disciplinas dedicadas al

análisis del lenguaje tienen un papel relevante en el marco de estos dos ámbitos en claro proceso de expansión.

4. Recursos lingüísticos

Los recursos lingüísticos (RL o LR, *Language Resources*) son, como ya se ha señalado, un elemento esencial para el desarrollo de las aplicaciones propias de las tecnologías del lenguaje (Cole, 1997b). Habitualmente, se agrupan en tres grandes categorías: corpus (apartado 4.1), diccionarios (apartado 4.2) y gramáticas (apartado 4.3), aunque estas últimas pueden considerarse también herramientas para el análisis o la generación de textos. Al igual que en el resto del trabajo, intentaremos, en los próximos apartados, poner de manifiesto el conocimiento lingüístico implicado en la constitución de recursos, siempre desde la perspectiva de su aplicación al procesamiento del lenguaje y del habla.

4.1. Corpus

Un corpus puede definirse como un conjunto estructurado de textos que constituyen una muestra lo más realista posible del uso de la lengua (Rafel y Soler, 2001; Sinclair, 1996; Torruella y Llisterri, 1999). Es preciso tener en cuenta que cualquier colección de materiales no constituye por sí misma un corpus si no cumple una serie de requisitos, entre los que señalaríamos un diseño coherente, la introducción de marcas en los textos que definan su estructura según unos estándares comúnmente aceptados, y una documentación completa que permita conocer la procedencia y las características de cada uno de los materiales. Para que sea realmente útil en el desarrollo de las aplicaciones que presentamos en este trabajo, un corpus tiene que estar también anotado o etiquetado, es decir, debe incorporar información lingüística adicional que un sistema automático de procesamiento del lenguaje o del habla pueda interpretar y emplear adecuadamente (Garside *et al*., 1997).

En las últimas décadas se ha desarrollado una disciplina, la lingüística de corpus (*corpus linguistics*), dedicada a la constitución y la explotación de los recursos lingüísticos escritos y orales (Biber *et al*., 1998; Kennedy, 1998; McEnnery y Wilson, 1996; Stubbs, 1996). Uno de los principales campos de trabajo de la lingüística de corpus es la aplicación de los recursos a la descripción del uso de la lengua, de modo que ésta se fundamente en materiales reales recogidos para tal fin y no únicamente en las intuiciones del investigador. No nos referiremos aquí a esta orientación, ampliamente documentada en los manuales citados, sino que nos centraremos en el papel de los corpus orales (apartado 4.1.1) y escritos (apartado 4.1.2) en el contexto de las tecnologías del lenguaje.

4.1.1. Corpus orales

En el ámbito de los corpus orales suele hacerse una distinción entre los que recogen la grabación de la señal sonora, denominados corpus orales o bases de datos orales (*speech corpora, speech databases*), y los corpus de lengua oral (*spoken language corpora*), que consisten en transcripciones ortográficas de la lengua hablada (Llisterri, 1996, 1999a). Esto no quiere decir que los corpus orales propiamente dichos no incluyan una representación ortográfica de los datos, ni que en los corpus de lengua oral la grabación original sea inaccesible. Tal división refleja más bien la diferencia entre los primeros recursos creados por los especialistas en tecnologías del habla, centrados, como es lógico, en la onda sonora y en el dominio para el que se diseñaba una determinada aplicación (Draxler, 2000; Gibbon *et al*., 1997; Lamel y Cole, 1997), y los que se recogieron desde una perspectiva más lingüística para el análisis del discurso y de la conversación. Sin embargo, desde hace ya tiempo se recurre a grandes corpus transcritos ortográficamente para el entrenamiento de los modelos lingüísticos de los sistemas de reconocimiento de habla (véase el apartado 2.2), mientras que desde la lingüística se reconoce la necesidad de disponer también de la señal sonora para el estudio de fenómenos sintácticos, semánticos o pragmáticos que se manifiestan fonéticamente mediante los elementos suprasegmentales.

Un corpus oral enfocado al desarrollo de tecnologías del habla responde, por lo general, a objetivos muy concretos, entre los que pueden citarse la extracción de unidades fonéticas para la síntesis, la obtención de conocimiento lingüístico para la conversión de texto en habla (véase el apartado 2.1), el entrenamiento de los modelos acústicos para el reconocimiento (2.2), o el diseño de escenarios para un sistema de diálogo (3.3). En función de su finalidad se establece el diseño del corpus y se definen los distintos niveles de etiquetado del mismo (Gibbon *et al*., 1997; Llisterri, 1999b).

Por ejemplo, un corpus para la síntesis requiere una trascripción fonética y una sincronización de la onda sonora con las fronteras entre segmentos y entre unidades de síntesis, tal como se muestra en la figura 10.

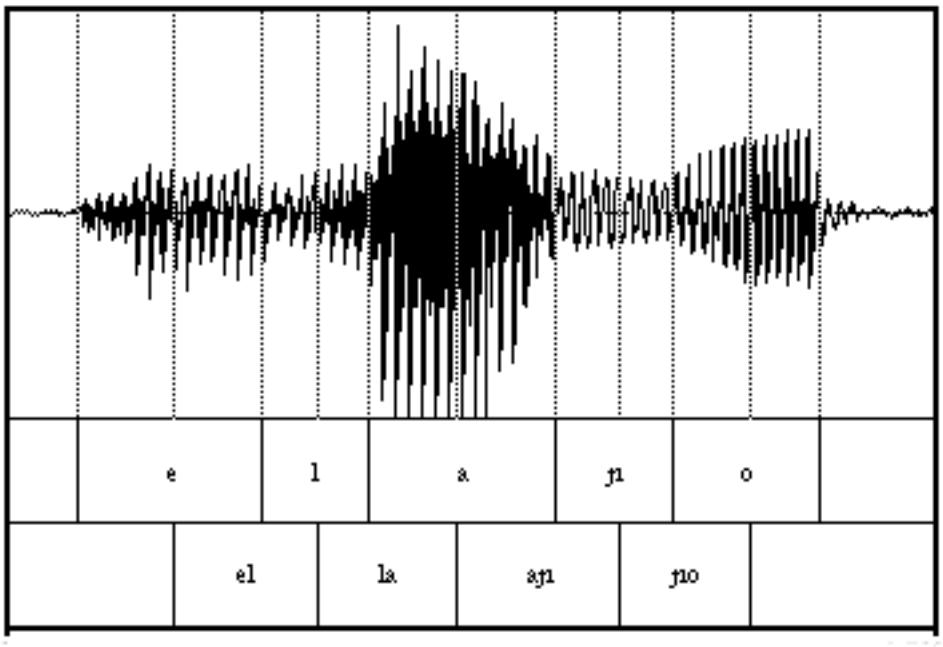

Figura 10: Etiquetado de alófonos y de difonemas en "el año", realizado con el programa Praat (http://www.praat.org).

Un corpus para un sistema de diálogo, en cambio, suele incorporar marcas que definen actos de habla o entidades propias del dominio de la aplicación (Dybkjaer *et al*., 2001; Leech *et al*., 1998; Mengel *et al*., 2000). En la siguiente transcripción de un diálogo entre un usuario y un sistema simulado mediante el protocolo del Mago de Oz (véase el apartado 2.3), se identifican los saludos, las peticiones de información, las confirmaciones que realiza el sistema y los nombres de las estaciones:

[usuario hombre]<saludo>Buenos días</saludo>.

[oz]¿Qué tipo de consulta desea realizar?

[usuario hombre]<solicitar información>Quiero saber cuánto dura el trayecto de <estación de origen>Gràcia</estación de origen> a <estación de destino>Les Planes</estación de destino ></solicitar información>.

[oz]<confirmación explícita>¿Me está solicitando información sobre la duración de un trayecto</confirmación explícita >?

[usuario hombre]<confirmación>Sí</confirmación>.

(Adaptado de Machuca *et al*., 2000)

Estas dos muestras ponen de manifiesto que la información con la que se enriquece un corpus durante el proceso de anotación es, en buena parte, de tipo lingüístico: para el etiquetado de la señal sonora es preciso definir y aplicar criterios de la segmentación, lo que requiere un buen conocimiento de la fonética acústica de la lengua; un corpus de diálogo, por su parte, incluye información de tipo pragmático, que puede, naturalmente, ser mucho más compleja que la que se ofrece en el ejemplo anterior.

También es relevante la participación del lingüista en el momento de diseñar el corpus. Como se mencionaba en apartados anteriores, los corpus utilizados para extraer unidades de síntesis o para entrenar sistemas de reconocimiento se preparan teniendo en cuenta el inventario de unidades fonéticas de la lengua y las posibles combinaciones entre estas unidades, determinadas por reglas fonotácticas. Cuando del corpus se quiere extraer información prosódica para la conversión de texto en habla, el diseño debe ser igualmente cuidadoso, teniendo en cuenta todos los factores que pueden afectar a la duración y la intensidad de un sonido o los que inciden en el patrón melódico de un enunciado.

Por citar un ejemplo en este último ámbito, para el diseño de un corpus prosódico orientado a la conversión de texto en habla (Riera, 2000) se tuvieron en cuenta variables como el número de oraciones por párrafo, la posición de la oración en el párrafo, la complejidad sintáctica, la modalidad y el número de grupos fónicos de cada oración; se prestó igualmente atención a la posición en la frase de los grupos de entonación, el tipo de límite sintáctico que les antecede o precede y el número de sílabas que contienen; finalmente, se consideraron también los grupos acentuales, buscando la representatividad en lo que se refiere al número de sílabas y a su posición en el grupo entonativo. Considerar estos y otros factores implica un buen conocimiento de la estructura prosódica y sintáctica de la lengua y requiere, por tanto, un cierto grado de especialización.

4.1.2. Corpus escritos

Un corpus escrito es una colección estructurada de textos en la que, por lo general, se han introducido marcas que definen su estructura y, en ocasiones, se ha enriquecido con anotaciones relativas a su contenido lingüístico. Para el primer propósito, existen desde hace tiempo estándares como los de la *Text Encoding Initiative* (TEI), basados en el uso del SGML (*Standard Generalized Markup Language*) y, más recientemente, del XML (*Extensible Markup Language*) La información que se añade al texto, en este caso, no es de tipo lingüístico, sino que señala aspectos estructurales como los títulos, subtítulos, la división en párrafos, etc., con un procedimiento análogo al que encontramos en el lenguaje HTML (*Hyper Text Markup Language*) en el que se codifican los

textos que se publican en forma de páginas web (Sperberg-McQueen y Burnard, 2002).

