Llisterri, J. (2007). El español y las nuevas tecnologías. In M. Lacorte (Ed.), *Lingüística aplicada del español*. (pp. 483-520). Madrid: Arco/Libros.

http://liceu.uab.cat/~joaquim/publicacions/Llisterri_07_Tecnologias_Linguisticas_Espanol.pdf

# 14

# EL ESPAÑOL Y LAS NUEVAS TECNOLOGÍAS

JOAQUIM LLISTERRI
*Universitat Autònoma de Barcelona*

## 14.1. INTRODUCCIÓN

Las tecnologías lingüísticas (TL o, en inglés, LT, *Language Technologies*), tecnologías de la lengua o tecnologías para el lenguaje humano (TLH o HLT, *Human Language Technologies*), son todas aquellas que se integran en programas informáticos de uso local, en la red o en entornos que requieran la interacción entre personas y ordenadores, para permitir el tratamiento de las lenguas, sea en su vertiente oral o escrita. Como veremos en este capítulo, las tecnologías lingüísticas pretenden facilitar el uso de las computadoras y el acceso a las redes que configuran la sociedad de la información y del conocimiento, sin que por ello tengamos que renunciar a nuestro uso habitual del lenguaje (Llisterri y Martí, 2002; Martí, 2003).

En este contexto se encuentra también el término "ingeniería lingüística" (IL o LE, *Language Engineering*), referido al empleo de las técnicas propias de la informática para desarrollar sistemas que incluyen componentes relacionados con el tratamiento de los textos escritos y del habla. 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 "lingüística informática", denominación que suele hacer referencia al uso de herramientas informáticas en la investigación lingüística o filológica (Gómez y Lorenzo, 1996; Blecua *et al.*, 1999). Finalmente, la "lingüística computacional" (LC o CL, *Computational Linguistics*) sería 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, 2000; Martí y Castellón, 2000; Mitkov, 2003; Lavid, 2005; Ruiz, 2005).

Las tecnologías que se ocupan específicamente del tratamiento de la lengua oral son las llamadas tecnologías del habla (tratadas en la sección 2), mientras que las que tienen como objeto los textos escritos se enmarcan en el procesamiento del lenguaje natural (sección 4), aunque también podrían definirse como "tecnologías del texto". En ambos casos, su desarrollo requiere el uso de recursos lingüísticos (sección 6), entre lo que cabe incluir los corpus, las bases de datos léxicos y las gramáticas computacionales[1].

En lo que se refiere a las tecnologías lingüísticas para el español, debe señalarse que no se ha alcanzado, por el momento, el mismo nivel de desarrollo global que para el inglés o para algunas otras lenguas. La razón no estriba, ciertamente, en la calidad de la investigación llevada a cabo, sino en las diferencias en el peso económico entre las lenguas (véase el capítulo 15) que dificultan que los resultados de la investigación lleguen al mercado y, por tanto, a los usuarios finales de las tecnologías. Debe reconocerse, sin embargo, que se trata de un campo con un amplio potencial, apoyado en un conjunto de equipos de investigación en las universidades, y que cuenta con algunas empresas que comercializan sus productos[2].

## 14.2. Las tecnologías del habla

Las tecnologías del habla (*Speech Technologies*) tienen como finalidad el tratamiento informático de la lengua oral. Hacen posible que un ordenador ofrezca información hablada –síntesis del habla–, reconozca los enunciados emitidos por una persona –reconocimiento

---

[1] En la presentación de las tecnologías del lenguaje que se realiza en este capítulo se han primado los aspectos lingüísticos sobre los informáticos, siguiendo la perspectiva adoptada en Llisterri (2003). Los lectores interesados en profundizar en los temas que aquí se exponen pueden encontrar información útil en manuales como los de Cole *et al.* (1997), Martí (2001), Coleman (2005) y Lavid (2005), o en textos más avanzados como los de Dale *et al.* (2000), Jurafksy y Martin (2000), Huang *et al.* (2001) y Farghaly (2003).

[2] A lo largo del presente trabajo se hará referencia a algunos de los logros obtenidos, pero para una valoración del campo de las tecnologías lingüísticas en español remitimos al lector a planteamientos generales como los que se encuentran en Castillo (1995), Pascual (1995) y Blecua (2001), a estudios específicos como Llisterri y Garrido (1998), Llisterri (1999 y 2004a), Moreno (2004) y Sánchez (2004), y a la recopilación de Rubio y Hernáez (2005). Una fuente de información para conocer los trabajos realizados en España son las actas de los congresos anuales de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN), publicadas en la revista de la asociación y accesibles a través de las páginas en Internet de la SEPLN.

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 en varias lenguas.

En esta sección nos centraremos en las tres vertientes básicas que configuran las tecnologías del habla y que acabamos de mencionar: la síntesis, con especial atención a la conversión de texto en habla, el reconocimiento y los sistemas de diálogo. Las aplicaciones se tratarán en la sección 6.3, mientras que los recursos lingüísticos necesarios para desarrollarlas se abordan en la sección 6.1[3].

### 14.2.1. *La síntesis del habla*

El propósito de la síntesis del habla es la generación automática de mensajes orales con el fin de dotar a los ordenadores de una salida vocal. La técnica más habitualmente empleada para muchas de las aplicaciones de la síntesis es la conversión de texto en habla (CTH o TTS, *Text-to-Speech Synthesis*), mediante la cual se transforma automáticamente cualquier texto escrito en su correspondiente realización sonora (Dutoit 1997; Llisterri *et al.*, 2004). La estructura de un conversor suele consistir en un conjunto de módulos, cada uno dedicado a una tarea específica en el proceso de convertir una cadena inicial de caracteres –el texto de entrada, en soporte electrónico– en una señal sonora lo más semejante posible a la lectura en voz alta del texto original.

El primer módulo de un conversor, conocido como módulo de preprocesamiento, tiene como misión deletrear elementos como las abreviaturas, los números, los símbolos especiales, etc. Una vez el texto inicial se ha convertido ya en una cadena de caracteres, es preciso acercar la forma ortográfica a una representación más cercana a la realidad sonora, tarea que se lleva a cabo en el módulo de transcripción fonética automática. Cuando se dispone del texto fonéticamente transcrito –o, en ocasiones, incluso en etapas anteriores del proceso–, es conveniente realizar un análisis lingüístico que com-

[3] Pueden encontrarse referencias más detalladas en trabajos introductorios como Llisterri (2001) o en manuales especializados como los de O'Shaughessy (2000), Holmes y Holmes (2001), Huang *et al.* (2001) o Coleman (2005); los aspectos más lingüísticos de las tecnologías del habla se abordan de un modo general en Llisterri *et al.* (2003a). En lo que respecta a la situación de la lengua española en relación con las tecnologías del habla, puede verse la reflexión general de Golderos (2001) o las panorámicas presentadas en Mora y Rodríguez (2001) y en Llisterri (2004ab).

plementa la tarea de otros módulos del conversor. Para tal fin pueden utilizarse las herramientas desarrolladas en el marco del procesamiento del lenguaje natural (véase la sección 4.1), como los analizadores morfológicos y los analizadores sintácticos.

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. El módulo prosódico de un conversor contiene modelos que especifican la duración –y, en algunos casos, la intensidad– de los segmentos, el contorno melódico de todo el enunciado, las modificaciones acústicas producidas por el acento y la colocación y la duración de las pausas (Llisterri *et al.*, 2003b). 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 de síntesis que constituirán la forma sonora del mensaje. Un diccionario de unidades de síntesis debe contener el inventario completo de segmentos –alófonos y fonemas– de la lengua sobre la que se trabaja, así como sus posibles combinaciones. En la síntesis han sido habituales los difonemas, que consisten en una combinación entre la mitad del primer sonido que lo forma y la mitad del segundo. Por razones de economía, las unidades de síntesis se guardan de forma parametrizada recurriendo, por ejemplo, al modelo de la fuente y el filtro propio de la fonética acústica.

Muchos sistemas comerciales de conversión de texto en habla emplean, en la actualidad, la técnica conocida como síntesis a partir de corpus, mediante la que tanto las unidades de síntesis como los datos prosódicos se extraen de un corpus de habla muy amplio, grabado por un locutor adecuadamente seleccionado. Con ello se consigue un alto grado de naturalidad, ya que se tiende a buscar en el corpus las realizaciones que coincidan en mayor medida con el texto que debe sintetizarse. Existen para el español diversos sistemas de conversión de texto en habla, algunos desarrollados por grupos de investigación en universidades (Politécnica de Cataluña, Politécnica de Madrid, Valladolid, Vigo, Zaragoza), y otros por empresas, tanto en España (Atlas, Telefónica I+D) como en el resto de Europa (Acapela, Loquendo) o en Estados Unidos (AT&T Labs, ScanSoft). Algunos de estos sistemas disponen incluso de versiones adaptadas a las distintas variedades geográficas de la lengua. Incorporan también el español sistemas como Festival, de la Universidad de Edimburgo, del que puede obtenerse libremente el código fuente.

### 14.2.2. *El reconocimiento del habla*

En el reconocimiento automático del habla (RAH o ASR, *Automatic Speech Recognition*) se plantea la tarea inversa a la síntesis, ya que se pretende transformar una señal sonora continua –el habla– en su correspondiente representación simbólica discreta que, en general, será un texto escrito. El principal problema de los sistemas de reconocimiento radica en que, para alcanzar su objetivo, deben ser capaces de tratar la diversidad de voces, de acentos, de estilos de habla y de entornos en los que puede encontrarse un usuario.

En lo esencial, los reconocedores pueden considerarse sistemas que, en una primera etapa, aprenden automáticamente de un extenso corpus de habla y, en el momento de reconocer un nuevo enunciado, lo comparan con los datos o modelos que previamente han extraído de ese corpus. Por ello, el desarrollo de un sistema de reconocimiento se inicia con el diseño y la recolección de lo que se conoce como corpus de aprendizaje (o de entrenamiento), a partir del cual el sistema adquirirá la información necesaria para crear modelos acústicos –"plantillas" o representaciones internas– de cada una de las unidades de reconocimiento, análogas en ocasiones a las de la síntesis descritas en la sección 2.1. El corpus de entrenamiento se emplea también para obtener la gramática o modelo de lenguaje del reconocedor, entendida, de un modo muy simplificado, como un modelo que recoge las probabilidades de aparición de palabras en un determinado punto.

Al igual que un conversor, un sistema de reconocimiento de habla también se concibe como un conjunto de módulos. La señal sonora se analiza, en una primera etapa, para extraer los parámetros acústicos que se consideraron relevantes en el momento del diseño (Mariño y Nadeu, 2004), y después se compara, en el módulo de reconocimiento, con los modelos acústicos de las unidades 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, como hemos indicado, la probabilidad de las secuencias de palabras que pueden encontrarse en el contexto de una determinada aplicación. Un reconocedor incorpora también un diccionario de pronunciación en que se encuentran transcritas fonéticamente las palabras que puede aceptar el sistema, tanto en su forma canónica como en las variantes que se documentan en el corpus de entrenamiento (véase la sección 6.2).

En el campo de las tecnologías del habla se investiga también sobre la comprensión de la lengua oral (SLU, *Spoken Language Understanding*), que no debe confundirse con el reconocimiento. Si este último convierte un enunciado hablado en su representación escrita, con la comprensión se pretende avanzar un paso más y determinar el contenido del enunciado. Para lograr este objetivo, se emplean técnicas propias del procesamiento del lenguaje natural que se basan en los resultados del reconocimiento y que utilizan, en algunos casos, parte de la información fonética que éste proporciona.

### 14.2.3. *Los sistemas de diálogo*

Mediante los sistemas de diálogo o sistemas conversacionales (SLS, *Spoken Language Systems*), se pretende facilitar la interacción oral entre una persona y un sistema informático (Waibel, 2001; Tapias, 2002; Dahl, 2004; Minker y Bennacef, 2004; López-Cózar y Araki, 2005). La principal aplicación de los sistemas de diálogo se encuentra, por el momento, en los servicios telefónicos que permiten obtener información o realizar transacciones sin la presencia de un operador humano. Por esta razón, el diseño de un sistema de diálogo comienza definiendo un dominio de aplicación y analizando interacciones auténticas entre personas, por ejemplo, grabaciones de llamadas a un servicio de venta de entradas. 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, en muchas ocasiones se recurre al procedimiento conocido como el "Mago de Oz". En este caso, la persona que realiza una llamada en la etapa de adquisición del corpus escucha una voz sintetizada que le proporciona las respuestas a sus consultas; estas respuestas las decide, en función de un conjunto de escenarios previamente establecidos, un investigador que sigue la conversación –el "mago"– y que envía a un conversor de texto en habla los mensajes más adecuados a cada situación. El corpus así obtenido proporciona unos datos más realistas que permiten refinar el diseño del sistema.

Un sistema de diálogo realiza su tarea mediante un conjunto de módulos. El primero es un reconocedor automático del habla (con una estructura como la presentada en 2.2), que procesa las preguntas del usuario y convierte la señal sonora en una representación simbólica accesible al sistema informático. A continuación, se efectúa la interpretación semántica del enunciado, tarea que corre a cargo

del módulo de comprensión (véase 4.3), que intenta identificar el contenido de la petición del usuario. Un tercer módulo genera un enunciado completo (4.2) que contiene los resultados de la consulta a una base de datos con la información relevante o que, en su caso, solicita al usuario que confirme un dato o proporcione una información adicional. Finalmente, un conversor de texto en habla (como los descritos en 2.1) se encarga de transformar los resultados del módulo de generación en su equivalente sonoro. Las tareas de estos módulos están coordinadas por lo que se conoce como un "gestor del diálogo" que establece, por ejemplo, los turnos de palabra, y que pone 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.

En la actualidad, existe un gran interés por aumentar las prestaciones de los sistemas conversacionales integrando otros tipos de información, especialmente visual –los gestos o las expresiones faciales–, aunque no se excluyen opciones como el empleo de pantallas táctiles o de otras modalidades de interacción. Los sistemas de diálogo multimodales (Granström *et al.*, 2002; Kuppevelt *et al.*, 2005; Minker *et al.*, 2005) suponen pues un paso más para facilitar el uso de los ordenadores en situaciones cotidianas.

## 14.3. Las aplicaciones de las tecnologías del habla

Las tecnologías del habla encuentran su aplicación en muchos contextos relacionados, en mayor o menor medida, con nuestra actividad diaria. Ya ha dejado de ser una novedad encontrar un sistema de diálogo al otro lado del teléfono cuando, por ejemplo, llamamos a un servicio de reserva de billetes y, para algunos profesionales como médicos o traductores, el dictado automático es una herramienta de gran utilidad. Empresas españolas como Telefónica I+D han sido pioneras en el desarrollo de una amplia gama de aplicaciones que, en la actualidad, se encuentran al alcance de sus clientes, tanto en España como en Latinoamérica (Villarrubia *et al.*, 2002 y 2003). En esta sección se presentan algunos campos de aplicación de las tecnologías del habla –dictado automático, sistemas conversacionales, traducción automática del habla, recuperación de información a partir de documentos sonoros y reconocimiento automático del locutor y de la lengua– y se mencionan también algunos de los logros conseguidos en lo que se refiere al español. Por razones de espacio, no se tratan las

aplicaciones más orientadas a personas con necesidades especiales o que sufren algún tipo de discapacidad, pese a que éste es uno de los ámbitos de mayor interés humano y social (Aguilera *et al.*, 2001).

### 14.3.1. *El dictado automático*

Si bien el reconocimiento del habla puede emplearse para el control de sistemas –p. ej., de un ordenador, al sustituir los menús y las acciones del ratón, de ciertas funciones de un vehículo o de algunos elementos del entorno doméstico– con palabras aisladas o comandos breves, la aplicación más conocida es tal vez el dictado automático, con el que un usuario puede dictar un texto mediante un micrófono conectado al ordenador y éste queda almacenado en un programa de tratamiento de textos. Se trata de una posibilidad especialmente útil cuando se realizan a la vez otras tareas que ocupan las manos, cuando la entrada de datos mediante un teclado no es muy cómoda –sería el caso de los ordenadores de mano o de los teléfonos celulares– o, naturalmente, para personas con limitaciones de movilidad.

En español existen, al menos, dos productos comerciales de dictado automático para ordenadores personales: ViaVoice, desarrollado por IBM con la colaboración de su Centro de Tecnología de la Lengua en Sevilla, actualmente distribuido por ScanSoft, y Dragon Naturally Speaking de ScanSoft. ViaVoice se comercializa para el español en la versión 8 para Windows, mientras que las versiones posteriores, tanto para Windows como para Mac OS X, se encuentran, por el momento, únicamente en inglés, alemán, italiano y japonés. En ambos sistemas es posible incorporar nuevas palabras que se graban con la voz del usuario y se almacenan; Dragon Naturally Speaking cuenta también con diccionarios adicionales a los que se ha incorporado vocabulario médico y legal. Se trata de programas que, por razones ya explicadas (sección 2.2), requieren un entrenamiento por parte del usuario, por lo que cada persona que lo emplea debe definir su propia identidad. En general, las prestaciones suelen mejorar cuanto más se emplea el sistema.

### 14.3.2. *Las interfaces conversacionales*

Como se ha señalado en la sección 2.3, los sistemas de diálogo son útiles en todas aquellas situaciones en las que se desea obtener una

información o realizar una transacción a través del teléfono, sin contar con la presencia de un operador humano (Tapias y Hernández, 2004). Se emplean, pues, en los servicios de atención al cliente, en la banca electrónica, para la venta de billetes o de entradas y en otros servicios relacionados con el comercio electrónico; constituyen también la base de los llamados "portales de voz", equivalentes telefónicos de los portales de Internet.

Existen en España diversas empresas especializadas en el desarrollo y la integración de sistemas conversacionales, entre las que pueden citarse Atlas, Grupo Voice, InfoSpeech, NaturalVox, Porfinya, Telefónica I+D, VoxSmart o Ydilo. Por otra parte, sistemas pioneros como Saplen y proyectos recientes como Basurde, TelCorreo, Diana o Gemini han dado como resultado prototipos de sistemas de diálogo en dominios como la venta de billetes de tren, la consulta del correo electrónico o la banca telefónica, y han contribuido al avance de este sector desde el entorno universitario.

### 14.3.3. *La traducción automática del habla*

El auge de los mercados internacionales y las posibilidades abiertas por los sistemas de diálogo ha planteado un nuevo reto: hacer posible la traducción automática en tiempo real de conversaciones telefónicas (Wahlster, 2000; Waibel, 2000; Casacuberta, 2004). La traducción de la lengua oral (SLT, *Spoken Language Translation*) requiere combinar un reconocedor automático del habla para procesar los enunciados de cada uno de los interlocutores, un módulo de traducción automática, un gestor del diálogo y un conversor de texto en habla, de modo que el resultado de la traducción sea accesible oralmente.

Algunas de las dificultades lingüísticas que se plantean en la traducción automática del habla aparecen también en los sistemas conversacionales, y radican en las llamadas "disfluencias" o discontinuidades, fenómenos propios del habla espontánea que se manifiestan en forma de elementos vocales como "eh" o "mm", falsos principios, repeticiones y construcciones a menudo muy alejadas de las de la lengua escrita. Por ello, algunos de los enfoques para la traducción del habla se basan en técnicas estadísticas y en el aprendizaje a partir de corpus en varias lenguas, en lugar de recurrir a los sistemas tradicionales de traducción automática (véase el apartado 5.2).

Entre los proyectos que incluyen el español cabe destacar EuTrans, SisHiTra y Ametra, desarrollados en la Universidad Politécnica de Valencia, o TC-STAR y FAME, llevados a cabo en la Politécnica de Cataluña. Hasta la fecha, los resultados consisten en prototipos que se centran en dominios restringidos como las reservas de alojamiento o las consultas a la recepción de un hotel.

### 14.3.4. *La recuperación de información a partir de documentos sonoros*

Los medios de comunicación orales disponen hoy en día de grandes archivos de grabaciones digitales; puesto que no es posible, por motivos de tiempo y de economía, transcribir su contenido, se han propuesto técnicas basadas en el reconocimiento del habla y en la recuperación de información (véase el apartado 5.3) que facilitan el acceso automático a documentos sonoros (SDR, *Spoken Document Retrieval*). Desde el punto de vista de las tecnologías del habla, parte de la dificultad de la tarea se debe a las condiciones de la grabación –p. ej., puede contener música de fondo, tratarse de una retransmisión en directo en un ambiente ruidoso o presentar intervenciones simultáneas de más de un hablante–, a la diversidad de locutores y a los problemas que plantea el reconocimiento del habla espontánea. En este sentido, en la Universidad de Vigo se está llevando a cabo un proyecto de transcripción e indexación de programas de noticias, Transcrigal, orientado precisamente hacia futuros sistemas de recuperación de información hablada.