En el campo de las tecnologías lingüísticas, este tipo de codificación es prácticamente imprescindible (Arrarte, 1999), pero también es primordial que los corpus escritos estén adecuadamente anotados. La anotación o etiquetado (Civit *et al*., 2001; Aguirre *et al*., 2001; Sánchez y Nieto, 1995) se realiza en el nivel morfológico, sintáctico, semántico, pragmático o textual, utilizando, cuando están disponibles, herramientas como las descritas en el apartado 3.1.2. y siguiendo unos estándares cada vez más difundidos (Leech y Wilson, 1999; Pérez, 1999; Pino y Santalla, 1996; Vivaldi *et al*., 1996)

Para que un analizador automático realice eficazmente su tarea en el momento de anotar un corpus en cualquiera de los niveles mencionados, se requiere una fase de entrenamiento de la herramienta con un corpus anotado manualmente. Los lingüistas que colaboran con grupos dedicados al procesamiento del lenguaje suelen dedicar mucho tiempo a esta labor, ya que cuanto mayor es el corpus y cuantos menos errores se encuentran en la anotación, mejor se lleva a cabo el entrenamiento de la herramienta y más fiables son los resultados que se obtienen al utilizarla con nuevos materiales. Por otra parte, las anotaciones automáticas requieren a menudo una revisión manual a cargo de un experto para detectar los errores que haya podido cometer la herramienta de análisis. A su vez, esta revisión sirve para mejorar el analizador, refinando sus reglas e incorporando conocimientos que se aplican a la anotación de nuevos corpus.

Uno de los principales problemas que surgen en la anotación morfológica de un corpus es la ambigüedad, que se mencionaba también como una de las dificultades que se plantean en el desarrollo de una herramienta de análisis (véase el 3.1.2). La ambigüedad se encuentra tanto en el caso de una palabra que puede pertenecer a más de una categoría –sería el caso de "joven", nombre o adjetivo-, como en el de palabras que presentan algún rasgo morfológico ambiguo –por ejemplo, "cólera", masculino o femenino- (Civit *et al*., 2002). Para que un analizador pueda tratar estas cuestiones a la hora de anotar automáticamente un corpus real, es necesario recorrer al conocimiento lingüístico experto, sea mediante la desambiguación manual de los casos que no pueden procesarse automáticamente, sea incorporando progresivamente nueva información a la herramienta.

La anotación sintáctica se lleva a cabo actualmente en los llamados *treebanks* o bancos de árboles sintácticos, en los que se marca la categoría y la función de los constituyentes (Civit y Martí, 2002; Leech y Eyes, 1997; Moreno *et al*., 2002). La anotación puede responder a una determinada teoría sintáctica, con lo que el conocimiento lingüístico se refleja directamente en el sistema de

anotación. Desde otras aproximaciones, se prefiere una anotación menos dependiente de un modelo, a veces denominada anotación neutra, de modo que el corpus sea útil para investigaciones de naturaleza más general. En ambos casos, es preciso también considerar los distintos grados de detalle con que se marcará la estructura de los constituyentes, que puede ir desde un análisis muy superficial (*shallow parsing*) a uno mucho más profundo (*deep parsing*). Al igual que en la anotación morfológica, en una primera etapa se requiere la corrección manual por parte de un especialista con objeto de mejorar gradualmente el sistema.

Consideraciones similares podrían hacerse sobre la anotación semántica y pragmática de corpus: llegar a desarrollar un sistema automático exige, por una parte, el conocimiento lingüístico necesario para definir las categorías y las etiquetas que se emplearán en la anotación y, por otra, disponer previamente de un corpus etiquetado manualmente y validado por un experto.

Un tipo especial de corpus escrito lo constituyen los llamados corpus paralelos (Abaitua, 2002b; de Yzaguirre *et al*., 2000), que contienen el mismo texto en dos o más lenguas y se emplean especialmente en el desarrollo de los sistemas de traducción automática. Para ello, se lleva a cabo previamente un proceso conocido como alineación, que hace corresponder un segmento en una lengua –una frase, un párrafo o el texto que se haya establecido como una unidad de traducción– con el segmento equivalente en la otra. La existencia de estos recursos ha permitido abordar la traducción automática entrenando los sistemas con corpus paralelos alineados, recurriendo a técnicas estadísticas similares a las utilizadas en el reconocimiento del habla. Tal aproximación contrasta con las estrategias basadas en reglas que se exponían en el apartado 3.2.3, pues la traducción se realiza en este caso basándose en los ejemplos que aparecen en el corpus y, aplicando, en ciertos sistemas, procedimientos basados en la analogía (Abaitua, 2002).

En relación con los corpus paralelos podrían mencionarse las memorias de traducción, que se han incorporado a las herramientas comerciales de más éxito en el ámbito de la traducción asistida (Abaitua, 2002). Tales memorias son el resultado de almacenar sistemáticamente en su versión final los textos con los que trabaja un traductor, de modo que cuando éste debe traducir un texto nuevo, sea posible reutilizar las traducciones guardadas de textos similares; en este sentido, una memoria de traducción constituye también un recurso lingüístico.

4.2. Recursos léxicos

Los recursos léxicos más utilizados en el campo de las tecnologías lingüísticas son los léxicos computacionales, monolingües o multilingües (Atkins *et al*., 2002; Leci *et al*., 2000), y las llamadas redes léxico-semánticas.

Un léxico computacional (Vázquez *et al.* 2002; Villegas *et al.*, 1998) difiere en muchos aspectos de un diccionario clásico, pues más que la definición de la palabra contiene la información morfológica, sintáctica y semántica relevante para las diversas aplicaciones del procesamiento del lenguaje, para su incorporación a las herramientas de análisis automático (descritas en 3.1.2) y para la anotación de corpus textuales (4.1.2).

En el caso del verbo "ir", por ejemplo, un léxico computacional contendría, entre otras, las siguientes informaciones (Llisterri y Martí, 2002):

<ir, [[-N], [+V], [SUBCAT <SP>]]

Se indica así que se trata de un verbo y no de un nombre, y que subcategoriza un sintagma preposicional Si se tratara de un léxico multilingüe, cada lema llevaría asociado su correspondencia con los equivalentes en otras lenguas, estableciendo también, si se emplea en traducción automática, las restricciones necesarias para lograr una buena traducción.

Se han desarrollado, para su aplicación a las tecnologías del habla, léxicos que contienen la transcripción fonética de las entradas, conocidos como léxicos de pronunciación (*pronunciation lexica*). Una de sus funciones es establecer, para cada palabra de un corpus, sus variantes de pronunciación y asociarlas a una forma canónica, por lo que en algunos casos se incluye también una representación fonológica más abstracta. Este tipo de léxicos se emplea, por ejemplo, para las excepciones a las reglas de un programa de transcripción fonética automática (2.1.1) o en los sistemas de reconocimiento del habla (2.2) (Adda-Decker y Lamel, 2000; Quazza y van den Heuvel, 2000).

Las redes léxico-semánticas estructuran el vocabulario en función de las relaciones semánticas entre palabras, basándose en conceptos propios de la semántica léxica como sinonimia, antonimia, hiponimia, hiperonimia o meronimia ("parte de"). El recurso más conocido en este ámbito es el *WordNet* desarrollado en la Universidad de Princeton (Fellbaum, 1998), del que se dispone de una versión en distintas lenguas europeas, conectadas entre sí, denominada *EuroWordNet* (Vossen, 1998)[13].

Como ilustración de la información recogida en una red léxico-semántica, en las figuras 12 y 13 se muestran los hipónimos y los hiperónimos de "silla" que aparecen en la versión española de *WordNet*

[13] *EuroWordnet* en español puede consultarse desde las páginas del CliC y del TALP.

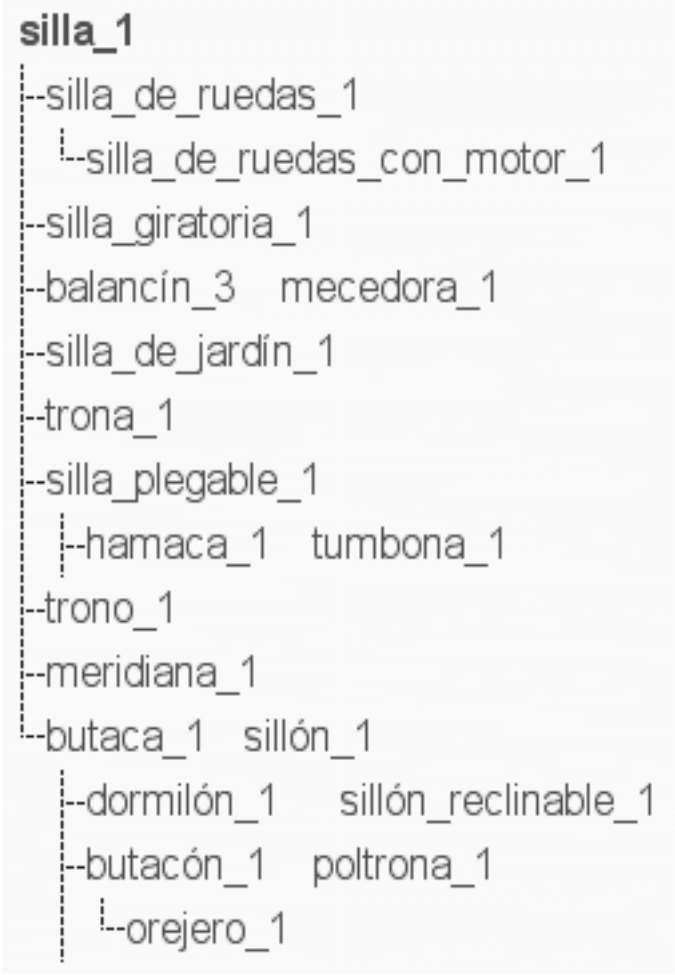

Figura 12: Hipónimos de "silla" en la versión española de *WordNet* (CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona)