### 14.3.5. *La identificación y verificación automáticas de la identidad del locutor y la identificación automática de la lengua*

La identificación automática de una persona a través de su voz y la verificación de su identidad por el mismo procedimiento son, probablemente, dos de las aplicaciones de las tecnologías del habla que más atraen la atención del gran público. El reconocimiento automático del locutor (ASR, *Automatic Speaker Recognition*) (Nolan, 1997; Rodríguez *et al.*, 1998) es necesario, por ejemplo, para efectuar transacciones bancarias por teléfono sin tener que recurrir a números personales de identificación; se plantea también como un procedimiento que sustituya las contraseñas al acceder a determinadas instalaciones o a sistemas informáticos y, en conjunto, representa un

sector destacado en el desarrollo de las técnicas biométricas para la identificación de personas. En el contexto legal ha adquirido también un papel muy relevante, pues permite incorporar una cierta objetividad a las decisiones que, hasta ahora, estaban en manos de expertos en fonética judicial (Hollien, 2002; Rose, 2002). La identificación automática de la lengua en la que se expresa un determinado hablante (LID, *Automatic Language Identification*) es también un problema que ha atraído la atención de los expertos, puesto que permite ofrecer servicios multilingües sin necesidad de que el usuario tenga que indicar explícitamente la lengua que desea emplear (Geoffrois, 2004).

En español contamos con productos comerciales de reconocimiento del locutor (CoreVox, de la empresa Agnitio, vinculada a la Universidad Politécnica de Madrid) y con desarrollos realizados tanto en empresas (Telefónica I+D) como en grupos universitarios que se han especializado en este ámbito en Barcelona, Madrid, Mataró o Vigo. Para aplicaciones judiciales existen también sistemas automáticos como BatVox (Agnitio) o IdentiVox. La Sociedad Española de Acústica Forense (SEAF), cuyos congresos, iniciados en el año 2000, se celebran cada tres años, reúne a buena parte de los expertos del país en este campo.

## 14.4. Las tecnologías del texto

Las que aquí denominamos tecnologías del texto son las que se centran, como hemos indicado, en el tratamiento de la lengua escrita, y suelen englobarse habitualmente bajo el término "procesamiento del lenguaje natural" (PLN o NLP, *Natural Language Processing*). Quizás no deje de ser pertinente recordar que, para los lingüistas, el lenguaje siempre es un fenómeno "natural", que se puede manifestar tanto en su forma oral –la primaria–, como escrita, en cierto modo subsidiaria de la hablada. El adjetivo "natural" obedece a que el PLN nació estrechamente ligado a la inteligencia artificial, y se hacía necesario distinguir el lenguaje humano de los lenguajes de programación. Debido a la separación inicial entre estos estudios y los que se ocupan de la lengua hablada –surgidos de la mano de la ingeniería de las telecomunicaciones–, se ha consolidado una distinción entre "lenguaje" (*language*) y "habla" (*speech*), que ha separado a los expertos que se ocupan básicamente del texto de los que lo hacen de la señal sonora. Sin embargo, esta dicotomía se difumina cada

vez más en la actualidad, con el empleo de técnicas estadísticas comunes y con la aparición de aplicaciones como los sistemas de diálogo que requieren el uso conjunto de conocimientos sobre el texto y sobre el habla.

En el campo de las tecnologías del texto, podríamos distinguir entre las herramientas con las que se procesa la lengua escrita, que se presentan en la sección 4.1, y las tecnologías empleadas en el desarrollo de aplicaciones, descritas en las secciones siguientes: generación del lenguaje (4.2) por una parte, y comprensión del lenguaje (4.3) por otra. Las aplicaciones de las tecnologías del texto se abordan en la sección 5[4].

### 14.4.1. *Herramientas de análisis lingüístico*

La primera operación que suele llevarse a cabo en el procesamiento de la lengua escrita es la lematización, es decir, la detección automática del radical de una palabra. Un lematizador es, por lo tanto, una herramienta que asocia una forma flexionada o derivada de una palabra con su correspondiente lema –la forma que aparece en las entradas de los diccionarios–, separando la raíz de los afijos. El siguiente paso en un tratamiento computacional sería el análisis morfológico completo, en el que se indicarían la categoría léxica de la palabra y las categorías gramaticales (género, número, persona, tiempo, modo, etc.) representadas en cada uno de sus morfos. Este análisis se refleja en los resultados que proporcionan los etiquetadores morfológicos (*tagger* o POS, *part of speech tagger*), empleados también, por ejemplo, en el tratamiento de corpus textuales (véase 6.1). Estos etiquetadores pueden basarse en reglas o en probabilidades, aunque existen sistemas híbridos que combinan ambos tipos de aproximación; los que se fundamentan en probabilidades requieren, al igual que sucede en los sistemas de reconocimiento de habla, un entrenamiento a partir de un corpus correctamente etiquetado.

---

[4] El lector interesado puede consultar las presentaciones generales de Gómez (2000), Rodríguez (2000), Badia (2001) y Ruiz (2005), o manuales como los de Martí (2001) y Lavid (2005); para profundizar en sus conocimientos, obras como las de Cole *et al.* (1997), Moreno *et al.* (1999), Dale *et al.* (2000), Jurafsky y Martin (2000) o Coleman (2005) pueden resultarle de utilidad. En lo que respecta al uso de técnicas estadísticas para el procesamiento del lenguaje natural, véanse las reflexiones de Rodríguez (2004) o el manual, ya de un nivel avanzado, de Manning y Schütze (1999).

Un analizador sintáctico (*parser*) ofrece la estructura de constituyentes de una oración, representada, por lo general, en forma de árboles con los nodos etiquetados, tal como suele hacerse habitualmente en sintaxis. Para ello es preciso disponer de un conjunto de reglas, similares a las reglas de reescritura propias de la gramática generativa, que se aplican, de modo ascendente o descendente, para determinar las relaciones jerárquicas entre los constituyentes de la oración y las categorías y funciones sintácticas de cada elemento (Rodríguez, 2002). El interés por disponer de corpus etiquetados sintácticamente –los llamados *treebanks*, tratados en la sección 6.1– ha contribuido en los últimos años a notables avances en el campo del análisis sintáctico automático. El principal problema al que tienen que enfrentarse tanto los analizadores morfológicos como los sintácticos es la ambigüedad (Civit *et al.*, 2002). En el nivel morfológico y léxico, por ejemplo, "casas" puede ser tanto un nombre como un verbo, y "bajo" puede ser nombre, adjetivo, adverbio o preposición; en el sintáctico, son conocidas ambigüedades del tipo "Juan vio a su amigo nadando", "Habló a sus alumnos de lingüística" o "María está pensando en su casa". Se ha propuesto que los enfoques basados en la probabilidad podrían favorecer el rendimiento de los analizadores en estos casos (Ruiz, 2005).

El análisis semántico automático presenta una mayor complejidad, ya que su objetivo es crear una representación abstracta que permita alcanzar un cierto grado de "comprensión" del enunciado, estableciendo las relaciones de significado entre los elementos léxicos. Suele recurrirse a representaciones basadas en la lógica formal, aunque en casos en los que el dominio de la aplicación es muy restringido –p. ej., la consulta a una base de datos–, se emplean procedimientos más simples pero únicamente válidos para una aplicación en concreto. El análisis semántico se enfrenta también al problema de la ambigüedad, tanto en lo que se refiere a la polisemia como a otros fenómenos que se relacionan con los cuantificadores –p. ej., el clásico "Todos los hombres aman a una mujer"– o la referencia (Ruiz, 2005). Un analizador semántico necesita apoyarse en recursos léxicos como las redes léxico-semánticas o las ontologías, descritas en la sección 6.2, y es también una herramienta que se emplea para la anotación semántica de corpus (6.1).

Finalmente, las herramientas de análisis pragmático son necesarias en aplicaciones que implican una interacción entre el usuario y un ordenador, como veíamos en el caso de los sistemas de diálogo. Dos de los principales problemas lingüísticos son las anáforas

y la identificación del valor de los actos de habla. En lo que respecta a la anáfora, debe tenerse en cuenta que, en el curso de un diálogo, puede suceder que el referente y su correspondiente elemento anafórico no se encuentren en el mismo enunciado; en cuanto a los actos de habla, es bien sabido que lo que formalmente constituye una enunciativa –p. ej., "Quiero saber los horarios de trenes entre Madrid y Barcelona"– debe interpretarse, en realidad, como una petición, por lo que un sistema informático tiene que ser capaz de detectar cuál es el objetivo del usuario y actuar en consecuencia.

En lo que se refiere al español, las páginas del CLiC de la Universidad de Barcelona ofrecen ejemplos interactivos de lematizadores, analizadores morfológicos y analizadores sintácticos; en las del GEDLC de la Universidad de Las Palmas de Gran Canaria, pueden encontrarse herramientas en línea de análisis y de generación en el nivel morfológico, y en las del IULA de la Universitat Pompeu Fabra se muestran sistemas como PALIC y AMBILIC. También la empresa ecuatoriana Signum, la finlandesa Connexor y la multinacional estadounidense Xerox proporcionan algunas muestras de herramientas en línea. Existen igualmente para el español otras herramientas de análisis morfológico y sintáctico, desarrolladas principalmente para la anotación de corpus o para integrarse en plataformas para el procesamiento del lenguaje; el lector interesado puede consultar las actas de los congresos de la SEPLN (Sociedad Española para el Procesamiento del Lenguaje Natural), en las que encontrará abundante información sobre este tipo de herramientas. Desde la perspectiva de la enseñanza de la gramática, el proyecto VISL (*Visual Interactive Syntax Learning*) ofrece una buena muestra de las posibilidades didácticas de los analizadores sintácticos combinados con otros recursos.

### 14.4.2. *La generación del lenguaje*

La generación del lenguaje (NLG, *Natural Language Generation*) es una técnica que permite la creación automática de textos escritos a partir de una representación conceptual (Lavid, 2005). En este sentido, el módulo de generación de respuestas de un sistema de diálogo (sección 2.3) sería una aplicación particular de la técnica que aquí abordamos. Sin embargo, la generación se ha empleado principalmente para otros tipos de textos como partes meteorológicos, informes o documentación técnica.

Un sistema de generación de lenguaje consta, en general, de tres módulos dedicados, respectivamente, a la macroplanificación, la microplanificación y la realización superficial (Lavid, 2005). El primero de ellos se ocupa de determinar el contenido del texto y de estructurarlo para que pueda comprenderse adecuadamente; para ello es necesario contar con conocimientos sobre el dominio de la aplicación, pero también es esencial disponer de información sobre el usuario del sistema y sobre la finalidad –es decir, la intención comunicativa– de los textos. El componente centrado en la microplanificación, en cambio, se encarga de estructurar las oraciones que formarán el texto, combinando varios mensajes en una oración, seleccionando el léxico apropiado y estableciendo las relaciones de correferencia entre elementos. Finalmente, en la etapa de realización superficial se aplican las reglas gramaticales que dan como resultado oraciones bien formadas y se establece la forma final del texto. Como argumenta Badia (2001), uno de los principales problemas de la generación reside en que el contenido de una misma representación puede manifestarse en la lengua de diversos modos; al igual que un mismo acto de habla puede realizarse de maneras muy diversas, la expresión, por ejemplo, de la impersonalidad, puede también llevarse a cabo a través de distintos procedimientos gramaticales.

Existen diversos métodos para la generación de textos (Lavid, 2005), desde los que contienen mensajes en cierto modo "prefabricados" o emplean plantillas cuyo contenido se modifica ligeramente, los que se basan en patrones de frases y de textos, hasta los sistemas que emplean rasgos de tipo sintáctico-semántico. Como bien puede suponerse, los primeros son los más simples pero los más dependientes de la aplicación –serían, en este sentido, equivalentes a los sistemas de síntesis con mensajes pregrabados–, mientras que los últimos son mucho más generales aunque, a la vez, lingüística y computacionalmente más complejos.

### 14.4.3. *La comprensión del lenguaje*

Cuando los lingüistas computacionales se refieren a comprensión del lenguaje (NLU, *Natural Language Understanding*), emplean el término "comprensión" en un sentido restringido y, generalmente, centrado en un dominio concreto. El proceso que se realiza en este caso consiste en extraer de un texto escrito una representación abstracta del contenido que contenga la información necesaria para realizar

otras operaciones; en el contexto, por ejemplo, de una interfaz en lenguaje natural que proporcione información sobre horarios de trenes, la "comprensión" se limita a determinar la estación de origen, la de llegada, la fecha y la franja horaria en que el usuario desea viajar, que son los elementos imprescindibles para proporcionar una respuesta. Como es lógico, la comprensión depende fundamentalmente de las herramientas de análisis morfológico, sintáctico y, especialmente, semántico descritas en la sección 4.1, así como también del desarrollo de gramáticas (6.3) y de diccionarios (6.2).

## 14.5. Las aplicaciones de las tecnologías del texto

Las tecnologías básicas sucintamente expuestas en la sección anterior encuentran múltiples aplicaciones en todos aquellos casos en los que se requiere el tratamiento automático de textos escritos. Mencionaremos, en primer lugar, las herramientas de ayuda a la escritura; a continuación describiremos brevemente los fundamentos de la traducción automática, y por último nos referiremos a las aplicaciones relacionadas con la información contenida en los textos.

### 14.5.1. *Herramientas de ayuda a la escritura*

Una de las aplicaciones más extendidas de las tecnologías lingüísticas son los correctores ortográficos y gramaticales, que se encuentran en la mayoría de los procesadores de textos, y que pueden describirse genéricamente como herramientas de ayuda a la escritura. Es frecuente distinguir, en este ámbito, tres niveles de complejidad creciente: verificación ortográfica, verificación gramatical y verificación de estilo (Gómez, 2000 y 2001). Los correctores ortográficos presentan aún ciertas limitaciones, principalmente debidas a que, en muchas ocasiones, no incorporan información lingüística y se limitan a comparar cadenas de caracteres con palabras almacenadas en un diccionario. A la hora de proponer una alternativa, el programa de corrección busca en el diccionario posibles formas existentes, partiendo de la base de que puede haberse insertado o eliminado una letra, cambiado una letra por otra o intercambiado dos letras consecutivas (Gómez, 2000). Lógicamente, la posibilidad de que se encuentre la palabra adecuada depende de un modo muy directo de la riqueza del diccionario.

En casos como, por ejemplo, "Los alumnos estudia", es probable que un verificador ortográfico no señale ningún error, ya que la forma "estudia" se encontraría en el diccionario. Para resolver problemas de este tipo es preciso recurrir a un corrector gramatical o verificador sintáctico. Los sistemas actuales se basan, generalmente, en la comparación de secuencias de palabras con unos patrones de errores que se han determinado antes; en ocasiones, establecer patrones que tengan una cierta validez general –de lo contrario, debería contarse con un patrón por cada posible error– es una operación compleja que requiere un cierto grado de abstracción lingüística (Gómez, 2000).

Finalmente, los verificadores de estilo comprueban la adecuación del texto a un conjunto de reglas previamente definidas. Una vez que el usuario ha establecido el tipo de texto que desea corregir –general, técnico, literario, etc.–, el corrector detecta aquellos elementos que no se corresponden con los rasgos válidos 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 la red se hallan demostraciones de correctores ortográficos en español como, por ejemplo, el de la empresa Signum, proveedora de Microsoft, DataSpell de Bitext, o el que se integra en el conjunto de herramientas COES de la Universidad Politécnica de Madrid. Por otra parte, se comercializan programas de corrección ortográfica, gramatical y de estilo como Stilus, de Daedalus –del que puede verse una demostración en línea en el sitio de la empresa–, y herramientas de revisión como la de Signum, que puede descargarse para un periodo de prueba.

### 14.5.2. *La traducción automática*

La traducción automática (TA o MT, *Machine Translation*) fue, ya desde los años cincuenta, una de las primeras aplicaciones que se intentaron abordar en el campo del procesamiento informático del lenguaje (Alonso, 2001; Abaitua, 2002a). En la actualidad es, tal vez, una de las más populares, pero pese a su difusión, sigue siendo una tecnología sujeta a ciertas limitaciones en lo que se refiere a la gama de contenidos que pueden traducirse con éxito, y a las dificultades que plantea la traducción entre pares de lenguas tipológicamente muy

lejanas. 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 por ordenador (TAO o CAT, *Computer Assisted Translation*) que mejoran notablemente la labor del traductor.

Los problemas de la traducción automática son, como indica Alonso (2001), los mismos que surgen a la hora de comprender un enunciado: por una parte, se necesitan conocimientos morfológicos, sintácticos, léxicos y semánticos, mientras que, por otra, es imprescindible en ciertos casos contar con lo que se denomina el conocimiento del mundo, información que difícilmente puede formalizarse, por el momento, en su totalidad. La combinación de todas estas fuentes de información contribuye a deshacer las potenciales ambigüedades de un texto (véase la sección 4.1); pero otro de los aspectos que dificultan la traducción automática es la diferencia estructural entre las lenguas. Compárense, por ejemplo, las construcciones "Juan cruzó el río a nado" y "John swam across the river" (Lavid, 2005), en las que se pone de relieve que es preciso recurrir a representaciones con un elevado nivel de abstracción para lograr un tratamiento adecuado de estos casos.

Entre las estrategias para la traducción automática suelen distinguirse la traducción directa, la transferencia y la interlingua. La traducción directa, que recurre únicamente a léxicos monolingües y bilingües, ofrece, como es de esperar, 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. En la traducción automática mediante 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 definidas en 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 con la ayuda de un léxico bilingüe, teniendo también en cuenta toda la información estructural acumulada durante el análisis. Al final, en la fase de generación, se convierten los resultados de la transferencia en oraciones gramaticalmente aceptables en la lengua de destino (Alonso, 2001). Los sistemas que utilizan lo que se conoce como interlingua basan la traducción en una representación abstracta del significado, extraída durante la fase de análisis, y utilizada como base para la generación. La principal dificultad estriba, por tal motivo, en la repre-

sentación 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.

Al igual que en otros campos del procesamiento del lenguaje natural, se han introducido también, en los últimos años, técnicas estadísticas para intentar que la traducción automática no dependa de un elaborado y complejo sistema de reglas. Para el entrenamiento del sistema debe disponerse de corpus paralelos alineados (véase 6.1), de modo que puedan calcularse las probabilidades de que una oración dada en la lengua de llegada sea la traducción de otra que se encuentra en un texto en la lengua de origen, creándose así el llamado modelo de traducción. Como hemos mencionado en referencia a otros ámbitos, hoy en día se tiende a desarrollar sistemas híbridos que complementan el conocimiento lingüístico con métodos estadísticos. También se utiliza la llamada traducción basada en ejemplos, fundamentada en corpus paralelos alineados que constituyen una "memoria de traducción" en que se buscan los enunciados que deben traducirse.

Existe en el mercado un gran número de sistemas de traducción automática y de herramientas de traducción asistida o TA (Hutchins, 2003 y 2005). Entre los sistemas de TA que incluyen el español y cuentan con demostraciones en la web pueden señalarse, además de los mencionados en el capítulo de Mar Cruz Piñol, los de Comprendium, AutomaticTrans o InterNOSTRUM, desarrollados en España, y Systran –incorporado a las herramientas lingüísticas de Google y al traductor Babel Fish de Altavista–, Reverso, o WebSphere Translation Server, creados por empresas europeas o americanas.

### 14.5.3. *La recuperación y extracción de información y la respuesta a preguntas*

El crecimiento de la información disponible en la web y la digitalización cada vez más habitual de grandes fondos documentales han creado la necesidad de disponer de un acceso automático a los datos y a los documentos, puesto que su volumen hace imposible una búsqueda manual. Las técnicas de recuperación y de extracción de información, así como las de respuesta a preguntas, algunas de las

cuales incorporan elementos tomados del procesamiento del lenguaje natural, aportan una respuesta a este problema y, por ello, son unas de las áreas que más atención reciben actualmente en el ámbito de las tecnologías lingüísticas.

La recuperación de información (RI o IR, *Information Retrieval*) consiste en seleccionar, en un conjunto de documentos, aquellos que contienen la información que un usuario solicita mediante una consulta (Gonzalo y Verdejo, 2001). Un ejemplo de la aplicación de esta técnica se encuentra en los buscadores más conocidos en Internet, que proporcionan un listado de páginas potencialmente relevantes en función de las palabras utilizadas en la búsqueda. Pese a que la recuperación de información ha sido un campo tradicionalmente alejado del procesamiento del lenguaje natural, algunas empresas que desarrollan sistemas de búsqueda y recuperación de información emplean ya en sus productos comerciales técnicas lingüísticas, y es previsible que su uso se incremente en el futuro (Molano, 2002), especialmente con la incorporación de recursos como las redes léxico-semánticas descritas en la sección 6.2 (Vossen, 2001). La introducción de estas técnicas viene también determinada por el requisito de tratar documentos en más de una lengua. Así, 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 que haya realizado su consulta (López *et al.*, 2004).

Mucho más compleja que la recuperación es la extracción de información (IE, *Information Extraction*). La finalidad de la búsqueda, en este caso, no es sólo 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. La extracción de información se efectúa a partir 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 específico. 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 a la persona que ha realizado la consulta (Gonzalo y Verdejo, 2001). Estos sistemas deben tratar problemas lingüísticos de naturaleza muy diversa, entre los que destacan el reconocimiento de los nombres propios, la asignación de correferencia y, en relación con ésta última, las anáforas.