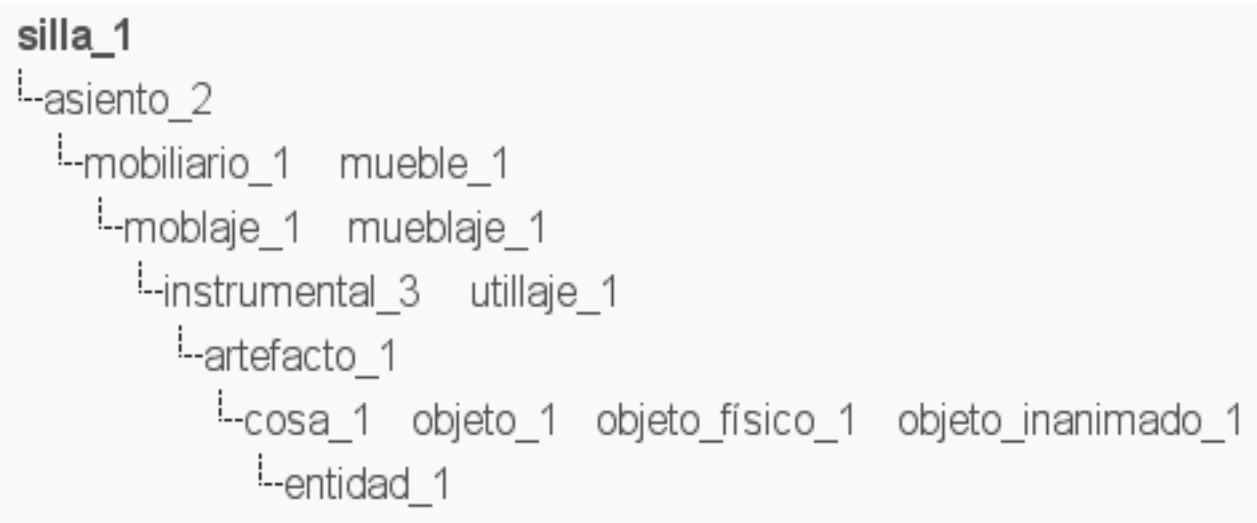

Figura 13: Hiperónimos de "silla" en la versión española de *WordNet* (CLiC, Centre de Llenguatge i Computació, Universitat de Barcelona)

Las redes léxico-semánticas constituyen hoy día un recurso ampliamente utilizado en la anotación semántica de corpus, y muestran también un gran potencial, como ya se ha señalado en el apartado 3.2.4, para las aplicaciones orientadas a la recuperación y extracción de información (Vossen, 2001).

Los recursos al estilo de *WordNet* podrían considerarse también ontologías (Feliu *et al*., 2002), teniendo en cuenta que establecen una organización de los conceptos, especialmente en el caso de los nombres. Dada la dificultad de reflejar la totalidad de la organización conceptual de una lengua, en el procesamiento de textos restringidos a un determinado dominio o campo del saber se recurre a recursos terminológicos, monolingües o multilingües, extraídos de corpus especializados (Bach *et al*., 1997; Cabré y Feliu, 2001).

Observamos, pues, que tanto los léxicos computacionales como las redes léxico-semánticas conllevan una formalización y una estructuración del conocimiento léxico, morfosintáctico, semántico y, en ciertos casos, fonético y fonológico. En este sentido, tanto en su diseño como en su realización y revisión deben intervenir los lingüistas para garantizar una coherencia interna y un nivel de calidad que haga posible el uso efectivo de estos recursos.

4.3. Gramáticas

Una gramática computacional puede entenderse como una descripción formalizada del conocimiento lingüístico que, en el marco del procesamiento del lenguaje, puede ser empleada tanto por las herramientas de análisis como por las de generación de textos. Por esta razón se considera un recurso, junto a los corpus y a los léxicos, para el desarrollo de sistemas que realicen las operaciones que hemos descritos anteriormente.

Una gramática requiere un formalismo para expresar la información lingüística que contiene. Si en una primera etapa se utilizaron las gramáticas de estructura sintagmática –o gramáticas de estructura de frase-, el procesamiento sintáctico en los últimos años se he centrado en las gramáticas de unificación y en la gramática de restricciones (*Constraint Grammar*). Las primeras reciben este nombre por el procedimiento que se aplica para combinar la información contenida en las categorías gramaticales, y tienen como principal característica la codificación de la máxima información posible en el léxico, al que se incorporan rasgos sintácticos y semánticos. Las gramáticas de restricciones parten de la anotación de la posibles funciones sintácticas de una palabra, para realizar después una desambiguación y seleccionar la función adecuada en una oración concreta (Badia, 2001; Balari, 1999; Rodríguez, 2002a).

Existen también aproximaciones, como la de la sintaxis léxica, que integran gramáticas y diccionarios electrónicos. Los diccionarios contienen, para cada forma, el lema a la que está asociada, la clase distribucional a la que

pertenece y sus propiedades morfológicas. Las gramáticas consisten, en este marco teórico, en una formalización de las propiedades de los predicados que se encuentran en el diccionario (Subirats y Ortega, 2000).

No hace falta insistir de nuevo en que la construcción de una gramática computacional es una tarea imposible de abordar sin un detallado conocimiento de la estructura morfológica, sintáctica y semántica de la lengua. Cabe destacar también que los formalismos de unificación están teniendo una cierta repercusión en la teoría sintáctica como alternativas a otros modelos, con lo que se establece un diálogo entre los planteamientos propios de la lingüística computacional y los de la teoría lingüística (Badia, 2001; Balari, 2000).

5. Las "nuevas profesiones"

Hace ya más de diez años, Cabré y Payrató (1990) se referían a las "nuevas profesiones" de los lingüistas en la recopilación de un ciclo de conferencias sobre Lingüística Aplicada que contenía, además, contribuciones sobre la traducción automática y las interfaces en lenguaje natural, áreas que, como acabamos de ver, son una parte esencial de las actuales tecnologías del lenguaje. En la misma línea que hemos intentado mantener a lo largo de este trabajo, Cabré y Payrató incluyen el procesamiento del lenguaje entre aquellos campos en los que "una «presencia lingüística» es, por definición, inexcusable" (p. 21), lo que les lleva a argumentar que las necesidades relacionadas con el tratamiento de la información abren nuevas oportunidades profesionales para el especialista con una formación adecuada.

Más recientemente, en noviembre de 2002, se celebró en Pavía y en Lugano un congreso titulado "Linguistics and the New Professions"[14], con objeto de reflexionar sobre el papel profesional del lingüista en el campo de las tecnologías de la información y de las comunicaciones. Entre los temas que figuraban en el programa se encontraban la extracción de información, la gestión documental y del conocimiento, la web semántica, la traducción automática, el papel de la pragmática y del análisis del diálogo para modelar el comportamiento del usuario, o la utilidad de la lingüística en los sistemas de comunicación persona-máquina y de comunicación persona-persona mediatizada por ordenador. Como puede observarse, casi todos estos ámbitos de trabajo se relacionan con diversos aspectos de las tecnologías lingüísticas, tanto con el procesamiento del lenguaje natural –extracción y gestión de información, traducción automática– como con las tecnologías del habla –comunicación

[14] Organizado conjuntamente por el Departamento de Lingüística de la Universidad de Pavía (Italia) y el Instituto de Lingüística y Semiótica de la Facultad de Ciencias de la Comunicación de la Universidad de Lugano (Suiza).

persona-máquina-, a la vez que con disciplinas lingüísticas como la semántica o la pragmática.

Estos dos hechos no son, ciertamente, anécdotas aisladas, sino que reflejan la tendencia general en las últimas décadas a considerar, al menos en ciertos sectores, que el lingüista no es únicamente, como se mencionaba al principio, un profesor de "gramática" (Payrató, 1997). Como hemos procurado mostrar, el desarrollo de las tecnologías del lenguaje no puede llevarse a cabo sin un importante bagaje de conocimientos lingüísticos; de ahí que la necesidad de aplicar estas tecnologías a ámbitos como la recuperación y extracción de información, la traducción automática o el acceso telefónico a servicios de información y de comercio electrónico permita pensar que los lingüistas pueden ubicarse profesionalmente realizando tareas relacionadas con las tecnologías del lenguaje.

En realidad, existen ya desde hace tiempo equipos de investigación y desarrollo integrados por especialistas en informática -en general, procedentes de la inteligencia artificial- o en ingeniería de telecomunicaciones -dedicados, primordialmente, al tratamiento digital de señales- y por lingüistas que han orientado su trabajo hacia el procesamiento del lenguaje o las tecnologías del habla.

Aun así, la alternativa más habitual en estos momentos, al menos en nuestro contexto más cercano, son los proyectos conjuntos entre equipos con perfiles bien diferenciados, en los que unos desarrollan las aplicaciones informáticas y otros proporcionan el conocimiento lingüístico necesario. En el contexto español[15] diversos grupos en departamentos de Lingüística o de Filología colaboran regularmente con otros procedentes del área de la informática o del tratamiento de señales, situación que se refleja, por ejemplo, en varias de las publicaciones que se citan en el presente trabajo.

Fuera ya del ámbito académico, determinadas empresas del sector de las tecnologías del lenguaje cuentan con lingüistas para la creación y la validación de herramientas y recursos. Este fue el caso, en su momento, de IBM (Madrid) o de Incyta (Barcelona), y lo es ahora en empresas relativamente jóvenes o en otras ya bien establecidas. Cabe contar, además, con las editoriales especializadas en obras de referencia que incorporan lexicógrafos computacionales a sus equipos y con las empresas de traducción que cuentan con lingüistas familiarizados con el uso de herramientas informáticas entre sus colaboradores.

[15] Llisterri y Garrido (1998) y Llisterri (1999c) presentan panorámicas generales de la situación de la ingeniería lingüística en España en el momento de publicación de los trabajos.

Para cerrar esta breve panorámica, es importante señalar la reciente creación de empresas de tecnologías lingüísticas como resultado de la actividad de grupos universitarios dedicados al procesamiento del lenguaje y del habla. Atlas-*Applied Technologies on Language and Speech* (Barcelona) en el campo de las tecnologías del habla y Daedalus–*Data, Decisions and Language* (Madrid), Thera (Barcelona) o Eleka (San Sebastián), centradas en el tratamiento de la lengua escrita, son algunos ejemplos. La presencia de estas empresas abre también oportunidades para la incorporación de lingüistas al mercado laboral.

6. La formación

La actual formación básica de un lingüista suele realizarse en España en el marco de una licenciatura en Filología y, en algunas universidades, en la licenciatura de segundo ciclo de Lingüística.

Las directrices generales que establecen el plan de estudios de esta especialidad incluyen como troncal una asignatura de Lingüística Computacional de seis créditos, para la que se encuentra el siguiente descriptor:

> "Procesamiento automático del lenguaje. Criterios formales de modelización lingüística. Reconocimiento automático del habla".

No es difícil percatarse de que existe una cierta disparidad entre los temas que se proponen y la descripción de las tecnologías lingüísticas que hemos realizado a lo largo de este trabajo: por un lado, entre las tecnologías del habla sólo se contempla el reconocimiento, sin que la síntesis o los sistemas de diálogo encuentren su lugar en los estudios; por otro, "procesamiento automático del lenguaje" podría considerarse equivalente a "procesamiento del lenguaje natural", en el sentido que se le da habitualmente (véase el apartado 3.2), o podría incluir también el tratamiento del habla si abarca el lenguaje en todas sus manifestaciones, con lo que sería redundante mencionar el reconocimiento. Las áreas a las que se asigna la docencia de esta asignatura son "Ciencias de la Computación e Inteligencia Artificial", "Lenguajes y Sistemas Informáticos" y, naturalmente, "Lingüística General"; en cambio, el reconocimiento del habla es uno de los campos de trabajo del área de Teoría de la Señal y Comunicaciones, al menos si atendemos a la práctica habitual en nuestro país.

El examen de esta "realidad oficial" que aparece en el BOE no lleva, pues, a un exceso optimismo en lo que se refiere a la definición de la materia. Es muy positivo, en cambio, desde nuestra perspectiva, que la Lingüística

Computacional aparezca separada de la Lingüística Aplicada, dotándole de una identidad específica[16].

Al margen de la formación inicial, profesores y alumnos reconocen que una licenciatura no proporciona sino unas bases generales, y que la preparación más orientada a una actividad profesional concreta suele adquirirse en cursos de postgrado. En el campo que nos ocupa han existido y existen aún algunas iniciativas orientadas a ofrecer una formación en tecnologías lingüísticas en cursos de postgrado, y varios programas de doctorado o de máster ofrecen asignaturas relacionadas con el procesamiento del lenguaje y del habla; también universidades como la Universidad Politécnica de Cataluña participan en el Máster Europeo en Lenguaje y Habla auspiciado por ELSNET (*European Network of Excellence in Human Language Technologies*), abriendo así una vía de integración en un mercado más amplio a los estudiantes.

Estos cursos permiten, efectivamente, obtener algunos de los conocimientos necesarios para incorporarse a equipos de trabajo implicados en el desarrollo de tecnologías o recursos lingüísticos. Aun así, no resuelven los problemas que se derivan del modo en que se enfoca la formación que recibe un futuro lingüista. Por lo general, en las licenciaturas en Filología y en Lingüística no se ofrece una educación que ayude a comprender el funcionamiento básico de las tecnologías del lenguaje; no se trata, obviamente, de convertir al lingüista en un experto en programación o en procesamiento digital de señales, pero es necesario dotarle de la información esencial sobre los sistemas informáticos para que pueda dialogar –y, en su momento, colaborar- con un especialista en estas materias. En un plano más aplicado, en la práctica profesional o en la investigación en tecnologías lingüísticas es del todo imprescindible el manejo sin dificultades de aplicaciones como editores de textos y bases de datos, la familiaridad con los procedimientos de búsqueda de información en Internet y un conocimiento básico de los entornos informáticos en los que se encuentran muchas de las herramientas que se usan en procesamiento del lenguaje y del habla.

Desde otro punto de vista, la integración en un equipo que desarrolla tecnologías lingüísticas exige un cierto cambio de perspectiva. El conocimiento lingüístico no puede, por lo general, integrarse en una aplicación del mismo modo en que se presenta en las descripciones de la lengua que se recogen en una gramática o en una monografía especializada. En ocasiones, estas descripciones no son lo suficientemente explícitas; en otras, no son lo bastante

[16] En este sentido la Lingüística Computacional puede considerarse una disciplina con mayor fortuna que la Fonética y la Fonología, que se reúnen en una única asignatura, mientras que, en cambio, la Morfología, la Sintaxis y la Semántica mantienen su individualidad.

exhaustivas, pues consisten más bien en la ilustración de un modelo o de una teoría. Justamente por tal motivo se recurre con frecuencia al uso de corpus (véase el apartado 4.1). Una formación filológica tradicional no siempre prepara necesariamente para pasar del conocimiento de la teoría gramatical al desarrollo de un analizador morfológico o sintáctico, para convertir una descripción prosódica en un conjunto de patrones aptos para un conversor de texto en habla, o para aplicar el análisis de la conversación a los problemas de la interacción oral entre personas y sistemas informáticos, por mencionar únicamente tres ejemplos. Es necesario, pues, introducir en los estudios los procedimientos para adquirir la capacidad de organización, de estructuración y de formalización de datos lingüísticos que requiere la creación de aplicaciones y de recursos (Llisterri y Martí, 2002).

Por otra parte, algunas de las herramientas que hemos mencionado constituyen, adecuadamente empleadas, una excelente ayuda para el estudio de la lengua. El trabajo con un sintetizador por formantes, por ejemplo, ayuda a comprender las relaciones entre las propiedades acústicas y las características perceptivas de los sonidos; el diseño y la aplicación de reglas de transcripción fonética automática sirve para sistematizar el conocimiento de los procesos fonéticos y fonológicos; los errores de un sistema de dictado automático llevan a interesantes debates sobre la percepción del habla y sobre la interacción entre información fonética e información gramatical; la observación de los resultados que proporciona un analizador morfológico o sintáctico es una buena estrategia para explicar determinados aspectos de la morfología o la sintaxis de las lenguas, y la formalización en reglas explícitas de una pequeña parte de la gramática de una lengua ayuda a profundizar en la descripción; en el campo de la semántica léxica, las redes léxico-semánticas aportan datos sobre los que basar una discusión más teórica; finalmente, en lo que se refiere a la descripción lingüística, un corpus textual o de transcripciones de la lengua oral constituye hoy un recurso casi imprescindible para validar las hipótesis. Este uso de las herramientas y los recursos propios de las tecnologías lingüísticas no pretende, claro está, formar lingüistas computacionales, pero con independencia de su valor didáctico, supone al menos una primera toma de contacto con el tratamiento informático del lenguaje que puede ayudar a entender mejor el porqué de las limitaciones de algunas tecnologías actuales o el potencial de las futuras.

Las carencias en el conocimiento tecnológico y en el modo de aplicar el conocimiento lingüístico tienen como consecuencia que, a menudo, los lingüistas que logran entrar en el mundo de las tecnologías del lenguaje lo hagan en calidad de proveedores de datos o de revisores de la información obtenida por procedimientos automáticos, sin que tengan una intervención real en la concepción y el desarrollo del proyecto. Tal situación, que contrasta con la

inevitable necesidad de conocimiento lingüístico para el procesamiento del lenguaje y del habla, tiene, además, una cierta repercusión económica, pues cuando el lingüista es un "asesor" más que un "creador", la valoración de su trabajo en el mercado no siempre es equivalente a la de quienes desarrollan los recursos informáticos necesarios para el funcionamiento una aplicación.

6. Conclusión

A lo largo de este trabajo se ha intentado poner de manifiesto que el conocimiento lingüístico constituye una base indispensable en el diseño y el desarrollo de herramientas, aplicaciones y recursos en el campo de las tecnologías del lenguaje. Este hecho debería propiciar la incorporación de los lingüistas a las nuevas profesiones que surgen como consecuencia de la expansión de la Sociedad de la Información y del Conocimiento. Sin embargo, tal integración no se realizará en condiciones óptimas sin un profundo replanteamiento de la formación que actualmente se imparte. Por ello, se ha procurado mostrar, en varios de los ámbitos que configuran el procesamiento del lenguaje y del habla, cuáles son las tareas que competen al lingüista y qué tipo de conocimiento es necesario para realizarlas. Se trata, sin duda, de una primera reflexión, inevitablemente ni exhaustiva ni definitiva, pero que, en todo caso, pretende contribuir al debate sobre el futuro de quienes desean convertir en una profesión su interés por el estudio del lenguaje.

* * * * * * *

Referencias bibliográficas

ABAITUA, J. (1999) "Quince años de traducción automática en España", *Perspectives: Studies in Translatology* 7, 2: 221-1230.
http://sirio.deusto.es/abaitua/konzeptu/ta/ta15.htm

ABAITUA, J. (2002a) *Introducción a la traducción automática.* Grupo DELI, Universidad de Deusto.
http://sirio.deusto.es/abaitua/konzeptu/ta/mt10h_es/index.html

ABAITUA, J. (2002b) "Tratamiento de corpora bilingües", in MARTÍ, M.A.- LLISTERRI, J. (Eds.) *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita.* Barcelona: Fundación Duques de Soria - Edicions Universitat de Barcelona (Manuals UB, 53). pp. 60-90.

ADDA-DECKER, M.- LAMEL, L. (2000) "The use of lexica in automatic speech recognition", in VAN EYNDE, F.- GIBBON, D. (Eds.) *Lexicon Development for Speech and Language Processing*. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 12). pp. 235-266.

AGUILAR, L.- FERNÁNDEZ, J.M.- GARRIDO, J.M.- LLISTERRI, J.- MACARRÓN, A.- MONZÓN, L.- RODRÍGUEZ, M.A. (1994) "Diseño de pruebas para la evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla", in *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Córdoba, 20-22 de julio de 1994.

http://liceu.uab.es/~joaquim/publicacions/cordoba_94.html

AGUIRRE, J.L.- ANDIÓN, N.- GÓMEZ, J. (2001) "Aspectos ortográficos, léxicos y morfosintácticos del etiquetado lingüístico de un corpus de informática en lengua gallega", *Procesamiento del Lenguaje Natural, Revista* nº. 27: 13-20.

ALONSO, J.A. (2001) "La traducció automàtica", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 86-119.