Los métodos de búsqueda de respuestas (BR o QA, *Question Answering*) representan un paso más en las tecnologías que permiten obtener información de un amplio conjunto de textos (Vicedo *et al.*, 2003; Ferrández, 2004). En este caso, el usuario formula una pregunta concreta, que el sistema analiza para extraer las palabras clave y para realizar una selección de los documentos que las contienen; a partir de estos datos se construye una respuesta que incluye, ordenadas jerárquicamente, las secciones más relevantes de los documentos. Portales de Internet como Ask.com o Answers.com ilustran los usos de esta tecnología.

En el contexto español se han desarrollado proyectos interuniversitarios como ITEM, HERMES o R2D2 centrados en la recuperación y extracción de información y en la recuperación de documentos; las demostraciones en línea del buscador multilingüe ITEM y del sistema WTB (*Website Term Browser*) constituyen una muestra de los resultados obtenidos. Finalmente, cabe señalar entre las aplicaciones relacionadas con el tratamiento y la gestión de la información, el resumen automático de documentos (Climent, 2001; Lavid, 2005). Como primera aproximación a esta tecnología, resulta útil realizar pruebas en línea con el programa SweSum, que incluye el español entre las lenguas de trabajo.

## 14.6. Los recursos lingüísticos

Los recursos lingüísticos (RL o LR, *Language Resources*) representan, como hemos señalado en diversas ocasiones, un elemento esencial para el desarrollo de las aplicaciones propias de las tecnologías del lenguaje. Habitualmente, se agrupan en tres grandes categorías: corpus, recursos léxicos y gramáticas, aunque estas últimas pueden considerarse también herramientas para el análisis o para la generación de textos.

### 14.6.1. *Los corpus*

Un corpus puede definirse como un conjunto estructurado de textos que forman una muestra representativa del uso real de la lengua (Torruella y Llisterri 1999; Rafel y Soler, 2001). Sin embargo, cualquier colección de materiales no da pie por sí misma a un corpus si no cumple una serie de requisitos: un diseño coherente,

la presencia en los textos de marcas 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 contexto de las tecnologías lingüísticas, 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.

La lingüística de corpus (*Corpus Linguistics*) es la disciplina que se dedica a la creación y la explotación de los recursos lingüísticos escritos y orales. Aunque uno de sus principales cometidos es la descripción y el análisis de la lengua con una base empírica, en esta sección nos centraremos en los corpus empleados en el desarrollo de las tecnologías lingüísticas (McEnery, 2003)[5].

Existen básicamente dos tipos de corpus: orales y escritos. Entre los primeros, se distinguen los que recogen la grabación de la señal sonora, denominados corpus orales o bases de datos orales (*speech corpora, speech databases*), de los corpus de lengua oral (*spoken language corpora*), que consisten en transcripciones ortográficas de la lengua hablada. Esto no implica 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 recursos creados por los especialistas en tecnologías del habla, centrados en la onda sonora y en el dominio para el que se diseña una determinada aplicación (Schiel *et al.*, 2004), y los que se recogen desde una perspectiva más lingüística para el análisis de las manifestaciones orales de la lengua (Moreno Fernández, 1997). Sin embargo, en la actualidad se recurre a grandes corpus transcritos ortográficamente para el entrenamiento de los modelos lingüísticos de los sistemas de reconocimiento de habla, 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 a través de los elementos suprasegmentales. En los últimos años se han realizado esfuerzos importantes para reco-

[5] Los manuales de Biber *et al.* (1998), Kennedy (1998), McEnery y Wilson (2001) o Berber Sardinha (2004), entre otros, ofrecen una panorámica de la lingüística de corpus más orientada hacia los aspectos lingüísticos. Para más información sobre los corpus generales en español, remitimos al lector al capítulo 5 de esta obra.

ger corpus en los que a la señal sonora se añade una grabación en vídeo del locutor, sea únicamente de la cara, sea de todo el cuerpo, para poder así desarrollar los sistemas multimodales a los que aludimos en la sección 2.3.

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, el entrenamiento de los modelos acústicos para el reconocimiento, o el diseño de escenarios para un sistema de diálogo. En función de su finalidad, se establece el diseño del corpus y se definen los distintos niveles de etiquetado (Llisterri *et al.*, 2005). Así, en el caso de la síntesis y del reconocimiento se marcan, sincronizadas con la onda sonora y, en la medida de lo posible, con la representación ortográfica, las fronteras entre los segmentos y entre las unidades que se hayan establecido; es también útil efectuar un etiquetado prosódico, o diferenciar un nivel fonético y un nivel fonológico de etiquetado. En los corpus para desarrollar sistemas de diálogo suele incluirse, además, una anotación pragmática que refleja los actos de habla y otros aspectos esenciales en la interacción.

La anotación de los recursos multimodales plantea, por su parte, la necesidad de emplear sistemas de descripción de las expresiones faciales y de los gestos, tradicionalmente abordados desde otros campos, pero que han despertado un gran interés entre quienes se dedican a las tecnologías del habla (Serenari *et al.*, 2002). Por este motivo, junto a las herramientas de dominio público que permiten el etiquetado segmental y suprasegmental como Praat o Wavesurfer, se emplean cada vez más programas para la transcripción y anotación conjunta de audio y vídeo como Anvil, ELAN o Transana, también disponibles gratuitamente en la red[6].

Un corpus escrito es una colección estructurada de textos en la que se han introducido marcas que definen su estructura, y que se ha enriquecido con anotaciones relativas a su contenido lingüístico. Para el primer propósito, existen desde hace tiempo estándares

---

[6] La situación de los corpus orales para las tecnologías del habla en español se describe en Llisterri *et al.* (2005); cabe señalar que, dada la aplicación de estos recursos para el desarrollo de productos comerciales, no son abundantes los corpus a los que pueda accederse con facilidad. De hecho, algunos se distribuyen a través de ELDA (*Evaluations and Language resources Distribution Agency*) y del LDC (*Linguistic Data Consortium*) con precios diferentes según se trate de grupos de investigación o de empresas.

como los de la *Text Encoding Initiative* (TEI) (Sperberg-McQueen y Burnard, 2002), basados en el uso del SGML (*Standard Generalized Markup Language*) y, más recientemente, del XML (*Extensible Markup Language*), del que también se ha realizado una adaptación para las tecnologías del habla conocida como VoiceXML. 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.

En el campo de las tecnologías lingüísticas, este tipo de codificación es prácticamente imprescindible; pero también es necesario que los corpus escritos estén adecuadamente anotados desde un punto de vista lingüístico. La anotación o etiquetado se realiza en el nivel morfológico, sintáctico, semántico, pragmático o textual, utilizando, cuando están disponibles, herramientas como las descritas en la sección 4.1 (Sánchez *et al.*, 1999; Civit, 2003). En realidad, para que una herramienta de anotación automática realice su tarea de un modo eficaz, se requiere una etapa de entrenamiento con un corpus anotado manualmente. Por otra parte, las anotaciones automáticas requieren una revisión manual a cargo de un experto para detectar los errores que haya podido cometer el sistema 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.

Un tipo de corpus textual anotado sintácticamente lo forman los llamados *treebanks* o bancos de árboles sintácticos, en los que se marca la categoría y la función de los constituyentes de cada oración. Entre los recursos desarrollados para el español cabe mencionar el *treebank* de la Universidad Autónoma de Madrid y el proyecto 3LB, que incluye el catalán y el euskera incorporando, además, una anotación semántica. Realizada desde una perspectiva diferente y manualmente anotada, la BDS (Base de Datos Sintácticos del Español Actual) de la Universidad de Santiago de Compostela permite diversas consultas relacionadas con los esquemas sintácticos en los que aparece un determinado verbo. En lo que a la anotación semántica se refiere, además del corpus 3LB se cuenta también, por ejemplo, con MiniCors, creado en el marco de Senseval, una iniciativa internacional dedicada a la evaluación de herramientas automáticas para la desambiguación de palabras en la anotación de corpus escritos.

En el contexto de la traducción automática se encuentran los denominados corpus paralelos, que contienen el mismo texto en dos o más lenguas entre las que se establece una correspondencia entre segmentos equivalentes –frases, párrafos o textos– mediante un proceso conocido como alineación (Abaitua, 2002b). La existencia de estos recursos ha permitido abordar la traducción automática entrenando los sistemas con corpus paralelos alineados, a los que se aplican técnicas estadísticas similares a las empleadas en el reconocimiento del habla.

### 14.6.2. *Los recursos léxicos*

El desarrollo de las tecnologías lingüísticas requiere, además de los corpus, el uso de recursos léxicos, entre los que cabe mencionar los léxicos computacionales, monolingües o multilingües, y las llamadas redes léxico-semánticas. Existen también otros recursos electrónicos como los diccionarios en CD-ROM o en la web –véase el capítulo 5 de este volumen–, que en ocasiones establecen la base para la creación de los léxicos a los que hacemos referencia en este apartado.

Un léxico computacional (Martí, 1994; Vázquez *et al.*, 2002), a diferencia de los diccionarios convencionales, 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 4.1) y para la anotación de corpus textuales (6.1). Así, por ejemplo, en la entrada de un verbo se especificarían su categoría léxica y el tipo de elementos que puede subcategorizar. En el caso de un léxico multilingüe, cada lema llevaría asociada su correspondencia con los equivalentes en otras lenguas, y se indicarían también, si se emplea en traducción automática, las restricciones necesarias para lograr una buena traducción. En el ámbito de las tecnologías del habla se han desarrollado diccionarios de pronunciación (*pronunciation lexica*) que recogen todas las variantes encontradas en un corpus y las asocian a una forma canónica o a una representación fonológica con objeto de facilitar el reconocimiento.

Las redes léxico-semánticas son recursos que 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").

En este sentido, pueden considerarse también ontologías (Feliu *et al.*, 2002), ya que establecen una organización jerárquica de los conceptos, especialmente en el caso de los nombres. La iniciativa más conocida en el ámbito de las redes léxico-semánticas es WordNet, desarrollado en la Universidad de Princeton, del que se dispone de una versión en distintas lenguas europeas, conectadas entre sí, denominada EuroWordNet. La versión española de WordNet se llevó a cabo en colaboración entre la Universidad de Barcelona, la Politécnica de Cataluña y la Universidad Nacional de Educación a Distancia, y puede consultarse en las páginas del proyecto Meaning o en las del CLiC de la Universidad de Barcelona. Otro tipo de red, que actualmente se desarrolla para el español en la Universidad Autónoma de Barcelona, es el SFN (*Spanish FrameNet*), en que se integra información léxica y sintáctica en el marco de la semántica cognitiva.