ALLEN, J. (1988) *Natural Language Understanding.* Redwood City, CA: Benjamin / Cummings. Second Edition, 1995.

AMORES, J.G. (2000) "Sistemas de traducción automática", *Quark. Ciencia, Medicina, Comunicación y Cultura* 19: 46 - 52.
http://www.imim.es/quark/num19/019046.htm

ARRARTE, G. (1999) "Normas y estándares para la codificación de textos y para la ingeniería lingüística”, in 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. 17-44.

ATKINS, S.- BEL, N.- BERTAGNA, F.- BOUILLON, P.- CALZOLARI, N.- FELLBAUM, C.- GRISHMAN, R.- LENCI, A.- MacLEOD, C.- PALMER, M.- THURMAIR, G.- VILLEGAS, M.- ZAMPOLLI, A. (2002) “From Resources to Applications. Designing the Multilingual ISLE Lexical Entry”, in *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas de Gran Canaria, 2002.
http://www.ub.es/gilcub/lascosas/pubYreps/isle.pdf

ATSERIAS, J.- CARMONA, J.- CASTELLÓN, I.- CERVELL, S.- CIVIT, M.- MÀRQUEZ, L.- MARTÍ, M.A.- PADRÓ, L.- PLACER, R.- RODRÍGUEZ, H.- TAULÉ, M.- TURMO, J. (1998) "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text", in RUBIO, A.- GALLARDO, N.- CASTRO, R.- TEJADA, A. (Eds.) *Proceedings of the First International Conference on Language Resources and Evaluation*. May 28 - 30, 1998, Granada, Spain. Vol. II. pp. 1267-1271.
http://www.lsi.upc.es/~nlp/papers/1998/lrec98-a.al.ps.gz

BACH, C.- SAURÍ, R.- VIVALDI, J.- CABRÉ, M.T. (1997) *El corpus de l'IULA: descripció.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (Papers de l'IULA, Sèrie Informes, 17).
ftp://ftp.iula.upf.es/pub/publicacions/97inf017.pdf

BADIA, T. (2001) "Tècniques de processament del llenguatge", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 189-238.

BALARI, S. (1999) "Formalismos gramaticales de unificación y procesamiento basado en restricciones", in GÓMEZ GUINOVART, J.- LORENZO SUÁREZ, A.- PÉREZ GUERRA, J.- ÁLVAREZ LUGRÍS, A. (Eds.) Panorama de la investigación en lingüística informática. *RESLA, Revista Española de Lingüística Aplicada*, Volumen monográfico. pp. 117-152.

BALARI, S. (2000) "El procesamiento del lenguaje natural como problema computacional", *Quark. Ciencia, Medicina, Comunicación y Cultura* 19: 35-43.
http://www.imim.es/quark/num19/019035.htm

BATEMAN, J.- ZOCK, M. (2002) *Natural Language Generation Systems*. English Department, University of Bremen.
http://www.fb10.uni-bremen.de/anglistik/langpro/NLG-table/NLG-table-root.htm

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

BIBER, D.- CONRAD, S.- REPPEN, R. (1998) *Corpus Linguistics. Investigating Language Structure and Use.* Cambridge: Cambridge University Press (Cambridge Approaches to Linguistics).

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", *Jornadas en Tecnologías del Habla*, Universidad de Sevilla, Sevilla, Noviembre de 2000.
http://gps-tsc.upc.es/veu/basurde/download/Bon00a_sevilla.pdf

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

CABRÉ, M.T.- FELIU, J. (Eds.) (2001) *La terminología científico-técnica: reconocimiento, análisis y extracción de información formal y semántica (DGES PB96-0293)*. Barcelona: Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra (Sèrie Materials).

CABRÉ, M.T.- PAYRATÓ, L. (1990) "La lingüística aplicada avui", in *La lingüística aplicada. Noves perspectives- Noves professions-Noves orientaciones*. Barcelona: Fundació Caixa de Pensions / Universitat de Barcelona. pp. 11-30.

CASTELLÓN, I.- CIVIT, M.- ATSERIAS, J. (1998) "Syntactic Parsing of Unrestricted Spanish Text", in RUBIO, A.- GALLARDO, N.- CASTRO, R.- TEJADA, A. (Eds.) *Proceedings of the First International Conference on Language Resources and Evaluation*. May 28 - 30, 1998, Granada, Spain. Vol. I. pp. 603-610.
http://www.lsi.upc.es/~civit/PUBLICA/elra.doc.gz

CERDÀ, R. (1995) *Perspectivas en traducción automática*. LynX, Documentos de Trabajo, vol 2.

CIVIT, M.- CASTELLÓN, I.- MARTÍ, M.A. (2001) "Creación, etiquetación y desambiguación de un corpus de referencia del español", *Procesamiento del Lenguaje Natural), Revista* nº. 27: 21-28.
http://www.lsi.upc.es/~civit/PUBLICA/Cedcre-def.ps.gz

CIVIT, M. – CASTELLÓN, I. – MARTÍ, M.A. (2002) "*Joven periodista triste busca casa frente al mar* o la ambigüedad en la anotación de corpus", in LUQUE, J.D.- PAMIES, A.- MANJÓN, F.J. (Eds.) *Nuevas tendencias en la investigación lingüística. Actas del Congreso Internacional sobre Nuevas Tendencias de la Lingüística*. Granada: Universidad de Granada (Granada Lingvistica).
http://www.lsi.upc.es/~civit/PUBLICA/GRANADA01-DEF.zip

CIVIT, M. - MARTÍ, MA. (2002) "Design Principles for a Spanish Treebank", in *TLT02, First Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.
http://www.lsi.upc.es/~civit/PUBLICA/stb.ps.gz

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/

COLE, R. (Ed.) (1997a) "Spoken Output Technologies", in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch5node2.html#Chapter5

COLE, R. (Ed.) (1997b) "Language Resources", in 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..
http://cslu.cse.ogi.edu/HLTsurvey/ch12node2.html#Chapter12

COLE, R.- ZUE, V. (Eds.) (1997) "Spoken Language Input", in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch1node2.html#Chapter1

COLE, R.A.- MARIANI, J.- USZKOREIT, H.- ZAENEN, A.- ZUE, V. (Eds.) (1997) *Survey of the State of the Art in Human Language Technology*. Cambridge: Cambridge University Press.
http://cslu.cse.ogi.edu/HLTsurvey/HLTsurvey.html

CORTÁZAR, I.- RODRÍGUEZ, M.A.- GARRIDO , J.M.- CAMINERO, F.J.- BERNAT, J.- RELAÑO, J.- GARIJO, F.J.- HERNÁNDEZ, L.A. (2002) "Últimos desarrollos en tecnologías del voz y del lenguaje", *Comunicaciones de Telefónica I+D* 24: 25-64.
http://www.tid.es/presencia/publicaciones/comsid/esp/24/art2.pdf

CHEVREAU, K.- COCH, J.- GARCÍA-MOYA, J.A. - ALONSO, M. (1999) "Generación multilingüe de boletines meteorológicos", *Procesamiento del Lenguaje Natural,* Revista nº 25: 51-58.

de YZAGUIRRE, Ll. - RIBAS, M.- VIVALDI, J.- CABRÉ, M.T. (2000) "Some Technical Aspects About Aligning Near Languages", in GABRILIDOU, M. *et al*. (Ed.) Proceedings of the Second International Conference on Language Resources and Evaluation. Athens: National Technical University of Athens Press. Vol I. pp. 545-548.
http://terminotica.upf.es/CREL/atenes.ps

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

DRAXLER, C. (2000) "Speech databases", in VAN EYNDE, F.- GIBBON, D. (Eds.) *Lexicon Development for Speech and Language Processing*. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 12). pp. 169-206.

DUTOIT, T. (1997) *An Introduction to Text-to-Speech Synthesis*. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 3).

DUTOIT, T. (1999) *A Short Introduction to Text-to-Speech Synthesis*. TCTS Lab, Faculté Polytechnique de Mons.
http://tcts.fpms.ac.be/synthesis/introtts.html

DYBKJAER, L.- BERMAN, S.- KIPP, M.- WAGENER, M.- PIRRELLI, V.- REITHINGER, N.- SORIA, C. (2001) *Survey of Existing Tools, Standards and User Needs for Annotation of Natural Interaction and Multimodal Data*. ISLE Natural Interactivity and Multimodality Working Group. D11.1. January 2001.
http://isle.nis.sdu.dk/reports/wp11/

ENRÍQUEZ, E. (1991) "El problema de las ambigüedades fonéticas y su tratamiento automático", *Boletín de la Real Academia de la Lengua Española* tomo LXXI, cuaderno XXLII (Enero- abril, 1991) pp. 157-183.

ESCUDERO, D.- CARDEÑOSO, V. (2001) "Modelo cuantitativo de entonación del español", *Procesamiento del Lenguaje Natural, Revista* nº. 27: 233-240.
http://www.sepln.org/revistaSEPLN/revista/27/27-articulo27.pdf

FELIU, J.- VIVALDI, J.- CABRÉ, M.T. (2002) *Ontologies: a review.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (Papers de l'IULA, Sèrie Informes, 34).
ftp://ftp.iula.upf.es/pub/publicacions/02inf034.pdf

FELLBAUM, C. (Ed.) (1998) *WordNet. An Electronic Lexical Database*. Cambridge, Mass.: The MIT Press - Bradford Books (Language, Speech and Communication Series).

FERNÁNDEZ, X.- RODRÍGUEZ, E. (2000) "Proposición de un marco adecuado para el estudio de contornos de F0 para síntesis de voz", *Procesamiento del Lenguaje Natural*, *Revista* nº 26: 175-182.

FERNÁNDEZ, S.- ORTEGA, R.- SERRANO, R. (2000) "Portales de voz: Internet en el teléfono", *Comunicaciones de Telefónica I+D* 19: 15-24.
http://www.tid.es/presencia/publicaciones/comsid/esp/19/ART_2.PDF

GARRIDO, J.M. (2001) "La estructura de las curvas melódicas del español: propuesta de modelización", *Lingüística Española Actual* 23, 2: 173-209.

GARRIDO, J.M.- ORTÍN, I.- QUAZZA, S.- SALZA, P.L.- MANCINI, F. (2000) "Desarrollo de un módulo de asignación de parámetros prosódicos para la versión en español del sistema de conversión texto-habla ACTOR®", *Procesamiento del Lenguaje Natural*, *Revista* nº 26: 183-190.

GARSIDE, R.- LEECH, G.- McENERY, T. (Eds.) (1997) *Corpus Annotation: Linguistic Information from Computer Text Corpora*. Harlow: Addison Wesley Longman.

GIBBON, D. - MOORE, R.- WINSKI, R. (Eds.) (1997) *Handbook on Standards and Resources for Spoken Language Systems*. Berlin: Mouton De Gruyter.

GIBBON, D.- MERTINS, I.- MOORE, R. (Eds.) (2000) *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation*. Dordrecht: Kluwer Academic Publishers (Kluwer International Series in Engineering and Computer Science, 565).

GÓMEZ, X. (1999) *La escritura asistida por ordenador: problemas de sintaxis y de estilo*. Vigo: Servicio de Publicacións da Universidade de Vigo.

GÓMEZ, X. (2000a) "Lingüística computacional", in RAMALLO, F.- REI-DOVAL, G.- RODRÍGUEZ, X.P. (Eds.) *Manual de Ciencias da Linguaxe*. Vigo: Edicións Xerais de Galicia (Universitaria, Manuais, 4). pp. 221-268.

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

GÓMEZ, X. (2001) "Recursos d'ajut a l'edició. Ortografia, sintaxi i estil", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 15-26.

GONZÁLEZ, A.L.- GOÑI, J.M.- GONZÁLEZ, J.C. (1995) "Un analizador morfológico para el castellano basado en chart", in *Actas de la VI Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA'95)*. Alicante, Noviembre de 1995.
http://www.mat.upm.es/~aries/docs/aepia95.ps

GONZALO, J.- VERDEJO, M.F. (2001) "Recuperació i extracció d'informació", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 151-187.

GREENBERG, S. (2001) "From here tu utility - Melding phonetic insight with speech technology", in DALSGAARD, P.- LINDBERG, B.- NEMMER, H. (Eds.) *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology*. September 3-7, 2001, Aalborg, Denmark. Vol 4. pp. 2485-2488.
http://www.icsi.berkeley.edu/~steveng/PDF/Utility.pdf

GRISHMAN, R. (1986) *Computational Linguistics. An Introduction*. Cambridge: Cambridge University Press (Studies in Natural Language Processing). Trad. cast. de A. Moreno: *Introducción a la lingüística computacional*. Madrid: Visor (Lingüística y Conocimiento, 9).

HLT Central, Human Language Technologies. http://www.hltcentral.org/

HUANG, X.- ACERO, A.- HON, H.-H.- REDDY, R. (2001) *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. New Jersey: Prentice Hall.

Ingeniería lingüística. Cómo aprovechar la fuerza del lenguaje. Luxembourg: Language Engineering, Telematics Applications Programme. Versión española: Observatorio Español de Industrias de la Lengua, Instituto Cervantes.
http://www.hltcentral.org/usr_docs/Harness/harness-es.htm

JURAFSKY, D.- MARTIN, J.H. (2000) *Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*. New Jersey: Prentice Hall.
http://www.cs.colorado.edu/~martin/slp.html

KENNEDY, G. (1998) *An Introduction to Corpus Linguistics*. London: Longman (Studies in Language and Linguistics).

KURZWEIL, R. (1998) "When Will HAL Understand What We Are Saying? Computer Speech Recognition and Understanding", in 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.- COLE, R.A. (1997) "Spoken Language Corpora", in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch12node5.html#SECTION123

Language Technology World. National Language Technology Competence Center. DFKI, Saarbrücken.
http://www.lt-world.org/

LENCI, A. - BEL, N.- BUSA, F.- CALZOLARI, N.- GOLA, E.- MONACHINI, M.- OGONOWSKI, A.- PETERS, I.- PETERS, W.- RUIMY, N.- VILLEGAS, M.- ZAMPOLLI, A. (2000) “SIMPLE: A General Framework for the Development of Multilingual Lexicons”, *International Journal of Lexicography* 13.
http://www.ub.es/gilcub/lascosas/pubYreps/simple.doc

LEECH, G.- EYES, E. (1997) "Syntactic annotation: treebanks", in GARSIDE, R.- LEECH, G.- McENERY, T. (Eds.)*Corpus Annotation. Linguistic Information from Computer Text Corpora*. London & New York: Longman. pp. 34-52.

LEECH, G.- WEISSER, M.- WILSON, A.- GRICE, M. (1998) *Survey and Guidelines for the representation and annotation of dialogue*. LE-EAGLES- WP4-4 Integrated Resources Working Group. 16 October 1998.
http://www.ling.lancs.ac.uk/eagles/delivera/wp4final.htm

LEECH, G.- WILSON, A. (1999) "Standards for Tagsets", in van HALTEREN, H. (Ed.) *Syntactic Wordclass Tagging*. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 9). pp. 55-80.

LÓPEZ, E.- HERNÁNDEZ, L.A. (1995) "Automatic Data-Driven Prosodic Modeling for Text to Speech", in *Eurospeech'95. Proceedings of the 4th European Conference on Speech Communication and Technology*. Madrid, Spain, 18-21 September, 1995. Vol 1, pp. 585-588.
ftp://ftp.gaps.ssr.upm.es/pub/tts/EUROS95.ps

LLEIDA, E. (2000) *Tecnologías del habla.* Grupo de Tecnologías de las Comunicaciones, Centro Politécnico Superior, Universidad de Zaragoza.
http://www.gtc.cps.unizar.es/~eduardo/investigacion/voz/voz.html

LLISTERRI, J. (1996) *Preliminary Recommendations on Spoken Texts*. EAGLES Document EAG-TCWG-STP/P, May 1996.
http://www.ilc.pi.cnr.it/EAGLES96/spokentx/spokentx.html

LLISTERRI, J. (1999a) "Corpus orals per a la fonètica i les tecnologies de la parla", in *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/Resum_tarragona_99.html

LLISTERRI, J. (1999b) “Transcripción, etiquetado y codificación de corpus orales", in GÓMEZ, J.- LORENZO, A.- PÉREZ, J.- ÁLVAREZ, A. (Eds.) Panorama de la investigación en lingüística informática. *RESLA, Revista Española de Lingüística Aplicada*, Volumen monográfico. pp. 53-82.
http://liceu.uab.es/~joaquim/publicacions/RESLA_99.pdf

LLISTERRI, J. (2003) “Lingüística y tecnologías del lenguaje”, *Lynx. Panorámica de Estudios Lingüísticos* (Departament de Teoria dels Llenguatges, Universitat de València) 2: 9-71.
http://liceu.uab.es/~joaquim/publicacions/TecnoLing_Lynx02.pdf

LLISTERRI, J. (1999c) “Tecnologías lingüísticas y sociedad de la información”, *Economía Industrial* (*La sociedad de la información en España* I) 325: 37-56.
http://liceu.uab.es/~joaquim/publicacions/LengEsp_SocInfo.pdf

LLISTERRI, J. (2001a) "Les tecnologies de la parla", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 239-272.

LLISTERRI, J. (2001b) "La conversión de texto en habla", *Quark. Ciencia, Medicina, Comunicación y Cultura* 21: 79-89.
http://liceu.uab.es/~joaquim/publicacions/Quark2001/CTH_Quark_01.pdf

LLISTERRI, J. (2001c) "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. (2002) "Las tecnologías del habla: Entre la ingeniería y la lingüística", *Congreso Internacional La Ciencia ante el Público. Cultura humanística y desarrollo científico y tecnológico*. Universidad de Salamanca, Salamanca, 28-31 de octubre de 2002. Edición en CD-ROM. pp. 51-74.
http://liceu.uab.es/~joaquim/publicacions/TecnolHab_Salamanca_02.pdf

LLISTERRI, J.- AGUILAR, L.- GARRIDO, J.M.- MACHUCA, M.J.- MARÍN, R.- DE LA MOTA, C.- RÍOS, A. (1999) "Fonética y tecnologías del habla", in 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_TecnolHabla.pdf

LLISTERRI, J.- GARRIDO, J.M. (1998) "La ingeniería lingüística en España", in *El español en el mundo. Anuario del Instituto Cervantes. 1998*. Madrid: Instituto Cervantes - Arco/Libros SL. pp. 299-391.
http://cvc.cervantes.es/obref/anuario/anuario_98/llisterri/

LLISTERRI, J.- MACHUCA, M.J.- DE LA MOTA, C.- RIERA, M.- RÍOS, A. (2003) "Entonación y tecnologías del habla", en PRIETO, P. (Ed.) *Teorías de la entonación*. Barcelona: Ariel (Ariel Lingüística). (en prensa).

LLISTERRI, J. - MARTÍ, M.A. (2002) "Las tecnologías lingüísticas en la Sociedad de la Información", in MARTÍ, M.A.- LLISTERRI, J. (Eds.) *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita.* Barcelona: Fundación Duques de Soria - Edicions Universitat de Barcelona (Manuals UB, 53). pp. 13-28.

MACARRÓN, A.- ESCALADA, G.- RODRÍGUEZ, M.A. (1991) "Generation of duration rules for a Spanish text-to-speech synthesizer", in *Eurospeech'91. 2nd European Conference on Speech Communication and Technology*. Genova, Italy, 24-26 September 1991. Vol. 2. pp. 617-620.

MACHUCA, M -, BUENO, L. - CALONGE, R. - ESTRUCH, M. - RIERA, M. (2000) "Corpus de diàleg", *Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*, Barcelona, 4 d'abril de 2000. CREL, Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya - Institut d'Estudis Catalans.
http://liceu.uab.es/publicacions/SFI_UAB_Corpus_Dialeg.pdf

MARTÍ, M.A. (Coord.) (2001) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53).