Como ya se ha indicado, las redes léxico-semánticas y las ontologías encuentran uno de sus principales usos en la anotación semántica de corpus y muestran un importante potencial en las aplicaciones orientadas a la recuperación y extracción de información. Por otra parte, en el procesamiento del lenguaje natural se emplean también recursos terminológicos, monolingües o multilingües, que se extraen de corpus especializados en un determinado dominio.

### 14.6.3. *Las gramáticas computacionales*

Una gramática computacional se concibe como una descripción formalizada del conocimiento lingüístico que puede ser empleada tanto como una herramienta de análisis automático (véase la sección 4.1) como en el funcionamiento de algunas de las aplicaciones que hemos descrito anteriormente. Por esta razón se considera, junto con los corpus y los léxicos, un recurso para el desarrollo de las tecnologías del texto y, en ocasiones, del habla.

Muchos de los esfuerzos en el diseño de gramáticas computacionales se han centrado en encontrar el formalismo más adecuado para representar la información lingüística. Se han creado, para ellos, diversos procedimientos, entre los que destacan, en los últimos años, las gramáticas de unificación y las gramáticas de restricciones (*Constraint Grammars*). 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 las posibles funciones sintácticas de una palabra, para realizar después una desambiguación y seleccionar la función adecuada en una oración concreta (Balari, 1999; Badia, 2001; Rodríguez, 2002).

Existen otras 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).

## 14.7. Consideraciones finales

En las tecnologías lingüísticas confluyen, como se ha intentado poner de relieve a lo largo de este capítulo, saberes muy diversos, lo que hace inevitable que su desarrollo se efectúe en el marco de equipos pluridisciplinares. El especialista en lingüística puede y debe tener un papel activo en el campo de las tecnologías del lenguaje, puesto que éstas constituyen una de las ramas de la lingüística aplicada.

En un momento en que se debate la utilidad de la lingüística, no debería olvidarse el interés social y el potencial económico de las tecnologías del lenguaje, un dominio, como hemos visto, que se nutre del mundo universitario pero que no puede desarrollarse plenamente sin la existencia de un tejido empresarial; en este sentido, algunos grupos de investigación empezaron hace ya tiempo a crear sus propias compañías, conocidas como *spin-off*, a través de las que comercializan productos y servicios. Sin embargo, la formación actual de un graduado en lingüística, en lengua española o en filología hispánica no parece incorporar, por lo general, ni las destrezas ni los conocimientos imprescindibles para llevar a cabo un trabajo productivo en un equipo dedicado a las tecnologías del lenguaje. La consecuencia más inmediata de estas carencias es que las labores propias de los profesionales de la lingüística las realizan, en ocasiones, informáticos o ingenieros, plenamente competentes en sus especialidades, pero no siempre con los conocimientos adecuados en lo que a las disciplinas lingüísticas se refiere. Así, no sólo se pierden importantes oportunidades de integración laboral, sino que se relega al lingüista a un

papel de mero proveedor de datos –que no de conocimiento–, de etiquetador manual de corpus o de revisor de los resultados proporcionados por las herramientas informáticas.

El uso cada vez más creciente de ordenadores, la necesidad de emplearlos del modo más sencillo posible –es decir, recurriendo al lenguaje "natural"– y la ayuda que éstos nos proporcionan en muchas actividades cotidianas –desde dictar un texto o reservar un billete por teléfono hasta recuperar documentos en diversas lenguas y disponer de una traducción–, debería llevar a una concepción de la lingüística como una disciplina que, sin prescindir de su necesaria vertiente teórica, tuviera una presencia visible y útil en la sociedad de la información y del conocimiento. Esto implica, en primer lugar, reconocer el lugar de la lingüística aplicada en la universidad, dejando de lado la jerarquía implícita que en ocasiones se establece entre "los teóricos" y "los aplicados", y valorando adecuadamente la investigación que se realiza en los terrenos que más entroncan la lingüística con el mundo real. Con ello, sería posible plantear una formación que permitiera a los futuros lingüistas participar, en igualdad con otros profesionales, en equipos interdisciplinares dedicados al desarrollo de tecnologías del lenguaje.

Las tecnologías lingüísticas están alcanzando un nivel de madurez que permite ya su uso en diversas aplicaciones; pero se ha señalado a menudo que uno de los principales obstáculos que impiden realizar avances más significativos es la falta de conocimiento sobre el lenguaje. Hemos mencionado las dificultades para el reconocimiento del habla espontánea; las limitaciones de expresividad de los conversores de texto en habla; la falta de naturalidad en la interacción con los sistemas de diálogo; los problemas que todavía plantea el análisis o la traducción automática de cualquier tipo de texto, y los problemas aún mayores de la traducción de conversaciones; las posibilidades que se abren para la búsqueda y la recuperación de información en grandes bases de datos documentales multilingües, sean escritas u orales, y otros campos, como la multimodalidad, en los que se centrará la investigación en los próximos años. Por estos motivos, no cabe duda de que nos encontramos ante un futuro inmediato en el que los lingüistas pueden realizar contribuciones relevantes para lograr un uso más humano de las nuevas tecnologías. Sin embargo, si no estamos adecuadamente preparados, y si no sabemos reconocer que una de las dimensiones de la lingüística es, precisamente, la que se complementa con la tecnología, perderemos de nuevo una excelente oportunidad.

## 14.8. Preguntas de reflexión

1. A partir de las demostraciones en línea de los conversores de texto en habla para el español que se mencionan en la sección 2.1, cuyas direcciones en Internet encontrará en el apartado "Recursos, demostraciones y fuentes de información", obtenga con cada uno de ellos una versión sintetizada del mismo texto y compare los resultados. Puede tener en cuenta la realización de los elementos segmentales, de los elementos suprasegmentales –p. ej., entonación y ritmo–, la inteligibilidad y la calidad de la voz.
2. Transcriba ortográficamente un fragmento breve de habla espontánea –p. ej., la intervención de un oyente en un programa de radio– y tradúzcalo a otra lengua que conozca empleando alguno de los traductores automáticos disponibles en la red (encontrará las direcciones en el apartado "Recursos, demostraciones y fuentes de información"). Con los resultados obtenidos, elabore una lista de problemas que se plantean en la traducción automática del habla.
3. Seleccione un texto breve en una lengua extranjera y tradúzcalo al español empleando los sistemas de traducción automática disponibles en la red (direcciones en el apartado "Recursos, demostraciones y fuentes de información"). Clasifique, con criterios lingüísticos, los tipos de errores que ha encontrado y reflexione sobre sus posibles causas. Argumente los motivos por los que un determinado sistema le parecería preferible a otros.
4. Identifique en un texto en español las palabras o construcciones que el corrector ortográfico que utiliza habitualmente señala como errores. En el caso de que el corrector ofrezca alternativas, explique la relación entre el texto original y las alternativas mostradas; si no se ofrece ninguna, intente sugerir alguna hipótesis sobre las causas. Comente también aquellos casos en los que se producen errores no detectados por el corrector e intente explicar los motivos por lo que esto sucede. Si dispone también de un corrector gramatical, realice esta misma actividad.
5. Busque algunos ejemplos de ambigüedad morfológica y de ambigüedad sintáctica en español y analícelos con las herramientas en línea que se mencionan en la sección 4.1 (direcciones en el apartado "Recursos, demostraciones y fuentes de

información"). Comente los problemas que plantea en estos casos un análisis automático y las soluciones ofrecidas por los sistemas que haya probado.

6. Reflexione sobre la posible utilidad de las tecnologías lingüísticas y de los recursos lingüísticos expuestos en este capítulo para la enseñanza de lenguas asistida por ordenador en el contexto de la enseñanza del español como lengua extranjera. Puede consultar el capítulo 5 de este manual o trabajos sobre cuestiones específicas como Córdoba (2001), Jacobi (2001), Díaz y Ruggia (2004) o Morante (2004).
7. A partir de los contenidos de este capítulo y las lecturas complementarias que haya podido realizar, exponga qué conocimientos cree que debería tener un lingüista para integrarse en un equipo dedicado a las tecnologías del lenguaje. Puede documentarse también buscando información sobre programas de estudios en centros que imparten cursos sobre esta materia.
8. Justifique, con ejemplos concretos, la afirmación que se realiza en las consideraciones finales: "El especialista en lingüística puede y debe tener un papel activo en el campo de las tecnologías del lenguaje, puesto que éstas constituyen una de las ramas de la lingüística aplicada".

## Bibliografía

ABAITUA, J. (2002a): *Introducción a la traducción automática –en diez horas–*. Grupo DELI, Universidad de Deusto. http://sirio.deusto.es/abaitua/konzeptu/ta/mt10h_ es/index.html

ABAITUA, J. (2002b): "Tratamiento de corpora bilingües." En M. Martí y J. Llisterri (eds.). *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita* (pp. 61-90). Barcelona, Edicions Universitat de Barcelona - Fundación Duques de Soria. http://paginaspersonales.deusto.es/abaitua/konzeptu/ta/soria00.pdf

AGUILERA, S.; GODINO, J.; PALAZUELOS, S. y MARTÍN, J. (2001): Aplicaciones sociales de las tecnologías de la lengua. *Quark. Ciencia, Medicina, Comunicación y Cultura* 21: 90-94. http://www.imim.es/quark/21/021090.htm

ALONSO, J. (2001): "La traducció automàtica." En M. Martí (coord.). *Les tecnologies del llenguatge* (pp. 86-119). Barcelona, Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: "La traducción automática." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 94-129). Barcelona, Editorial UOC, 2003.]

BADIA, T. (2001): "Tècniques de processament del llenguatge." En M. Martí (coord.). *Les tecnologies del llenguatge* (pp. 189-238). Barcelona, Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: "Técnicas de procesamiento del lenguaje." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 193-248). Barcelona, Editorial UOC, 2003.]

BALARI, S. (1999): "Formalismos gramaticales de unificación y procesamiento basado en restricciones." En J. Gómez, A. Lorenzo, J. Pérez y A. Álvarez (eds.). *Panorama de la investigación en lingüística informática* RESLA, Revista Española de Lingüística Aplicada, Volumen monográfico. (pp. 117-152).

BERBER SARDINHA, T. (2004): *Lingüística de Corpus.* Barueri, São Paulo, Editora Manole.

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

BLECUA, J. (2001): Lengua española y tecnologías. *Archipiélago* 48: 100-197.

BLECUA, J.; CLAVERÍA, G.; SÁNCHEZ, C. y TORRUELLA, J. (eds.) (1999): *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.

CASACUBERTA, F. (2004): "Traducción automática del habla." En M. Martí y J. Llisterri (eds.). *Tecnologías del texto y del habla* (pp. 121-144). Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria.

CASTILLO, A. (1995): "El poder tecnológico de la lengua española." En Marqués de Tamarón (dir.). *El peso de la lengua española en el mundo* (pp. 173-194). Valladolid, Secretariado de Publicaciones de la Universidad de Valladolid - Fundación Duques de Soria - INCIPE, Fundación Instituto de Cuestiones Internacionales y Política Exterior.