MARTÍ, M.A.- LLISTERRI, J. (2001) "L'enginyeria lingüística en la societat de la informació", *Digit-HVM, Revista Digital d'Humanitats* (Universitat Oberta de Catalunya) 3.
http://www.uoc.es/humfil/digithum/digithum3/catala/Art_Llisterri_Marti/index.htm

MARTÍNEZ, P.- GARCÍA, A. (2002) "Utilizando recursos lingüísticos para mejora de la recuperación de información a través de la Web", *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial* 16: 55-64.
http://tornado.dia.fi.upm.es/caepia/numeros/16/pmf.pdf

McENERY, T.- WILSON, A. (1996) *Corpus Linguistics*. Edinburgh: Edinburgh University Press (Edinburgh Textbooks in Empirical Linguistics)
http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

MENGEL, A. - DYBKJAER, L., GARRIDO, J.M. - HEID, U.- KLEIN, M. - PIRRELLI V. - POESIO, M. - QUAZZA, S. - SCHIFFRIN, A. - SORIA, C. (2000) *MATE Dialogue Annotation Guidelines*. MATE Deliverable D2.1. 8 January 2000.
http://www.ims.uni-stuttgart.de/projekte/mate/mdag/

MINKER, W.- BENNACEF, S. (2001) *Parole et dialogue homme-machine.* Paris: Étidions Eyrolles - Éditions du CNRS (Sciences et techniques de l'ingénieur).

MOLANO, A. (2002) "Aplicación de las técnicas de Procesado del Lenguaje Natural en la próxima generación de herramientas de búsqueda de información", *Euromap Language Technologies*.
http://www.hltcentral.org/htmlengine.shtml?id=1060

MONAGHAN, A. (2002) "Prosody in synthetic speech: Problems, solutions and challenges", in KELLER, E. - BAILLY, G.- MONAGHAN, A.- TERKEN, J.- HUCKVALE, M. (Eds.) *Improvements in Speech Synthesis. Cost 258: The Naturalness of Synthetic Speech*. Chichester: John Wiley & Sons. pp. 89-92.

MONZÓN, L.- RODRÍGUEZ, M.A.- ESCALADA, G. (1993) "Módulo de análisis sintáctico para un sistema de conversión texto-voz en castellano", *Boletín de la Sociedad Española para el Procesamiento del Lenguaje Natural* 13: 367-379.

MORENO, A.- LÓPEZ, S.- SÁNCHEZ, F.- GRISHMAN, R. (2002) "Developing a syntactic annotation scheme and tools for a Spanish treebank", in ABEILLÉ, A. (Ed.) *Building and using syntactically annotated corpora*. Dordrecht: Kluwer (Text, Speech and LanguageTechnology).
http://treebank.linguist.jussieu.fr/pdf/9.pdf

MORENO, L.- PALOMAR, M.- MOLINA, A.- FERRÁNDEZ, A. (1999) *Introducción al Procesamiento del Lenguaje Natural.* Alicante: Servicio de Publicaciones de la Universidad de Alicante.

MOURE, T.- LLISTERRI, J. (1996) "Lenguaje y nuevas tecnologías. El campo de la lingüística computacional", en FERNÁNDEZ , M. (Coord.) *Avances en lingüística aplicada*. Santiago de Compostela: Universidade de Santiago de Compostela, Servicio de Publicacións e Intercambio Científico (Avances en, 4). pp. 147-228.

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

OLIVE, J.P. (1998) "'The Talking Computer': Text to Speech Synthesis", in 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

O'SHAUGHNESSY, D. (1987) *Speech Communication. Human and Machine*. Reading, Mass.: Addison Wesley. Second Edition: IEEE Press, 2000.

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", in *Workshop on Developing Language Resources for Minority Languages: Reusability and Strategic Priorities*, LREC'00, Athens, Greece, May 2000.
http://liceu.uab.es/~joaquim/publicacions/Paches_et_al_2000.pdf

PASTOR, M.- SANCHIS, A.- CASACUBERTA, F.- VIDAL, E. (2001) "Eutrans: a speech-to-speech translator prototype", in *Eurospeech 2001. Proceedings of the 7th European Conference on Speech Communication and Technology.* September 3-7, 2001, Aalborg, Denmark.
http://www.iti.upv.es/~prhlt/PAPERS/2001/Pastor01a.pdf.gz

PAYRATÓ, L. (1997) *De professió, lingüista. Panorama de la lingüística aplicada*. Barcelona: Empúries (Biblioteca Universal Empúries, 89). Trad. cast. de J. Giménez: *De profesión, lingüista. Panorama de la lingüística aplicada*. Barcelona: Ariel (Ariel Practicum), 1998.

PÉREZ, J. (1999) "Estándares de anotación en lingüística de corpus", in GÓMEZ, J.- LORENZO, A.- PÉREZ, J.- ÁLVAREZ, A. (Eds.) Panorama de la investigación en lingüística informática. *RESLA, Revista Española de Lingüística Aplicada*, Volumen monográfico. pp.25-52.

PIERREL, J.M. (Ed.) (2000) *Ingénierie des langues*. Paris: Hermes Sciences.

PINO, M.- SANTALLA, M.P. (1996) "Codificación de la anotación morfosintáctica en SGML", *Procesamiento del Lenguaje Natural, Revista* nº 19: 101-117.

POLS, L.C.W. (1996) "Speech Synthesis Evaluation", in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch13node9.html#SECTION137

POLS, L.C.W.- JEKOSCH, U. (1997) "A Structured Way of Looking at the Performance of Text-to-Speech Systems", in van SANTEN, J.P.H. - SPROAT, R.W.- OLIVE, J.P.- HIRSCHBERG. J. (Eds.) *Progress in Speech Synthesis*. New York: Springer. pp. 519-528.

PRICE, P. (1997) "Spoken Language Understanding", in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch1node10.html

QUAZZA, S.- GARRIDO, J.M. (1998) “Prosody”, in KLEIN, M. (Ed.) *Supported Coding Schemes*. MATE Deliverable D1.1. LE Telematics Project LE4 – 8370. July 1998.
http://liceu.uab.es/publicacions/MATED1.1.6Prosody/D11_6_Prosody.html

QUAZZA, S.- van den HEUVEL, H. (2000) "The use of lexica in text-to-speech systems", in VAN EYNDE, F.- GIBBON, D. (Eds.) *Lexicon Development for Speech and Language Processing*. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 12). pp. 207-234.

RAFEL, J.- SOLER, J. (2001) “El processament de corpus. La lingüística empírica", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 27-59.

RAMÍREZ, F.- SÁNCHEZ, F. (1996) "GramChek: un corrector gramatical para español", *Procesamiento del Lenguaje Natural, Boletín* nº 19: 30-37.

RAMÍREZ, F.- SÁNCHEZ, F.- DECLERK, T. (1998) "CON-TEXT, Un corrector gramatical de bajo nivel", *Procesamiento del Lenguaje Natural*, Revista nº 23: 165-170.

RENALS, S.- ROBINSON, T. (Eds.) (2000) Special Issue on Accessing Information on Spoken Audio, *Speech Communication* 32, 1-2.

RIERA, M. (2000) "Corpus de recerca: Corpus prosòdic", *Jornades del Centre de Referència en Enginyeria Lingüística (CREL)*, Barcelona, 4 d'abril de 2000. CREL, Centre de Referència en Enginyeria Lingüística de la Generalitat de Catalunya - Institut d'Estudis Catalans.
http://liceu.uab.es/publicacions/SFI_UAB_Corpus_Prosodic.pdf

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/

RODRÍGUEZ, C.- RUBIO, C.- SÁNCHEZ, A.- SOPEÑA, L. (1992) "Herramientas de ayuda a la redacción de textos: un sistema de verificación léxica, sintáctica y estilística", *Voz y Letras. Revista de Filología* 3,: 155-174.

RODRÍGUEZ, H. (2000) "Técnicas básicas en el tratamiento informático de la lengua", *Quark. Ciencia, Medicina, Comunicación y Cultura* 19: 26-34.
http://www.imim.es/quark/num19/019026.htm

RODRÍGUEZ, H. (2001) "Les interfícies en llenguatge natural", in MARTÍ, M.A. (Coord.) *Les tecnologies del llenguatge*. Barcelona: Edicions de la Universitat Oberta de Catalunya (Manuals, 53). pp. 121-149.

RODRÍGUEZ, H. (2002) "Técnicas de análisis sintáctico" in 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 (Biblioteca de la Universitat de Barcelona, Manuales, 53). pp. 91-132.

RODRÍGUEZ, M.A.- CORTÁZAR, I.- TAPIAS, D.- RELAÑO, J. (2001) "Estado del arte en tecnologías de voz", *Comunicaciones de Telefónica I+D* 20: 117-136.
http://www.tid.es/presencia/publicaciones/comsid/esp/20/8XX.PDF

SÁNCHEZ, F.- SERRANO, A. (1995) "Desarrollo de un etiquetador morfosintáctico para el español", *Procesamiento del Lenguaje Natural, Boletín* nº 17: 14-28.

SANTANA, O.- PÉREZ, J.- CARRERAS, F.- DUQUE, J.- HERNÁNDEZ, Z.- RODRÍGUEZ, G. (1999) "FLANOM: Flexionador y lematizador automático de formas nominales", *Lingüística Española Actual 2*1, 2: 253-297.
http://www.gedlc.ulpgc.es/art_ps/art29.pdf

SANTANA, O.- PÉREZ, J.- HERNÁNDEZ, Z.- CARRERAS, F.- RODRÍGUEZ, G. (1997) "FLAVER: Flexionador y lematizador automático de formas verbales", *Lingüística Española Actual* 19, 2: 229-282.
http://www.gedlc.ulpgc.es/art_ps/art28.pdf

SANTOS, A.- MUÑOZ, P.- MARTÍNEZ, M. (1988) "Diseño y evaluación de reglas de duración en la conversión de texto a voz", *Procesamiento del Lenguaje Natural*, *Boletín* nº 6: 69-92.

SIDOROV, G. (2001) "Problemas actuales de lingüística computacional", *Revista Digital Universitaria* (Universidad Autónoma Nacional de México), 2, 1.
http://www.revista.unam.mx/vol.2/num1/art1/

SINCLAIR, J. (1996) *Preliminary Recommendations on Corpus Typology*. EAGLES Document EAG-TCWG-CTYP/P, May 1996.
http://www.ilc.pi.cnr.it/EAGLES96/corpustyp/corpustyp.html

SPARCK-JONES, K. (1999) "What is the role of NLP in text retrieval?", in STRZALKOWSKI, T. (Ed.) *Natural Language Information Retrieval*. Dordrecht: Kluwer Academic Publishers. pp. 1-24.

SPERBERG-McQUEEN, C.M.- BURNARD, L. (Eds.) (2002). *TEI P4: Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative Consortium. XML Version: Oxford, Providence, Charlottesville, Bergen.
http://www.tei-c.org/P4X/

STRIK, H.- CUCCHIARINI, C. (1999) "Modeling pronunciation variation for ASR: A survey of the literature", Special Issue on Modeling Pronunciation Variation for Automatic Speech Recognition. *Speech Communication* 29, 2-4: 225-246.
http://lands.let.kun.nl/TSpublic/strik/a64b.html

STUBBS, M. (1996) *Text and Corpus Analysis. Computer Assisted Studies of Language and Culture*. Oxford: Basil Blackwell (Language in Society).

SUBIRATS, C.- ORTEGA, M. (2000) "Tratamiento automático de la información textual en español mediante bases de información lingüística y transductores", *Estudios de Lingüística Española* 10.
http://elies.rediris.es/elies10/

TAPIAS, D. (1999) "Sistemas de reconocimiento de voz en las telecomunicaciones", in GÓMEZ, J.- LORENZO, A.- PÉREZ, J.- ÁLVAREZ, A. (Eds.) Panorama de la investigación

en lingüística informática. *RESLA, Revista Española de Lingüística Aplicada*, Volumen monográfico. pp. 83-102.

TAPIAS, D. (2002) "Interfaces de voz con lenguaje natural". in 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 (Biblioteca de la Universitat de Barcelona, Manuales, 53). pp. 189-207.

TORRUELLA, J.- LLISTERRI, J. (1999) “Diseño de corpus textuales y orales”, in 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. 45-77.

TRUJILLO, A. (2000) "Estrategias de traducción automática", *Quark. Ciencia, Medicina, Comunicación y Cultura* 19: 53 - 57.
http://www.imim.es/quark/num19/019053.htm

UREÑA, L.A. (2002) *Resolución de la ambigüedad léxica en tareas de clasificación automática de documentos*. Alicante: Editorial Club Universitario (Monografías de la Sociedad Española para el Procesamiento del Lenguaje Natural, 1).

USZKOREIT, H. (2000) *What is Computational Linguistics?* Department of Computational Linguistics and Phonetics, Saarland University, Saarbrücken.
http://www.coli.uni-sb.de/~hansu/what_is_cl.html

USZKOREIT, H. (2002) *Language Technology. A First Overview*. Language Technology World. German Research Center for Artificial Intelligence (DFKI). Saarbrücken.
http://www.dfki.de/~hansu/LT.pdf

USZKOREIT, H. (Ed.) (1997) "Language Generation", in in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch4node2.html#Chapter4

van HALTEREN, H.- VOUTILAINEN, A. (1999) "Automatic Taggers: An Introduction", LEECH, G.- WILSON, A. (1999) "Standards for Tagsets", in van HALTEREN, H. (Ed.) *Syntactic Wordclass Tagging*. Dordrecht: Kluwer Academic Publishers (Text, Speech and Language Technology, 9). pp. 109-116.

van SANTEN, J.H.P. (1997) "Prosodic modelling in text-to-speech synthesis", in KOKKINAKIS, G.- FAKOTAKIS, N.- DERMATAS, E. (Eds.) *Eurospeech'97. 5th European Conference on Speech Communication and Technology.* Rhodes, Greece, 22-25 September 1997. Vol. 1. pp. KN-18 - KN-28.
http://www.bell-labs.com/project/tts/jphvs-keynote97.ps

VÁZQUEZ, G. - FERNÁNDEZ A., A.M.- MARTÍ, M.A. (2002) "Léxicos verbales computacionales" in 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 (Biblioteca de la Universitat de Barcelona, Manuales, 53). pp. 29-60.

VERDEJO, F.- GONZALO, J.- PEÑAS, A. (1999) *Information Retrieval & Natural Language Processing*. Grupo de Lenguaje Natural, Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia.

http://rayuela.lsi.uned.es/~ircourse/

VILLEGAS, M.- BROSA, I.- BEL, N. (1998) "El léxico PAROLE del español", *Procesamiento del Lenguaje Natural*, *Revista* nº 23: 84-89.
http://www.ub.es/gilcub/lascosas/pubYreps/seplen98.rtf

VIVALDI, J.- de YZAGUIRRE, Ll.- SOLÉ, X. (1996) *Marcatge estructural i morfosintàctic del corpus tècnic amb l'estàndard SGML*. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra (Papers de l'IULA, Sèrie Informes, 1).
ftp://ftp.iula.upf.es/pub/publicacions/96inf001.pdf

VOSSEN, P. (2001) "Oportunitats per a l'enginyeria lingüística", *Digit-HVM, Revista Digital d'Humanitats* (Universitat Oberta de Catalunya) 3.
http://www.uoc.edu/humfil/articles/cat/vossen/vossen.html

VOSSEN, P. (Ed.) (1998) *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

WAHLSTER, W. (2000) "Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System", in WAHLSTER, W. (Ed.) *Verbmobil: Foundations of Speech-to-Speech Translation*. Heidelberg - New York: Springer Verlag (Artificial Intelligence).
http://verbmobil.dfki.de/ww.html

WAIBEL, A. (1996) "Interactive Translation of Conversational Speech", *Computer* 29, 7, July:
http://www.ri.cmu.edu/pub_files/pub1/waibel_alex_1996_1/waibel_alex_1996_1.pdf

WAIBEL, A. (2000) "La traducción interactiva del habla", *Quark. Ciencia, Medicina, Comunicación y Cultura* 19: 58 - 65.
http://www.imim.es/quark/num19/019058.htm

ZAENEN, A. (Ed.) (1997) "Language Analysis and Understanding", in 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.
http://cslu.cse.ogi.edu/HLTsurvey/ch3node2.html#Chapter3

ZUE, V. (1999) "Talking with your computer", *Scientific American*, August 1999. pp. 40-41.
http://www.sciam.com/article.cfm?articleID=0009D2B7-F2E6-1C72-9B81809EC588EF21&catID=2

Grupos de investigación y páginas con demostraciones citadas en el texto

AhoLab, Universidad del País Vasco. http://bips.bi.ehu.es/ahoweb/

BASURDE - Sistema de Diálogo para Habla Espontánea en un Dominio Semántico Restringido, Grupo de Procesado del Habla, Universitat Politècnica de Catalunya.
http://gps-tsc.upc.es/veu/basurde/

Centro Ramón Piñeiro para a Investigación en Humanidades, Santiago de Compostela.
http://www.cirp.es/

CLiC – Centre de Llenguatge i Computació, Universitat de Barcelona.
http://clic.fil.ub.es/

COLE – Compiladores y Lenguajes, Universidade de La Coruña.
http://www.grupocole.org/

Dadedalus – Data, Decisions and Language SA., Madrid
http://gramatia.daedalus.es/; http://www.daedalus.es/STILUS/

DELI – Grupo de Lingüistas, Informáticos e Ingenieros de Deusto, Universidad de Deusto.
http://sirio.deusto.es/abaitua/deli/

ECA_SIM - Grupo de Computación Avanzada y Entornos de Comunicación Multimodal, Universidad de Valladolid. http://logos.dcs.fi.uva.es/

GEDLC - Grupo de Estructura de Datos y Lingüística Computacional, Universidad de las Palmas de Gran Canaria. http://www.gedlc.ulpgc.es/

gilcUB - Grup d'Investigació en Lingüística Computacional, Universitat de Barcelona.
http://www.ub.es/gilcub/

Grupo de Procesado del Habla, Universitat Politècnica de Catalunya. http://gps-tsc.upc.es/veu/

Grupo de Reconocimiento de Formas y Tecnología del Habla, Universidad del País Vasco.
http://grah.ehu.es/

Grupo de Sistemas Inteligentes, Universidad Politécnica de Madrid.
http://www.gsi.dit.upm.es/

Grupo de Tecnología del Habla, Universidad Politécnica de Madrid. http://www-gth.die.upm.es/

Grupo de Tecnologías de las Comunicaciones, Universidad de Zaragoza.
http://www.gtc.cps.unizar.es/

Grupo de Tratamiento de la Señal, Universidade de Vigo.
http://www.gts.tsc.uvigo.es

GSTC - Grupo de Investigación en Señales, Telemática y Comunicaciones, Universidad de Granada. http://ceres.ugr.es/

HERMES – Hemerotecas Multilingües: Recuperación Multilingüe y Extracción Semántica. Grupo de Lenguaje Natural de la Universidad Nacional de Educación a Distancia.
http://terral.lsi.uned.es/hermes/

IEC – Institut d'Estudis Catalans, Barcelona. Portal de Dades Lingüístiques.
http://pdl.iecat.net/

ILGA – Instituto da Lingua Galega, Universidade de Santiago de Compostela.
http://www.usc.es/~ilgas/

ITEM - Recuperación de Información Textual en un Entorno Multilingüe con Técnicas de Lenguaje Natural, Grupo de Lenguaje Natural de la Universidad Nacional de Educación a Distancia. http://sensei.ieec.uned.es/item/

IULA – Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
http://www.iula.upf.es/

IXA Taldea, Universidad del País Vasco. http://ixa.si.ehu.es/

LLISTERRI, J. (2003) “Lingüística y tecnologías del lenguaje”, *Lynx. Panorámica de Estudios Lingüísticos* (Departament de Teoria dels Llenguatges, Universitat de València) 2: 9-71. http://liceu.uab.es/~joaquim/publicacions/TecnoLing_Lynx02.pdf

PRHLT - Grupo de Reconocimiento de Formas y Tecnologías para el Lenguaje Humano, Universidad Politécnica de Valencia. http://www.iti.upv.es/~prhlt/index.html

Secció de Teoria del Senyal, Universitat Ramon Llull. http://www.salleurl.edu/Eng/elsDCTS/tsenyal/index.htm

Signum Cía. Ltda., Quito. http://www.lenguaje.com/

SLI – Seminario de Lingüística Informática, Universidade de Vigo. http://www.uvigo.es/webs/sli/

TALP – Centre de Tecnologies i Aplicacions del Llenguatge i la Parla, Universitat Politècnica de Catalunya. http://www.talp.upc.es/