CIVIT, M. (2003): *Criterios de etiquetación y desambiguación morfosintáctica de corpus en español.* Alicante, Sociedad Española para el Procesamiento del Lenguaje Natural. http://clic.fil.ub.es/personal/civit/PUBLICA/memoria.pdf.

CIVIT, M.; CASTELLÓN, I. y MARTÍ, M. (2002): "*Joven periodista triste busca casa frente al mar* o la ambigüedad en la anotación de corpus." En J. de D. Luque, A. Pamies y F. Manjón (eds.) *Nuevas tendencias en la investigación lingüística. Actas del I Congreso Internacional sobre Nuevas Tendencias de la Lingüística.* Granada, Universidad de Granada. http://clic.fil.ub.es/personal/civit/PUBLICA/GRANADA 01 -DEF.zip

CLIMENT, S. (2001): Sistemas de resumen automático de documentos. *Digithum, Revista digital d'humanitats* 3. http://www.uoc.edu/humfil/digithum/digithum3/catala/Art _Climent_esp/Climent/climent.html

COLE, R.; MARIANI, J.; USZKOREIT, H.; ZAENEN, A. y 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

COLEMAN, J. (2005): *Introducing Speech and Language Processing.* Cambridge, Cambridge University Press.

CÓRDOBA, F. (2001): El uso de los corpus lingüísticos en la enseñanza del español. *Boletín de la Asociación de Profesores de Español de la República Checa.* http://oldwww.upol.cz/res/ssup/ape/boletin2001/cordoba.htm

DAHL, D. (ed.) (2004): *Practical Spoken Dialog Systems.* Dordrecht, Kluwer.

DALE, R.; MOISL, H. y SOMERS, H. (eds.) (2000): *Handbook of Natural Language Processing.* Nueva York, Marcel Dekker.

DÍAZ, L. y RUGGIA, A. (2004): Cómo evaluar textos de fines específicos con ayuda de recursos informáticos: nuevas tecnologías al servicio del *feedback* en ELE. *RedELE, Revista Electrónica de Didáctica del Español como Lengua Extranjera.* http://www.sgci.mec.es/redele/revista/pdf/diaz_ruggia.pdf

DUTOIT, T. (1997): *An Introduction to Text-to-Speech Synthesis.* Dordrecht, Kluwer.

FARGHALY, A. (ed.) (2003): *Handbook for Language Engineers.* Stanford, CSLI Publication.

FELIU, J.; VIVALDI, J. y CABRÉ, T. (2002): *Ontologies: a Review.* Barcelona, Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. ftp://ftp.iula.upf.es/pub/publicacions/02inf034.pdf

FERRÁNDEZ, A. (2004): "Sistemas de pregunta y respuesta." En M. Martí y J. Llisterri (eds.). *Tecnologías del texto y del habla* (pp. 11-36). Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria.

GEOFFROIS, E. (2004): "Identification automatique des langues: techniques, ressources et évaluations." En *MIDL 2004. Modélisations pour l'identification des langues et des variétés dialectales* (pp. 43-44). (París, 29-30 de noviembre). http://www.limsi.fr/MIDL/actes/conference%20invitee%20I/Geoffrois_MIDL2004.pdf

GOLDEROS, F. (2001): Tecnologías del habla en español: convergencia con Internet. Ponencia en el II Congreso Internacional de la Lengua Española. El español en la Sociedad de la Información. (Valladolid, 16-9 de octubre). (http://cvc.cervantes. es/obref/congresos/valladolid/ponencias/el_espanol_en_la_sociedad/4_internet_en_espanol/golderos_f.htm).

GÓMEZ, X. (2000): "Lingüística computacional." En F. Ramallo, G. Rei-Doval y X. Rodríguez Yáñez (eds.). *Manual de Ciencias da Linguaxe* (pp. 221-268). Vigo, Edicións Xerais de Galicia. http://webs.uvigo.es/sli/arquivos/xerais.pdf.

GÓMEZ, X. (2001): "Recursos d'ajut a l'edició. Ortografia, sintaxi i estil." En M. Martí (coord.). *Les tecnologies del llenguatge* (pp. 15-26). Barcelona, Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: "Recursos de ayuda a la edición." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 30-40). Barcelona, Editorial UOC, 2003.]

GÓMEZ, X. y LORENZO, A. (eds.) (1996): *Lingüística e informática.* Santiago de Compostela, Tórculo Edicións.

GONZALO, J. y VERDEJO, F. (2001): "Recuperació i extracció d'informació." En M. Martí (coord.). *Les tecnologies del llenguatge* (pp. 151-187). Barcelona, Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: "Recuperación y extracción de información." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 157-192). Barcelona, Editorial UOC, 2003.]

GRANSTRÖM, B.; HOUSE, D. y KARLSSON, I. (eds.) (2002): *Multimodality in Language and Speech Systems.* Dordrecht, Kluwer.

HOLLIEN, H. (2002): *Forensic Voice Identification.* San Diego, Academic Press.

HOLMES, J. y HOLMES, W. (2001): *Speech Synthesis and Recognition* (2ª ed.). Londres, Taylor & Francis.

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

HUTCHINS, W. (2003): "Machine Translation: General Overview." En R. Mitkov (ed.). *The Oxford Handbook of Computational Linguistics* (pp. 501-511). Oxford, Oxford University Press. http://ourworld.compuserve.com/homepages/WJHutchins/ Mitk ov.pdf

HUTCHINS, W. (2005): *Compendium of translation software: directory of commercial machine translation systems and computer-based translation support tools.* Ginebra, EAMT, European Association for Machine Translation. http://ourworld.compuserve.com/homepages/WJHutchins/Compendium.htm

JACOBI, C. (2001): *Lingüística de Corpus e ensino de espanhol a brasileiros: Descrição de padrões e preparação de atividades didáticas (decir/hablar; mismo; mientras /en cuanto/ aunque).* Tesis de disertación de maestría, Pontifícia Universidades Católica de São Paulo. http://lael.pucsp.br/lael-inf/teses/tese_ claudia.zip

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

KENNEDY, G. (1998): *An Introduction to Corpus Linguistics.* Londres, Longman.

KUPPEVELT, J.; DYBKJAER, L. y BERNSEN, N. (eds.) (2005): *Advances in Natural Multimodal Dialogue Systems.* Dordrecht, Springer.

LAVID, J. (2005): *Lenguaje y nuevas tecnologías. Nuevas perspectivas, métodos y herramientas para el lingüista del siglo XXI.* Madrid, Cátedra.

LLISTERRI, J. (1999): 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/Llisterri_99_TecnolLing_SocInfo.pdf

LLISTERRI, J. (2001): "Les tecnologies de la parla." En M. Martí (coord.). *Les tecnologies del llenguatge* (pp. 239-272). Barcelona, Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: "Las tecnologías del habla." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 249-281). Barcelona, Editorial UOC, 2003.]

LLISTERRI, J. (2003): Lingüística y tecnologías del lenguaje. *Lynx. Panorámica de Estudios Lingüísticos* 2: 9-71. http://liceu.uab.es/~joaquim/publicacions/TecnoLing_Lynx 02.pdf

LLISTERRI, J. (2004a): "Las tecnologías del habla para el español." En R. Sequera (ed.). *Ciencia, tecnología y lengua española: La terminología científica en español* (pp. 123-141). Madrid, Fundación Española para la Ciencia y la Tecnología. http://liceu.uab.es/~joaquim/publicacions/TecnolHablaEsp_FECyT03.pdf

LLISTERRI, J. (2004b): "Las tecnologías lingüísticas en España." En *El español en el mundo. Anuario del Instituto Cervantes 2004* (pp. 229-251). Madrid, Instituto Cervantes - Círculo de Lectores - Plaza & Janés. http://cvc.cervantes.es/obref/anuario/anuario_04/llisterri/default.htm

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

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

LLISTERRI, J.; CARBÓ, C.; MACHUCA, M.; DE LA MOTA, C.; RIERA, M. y RÍOS, A. (2003a): "El papel de la lingüística en el desarrollo de las tecnologías del habla." En M. Casas (dir.) y C. Varo (ed.). *VII Jornadas de Lingüística* (pp. 137-191). Cádiz, Servicio de Publicaciones de la Universidad de Cádiz. http://liceu.uab.es/publicacions/ Linguistica_TH_Cadiz02.pdf

LLISTERRI, J.; MACHUCA, M.; DE LA MOTA, C.; RIERA, M. y RÍOS, A. (2003b): "Entonación y tecnologías del habla." En P. Prieto (ed.). *Teorías de la entonación* (pp. 209-243). Barcelona, Ariel.

LLISTERRI, J.; CARBÓ, C.; MACHUCA, M.; DE LA MOTA, C.; RIERA, M. y RÍOS, A. (2004): "La conversión de texto en habla: aspectos lingüísticos." En M. Martí y J. Llisterri (eds.). *Tecnologías del texto y del habla* (pp. 145-186). Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria. http://liceu.uab.es/publica cions/Linguistica_CTH_FDS02.pdf

LLISTERRI, J.; MACHUCA, M.; DE LA MOTA, C.; RIERA, M. y RÍOS, A. (2005): Corpus orales para el desarrollo de las tecnologías del habla en español. *Oralia. Análisis del discurso oral* 8 [en prensa]. http://liceu.uab.es/~joaquim/publicacions/Oralia_04.pdf

LÓPEZ, F.; GONZALO, J. y VERDEJO, F. (2004): Búsqueda de información multilingüe: Estado del arte. *Revista Iberoamericana de Inteligencia Artificial* 8: 11-35. http://nlp. uned.es/pergamus/pubs/EstadoDelArteCLIR.pdf

LÓPEZ-CÓZAR, R. y ARAKI, M. (2005): *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment.* Chichester, John Wiley.

MANNING, C. y SCHÜTZE, H. (1999): *Foundations of Statistical Natural Language Processing.* Cambridge, MA., The MIT Press. http://nlp.stanford.edu/fsnlp/

MARIÑO, J. y NADEU, C. (2004): "La representación de la voz para el reconocimiento del habla." En M. Martí y J. Llisterri (eds.). *Tecnologías del texto y del habla* (pp. 187-224). Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria.

MARTÍ, M. (1994): "Lexicografía computacional." En J. Gómez Guinovart (ed.). *Aplicaciones lingüísticas a la informática* (pp. 35-50). Santiago de Compostela, Tórculo Edicións.

MARTÍ, M. (coord.). (2001): *Les tecnologies del llenguatge.* Barcelona: Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: *Tecnologías del lenguaje.* Barcelona, Editorial UOC, 2003.]

MARTÍ, M. (2003): "Las tecnologías de la lengua y la sociedad de la información." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 9-29). Barcelona, Editorial UOC.

MARTÍ, M. y CASTELLÓN, I. (2000): *Lingüística computacional.* Barcelona, Edicions de la Universitat de Barcelona.

MCENERY, T. (2003): "Corpora." En R. Mitkov (ed.). *The Oxford Handbook of Computational Linguistics* (pp. 448-463). Oxford, Oxford University Press.

MCENERY, T. y WILSON, A. (2001): *Corpus Linguistics* (2ª ed.). Edimburgo, Edinburgh University Press. http://www.ling.lancs.ac.uk/monkey/ihe/linguistics/contents.htm

MINKER, W. y BENNACEF, S. (2004): *Speech and Human-Machine Dialog.* Dordrecht, Kluwer.

MINKER, W.; BÜHLER, D. y DYBKJAER, L. (eds.) (2005): *Spoken Multimodal Human-Computer Dialogue in Mobile Environments.* Dordrecht, Springer.

MITKOV, R. (ed.) (2003): *The Oxford Handbook of Computational Linguistics.* Oxford, Oxford University Press.

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.cervantes.es/seg_nivel/lect_ens/oesi/Articulos/Anastasio%20Molano/spanish%20version/Molano.htm

MORA, E. y RODRÍGUEZ, M. (2001): "Research Activities in and Applications of Speech Technologies in Latin America." En *COCOSDA Workshop 2001* (Aalborg, 2 de septiembre de 2001). http://www.slt.atr.co.jp/cocosda/cocosdaE-31-8-01.doc

MORANTE, R. (2004): VOCABLE: Una plataforma para el aprendizaje de vocabulario asistido por ordenador. *RedELE, Revista Electrónica de Didáctica del Español como Lengua Extranjera* 2. http://www.sgci.mec.es/redele/revista2/pdfs2/morante.pdf

MORENO FERNÁNDEZ, F. (1997): "La formación de corpus de lengua hablada." En F. Moreno Fernández (ed.). *Trabajos de sociolingüística hispánica* (pp. 93-114). Alcalá de Henares, Universidad de Alcalá, Servicio de Publicaciones.

MORENO, J. (2004): "La investigación en ingeniería lingüística en España." En R. Sequera (ed.). *Ciencia, tecnología y lengua española: La terminología científica en español* (pp. 97-122). Madrid, Fundación Española para la Ciencia y la Tecnología. http://www.fecyt.es/documentos/Libro%20CTL_web.pdf

MORENO, L.; PALOMAR, M.; MOLINA, A. y FERRÁNDEZ, A. (1999): *Introducción al procesamiento del lenguaje natural.* Alicante, Servicio de Publicaciones de la Universidad de Alicante.

NOLAN, F. (1997): "Speaker Recognition and Forensic Phonetics." En W. Hardcastle y J. Laver (eds.). *The Handbook of Phonetic Sciences* (pp. 744-767). Oxford, Blackwell.

O'SHAUGHNESSY, D. (2000): *Speech Communications: Human and Machines* (2ª ed.). Reading, MA, Addison Wesley.

PASCUAL, J. (1995): "Escándalo o precaución. Sobre el futuro de nuestra lengua." En Marqués de Tamarón (dir.). *El peso de la lengua española en el mundo* (pp. 135-171). Valladolid, Secretariado de Publicaciones de la Universidad de Valladolid - Fundación Duques de Soria - INCIPE, Fundación Instituto de Cuestiones Internacionales y Política Exterior. [Versión adaptada en *Index-net, Programa de apoyo al profesorado.* Editorial Santillana]. http://www.index-net.santillana.es/rcs/_archivos/Documentos/lenguadoc/futuroleng.pdf

RAFEL, J. y SOLER, J. (2001): "El processament de corpus. La lingüística empírica." En M. Martí (coord.). *Les tecnologies del llenguatge* (pp. 27-59). Barcelona, Edicions de la Universitat Oberta de Catalunya. [Trad. esp.: "El procesamiento de corpus." En M. Martí (coord.). *Tecnologías del lenguaje* (pp. 41-73). Barcelona, Editorial UOC, 2003.]

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/19/019026.htm

RODRÍGUEZ, H. (2002): "Técnicas de análisis sintáctico." En M. Martí y J. Llisterri (eds.). *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita* (pp. 91-132). Barcelona, Edicions Universitat de Barcelona - Fundación Duques de Soria.

RODRÍGUEZ, H. (2004): "Lingüística y estadística, ¿incompatibles?" En M. Martí y J. Llisterri (eds.). *Tecnologías del texto y del habla* (pp. 89-117). Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria.

RODRÍGUEZ, L; DOCÍO, L. y GARCÍA, C. (1998): "Panorámica de la tecnología en reconocimiento automático de locutores." En J. Gómez, A. Lorenzo, J. Pérez y A. Álvarez (eds.). *Panorama de la investigación en lingüística informática* (pp. 36-40). RESLA, Revista Española de Lingüística Aplicada (volumen monográfico).

ROSE, P. (2002): *Forensic Speaker Identification.* Londres, Taylor & Francis.

RUBIO, A. y HERNÁEZ, I. (2005): *Libro blanco de Tecnologías del Habla.* Granada, Red Temática en Tecnologías del Habla. http://www.rthabla.org/LibroBlanco-Tecno logiasDelHabla.pdf

RUIZ, J. (2005): "Lenguaje e informática / Lenguaje y ordenadores." En Á. López y B. Gallardo (eds.). *Comunicación y lenguaje* (pp. 401-436). Valencia, Universitat de València.

SÁNCHEZ, F. (2004): "Comentario del panel 'Tecnologías lingüísticas para el español'." En R. Sequera (ed.). *Ciencia, tecnología y lengua española: la terminología científica en español* (pp. 160-163). Madrid, Fundación Española para la Ciencia y la Tecnología. http://www.fecyt.es/documentos/Libro%20CTL_web.pdf

SÁNCHEZ, F.; PORTA, J.; SANCHO, J.; NIETO, A.; BALLESTER, A.; FERNÁNDEZ, A.; GÓMEZ, J.; GÓMEZ, L.; RAIGAL, E. y RUIZ, R. (1999): La anotación de los corpus CREA y CORDE. *Procesamiento del Lenguaje Natural* 25: 175-182. http://www.sepln.org/revista SEPLN/revista/25/25-Pag175.pdf

SCHIEL, F.; DRAXLER, C.; BAUMANN, A.; ELLBOGEN, T. y STEFFEN, A. (2004): *The Production of Speech Corpora.* Version 2.5. Munich, Bavarian Archive for Speech Signals. http://www.phonetik.uni-muenchen.de/Forschung/BITS/TP1/Cookbook/

SERENARI, M.; DYBKJAER, L.; HEID, U.; KIPP, M. y REITHINGER, N. (2002): *Survey of Existing Gesture, Facial Expression and Cross-Modality Coding Schemes.* NITE, Natural Interactivity Tools Engineering, Deliverable D2.1. September 2002. http://nite.nis. sdu.dk/deliverables/NITE-D2.1-sept02-F.pdf

SPERBERG-MCQUEEN, C. y 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/

SUBIRATS, C. y 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. (2002): "Interfaces de voz con lenguaje natural." En M. Martí y J. Llisterri (eds.). *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita* (pp. 189-207). Barcelona, Edicions Universitat de Barcelona - Fundación Duques de Soria.

TAPIAS, D. y HERNÁNDEZ, L. (2004): "Los sistemas de diálogo en los servicios telefónicos: evolución y consideraciones de diseño." En M. Martí y J. Llisterri (eds.). *Tecnologías del texto y del habla* (pp. 225-253). Barcelona, Edicions de la Universitat de Barcelona - Fundación Duques de Soria.

TORRUELLA, J. y LLISTERRI, J. (1999): "Diseño de corpus textuales y orales." En J. Blecua, G. Clavería, C. Sánchez y J. Torruella (eds.). *Filología e informática. Nuevas tecnologías en los estudios filológicos* (pp. 45-77). Barcelona, Seminario de Filología e Informática, Departamento de Filología Española, Universidad Autónoma de Barcelona - Editorial Milenio. http://liceu.uab.es/~joaquim/publica cions/Torruella_Llisterri_99.pdf

VÁZQUEZ, G.; FERNÁNDEZ, A. y MARTÍ, M. (2002): "Léxicos verbales computacionales." En M. Martí y J. Llisterri (eds.) *Tratamiento del lenguaje natural. Tecnología de la lengua oral y escrita* (pp. 29-60). Barcelona, Edicions Universitat de Barcelona - Fundación Duques de Soria.

VICEDO, J.; RODRÍGUEZ, H.; PEÑAS, A. y MASSOT, M. (2003): Los sistemas de búsqueda de respuestas desde una perspectiva actual. *Procesamiento del Lenguaje Natural* 31: 351-368. http://www.sepln.org/revistaSEPLN/revista/31/31-Pag351.pdf

VILLARRUBIA, L.; GARRIDO, J.; RELAÑO, J.; CAMINERO, J.; ESCALADA, J.; RODRÍGUEZ, M. y HERNÁNDEZ, L. (2002): Productos de tecnología del habla para Latinoamérica. *Comunicaciones de Telefónica I+D* 27: 53-72. http://www.tid.es/documentos/revista_comunicaciones_i%2Bd/numero27.pdf

VILLARRUBIA, L.; RODRÍGUEZ, A.; RELAÑO, J.; GARIJO, F.; BERNAT, J.; HERNÁNDEZ, L.; SAN SEGUNDO, R.; TAPIAS, D. y MARÍA, L. (2003): Tecnología del habla para aplicaciones multilingües, multiservicio y multiplataforma. *Comunicaciones*

*de Telefónica I+D* 30: 47-78. http://www.tid.es/documentos/revista_comunicaciones_i%2Bd/ numero30.pdf

VOSSEN, P. (2001): Oportunidades para la ingeniería lingüística. *Digithum, Revista Digital d'Humanitats* 3. http://www.uoc.es/humfil/articles/esp/vossen/vossen.html

WAHLSTER, W. (2000): "Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System." En W. Wahlster (ed.). *Verbmobil: Foundations of Speech-to-Speech Translation* (pp. 3-21). Heidelberg - Nueva York, Springer. http://verbmobil.dfki.de/ww.html

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

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

## *Recursos, demostraciones y fuentes de información citados en el texto*

Las direcciones de los recursos, las demostraciones y las fuentes de información disponibles en Internet y citados a lo largo del capítulo pueden consultarse en: http://liceu.uab.es/~joaquim/publicacions/linguistica_aplicada_del_espanol.html