Fundamentos de Lingüística Computacional: bases teóricas, líneas de investigación y aplicaciones

Javier Gómez Guinovart
Universidad de Vigo (Vigo)
jgomez@uvigo.es

RESUM

S'ofereix un estat de la qüestió de la recerca i desenvolupament en el camp de la lingüística computacional, presentant els fonaments, aplicacions i perspectives del seus principals vessants d'estudi.

RESUMEN

Se ofrece un estado de la cuestión de la investigación y desarrollo en el campo de la lingüística computacional, presentando los fundamentos, aplicaciones y perspectivas de sus principales vertientes de estudio.

1. El campo de la lingüística computacional

La lingüística computacional (o lingüística informática) es un campo científico interdisciplinar relativamente reciente –cerca de cincuenta años de investigación y desarrollo– cuyo objetivo radica en incorporar en los ordenadores la habilidad en el manejo del lenguaje humano.

Desde el punto de vista de su vinculación a la informática, y también por motivos históricos, la lingüística computacional suele ser considerada como una subdisciplina de la *inteligencia artificial*. La inteligencia artificial, por su parte, es una subdisciplina de la informática que se ocupa de la comprensión de la inteligencia y del diseño de máquinas inteligentes, es decir, de máquinas que presentan características asociadas con el entendimiento humano, como el raciocinio, la comprensión del lenguaje hablado y escrito, el aprendizaje o la toma de decisiones. En una afortunada definición de Minsky (1967), la inteligencia artificial es «la ciencia de hacer que las máquinas hagan cosas que, de haber sido hechas por seres humanos, requerirían inteligencia».

Por otra parte, desde el punto de vista de su vinculación a la lingüística, la lingüística computacional puede ser considerada una subdisciplina de la *lingüística teórica*, en tanto que uno de sus objetivos es la elaboración de modelos formales (e implementables informáticamente) del lenguaje humano.

Por último, en cuanto que disciplina experimental orientada a la elaboración de productos comerciales y de investigación, la lingüística computacional forma parte de las denominadas *industrias de la lengua*, un sector industrial cada vez más amplio que proporciona datos y programas informáticos aplicados al tratamiento del lenguaje. Datos tales como diccionarios electrónicos e impresos, bancos de datos terminológicos y tesauros; y programas tales como sistemas de traducción automática, interfaces de consulta a bases de datos en lenguaje natural, o correctores ortográficos y estilísticos.

Esta misma actividad, cuyos resultados se plasman en las aplicaciones lingüísticas de la informática, vincula la lingüística computacional con la *lingüística aplicada*, una rama de la lingüística dedicada a aplicar los resultados y métodos de la investigación lingüística a campos tales como la enseñanza de idiomas, la traducción e interpretación, o la logopedia.

Siendo la lingüística computacional una disciplina tan reciente, y abarcando objetivos tan variados, resulta bastante comprensible la gran vacilación terminológica que impera en su dominio. Una aproximación tentativa a la delimitación del campo de estudio exige el reconocimiento de un mínimo de tres vertientes. Estas tres grandes líneas de trabajo –ordenadas de la más vinculada a la lingüística, a la más vinculada a la informática– son:

a. La informática aplicada a la investigación lingüística
b. La implementación de teorías lingüísticas
c. Las aplicaciones lingüísticas de la informática.

En los siguientes apartados, expondré con algo más de detalle en qué consisten estas tres líneas de trabajo, y las ilustraré con ejemplos y orientaciones bibliográficas.

2. La informática aplicada a la investigación lingüística

La aplicación de los ordenadores a la investigación lingüística, es decir, al estudio científico del lenguaje, suele recibir el nombre de *lingüística informática* o de *informática aplicada a la lingüística* (adaptaciones del término inglés *linguistic computing*). Veamos, a título de ejemplo, una muestra de una investigación lingüística real de un aspecto morfológico concreto de la lengua inglesa, en un período determinado de su evolución histórica (Lezcano *et al.*, 1997).

Los investigadores de la filología inglesa interesados en conocer la vitalidad del sufijo *-able* en inglés moderno temprano, pueden acceder a una colección de textos escritos entre los años 1500 y 1710 de más de medio millón de palabras, con el objetivo de observar de manera intuitiva si hay muchas o pocas palabras formadas con tal sufijo, y establecer sus conclusiones (por ejemplo, «en inglés moderno temprano, a mi manera de ver, no se empleaba mucho este sufijo»). Si además de interés, la persona investigadora posee grandes dotes de observación, tiempo y paciencia, puede llegar a algunas conclusiones (de fiabilidad ciertamente cuestionable) sobre el porcentaje de palabras que contienen el sufijo *-able* en la colección de textos mencionada.

Sin embargo, si las 550.000 palabras de la colección de textos se hallan en soporte magnético (como lo están en la actualidad, formando parte del denominado Corpus de Helsinki), resulta posible obtener rápidamente, mediante un sencillo programa informático, la lista de todas las palabras de los textos compilados acabadas en *-able*, junto con diversas indicaciones sobre su frecuencia de aparición o sobre su emplazamiento concreto en los textos integrantes de la colección.

Si además estos textos estuvieran adecuadamente «etiquetados» (es decir, si las palabras de los textos llevaran adjunta una indicación sobre su categoría morfosintáctica), se podría conocer automáticamente el número total de adjetivos incluidos en la colección de textos y, lo que es más importante, se podría deducir la proporción de adjetivos acabados en *-able*, en relación al número total de adjetivos, e incluso se podría comparar la productividad de este sufijo con la de otros sufijos adjetivos. Así, basándose en datos empíricos (y no en suposiciones o cómputos subjetivos) y aplicando la potencia de cálculo y la capacidad de memoria de los ordenadores al estudio científico del lenguaje, los investigadores podrían llegar a establecer conclusiones bien fundamentadas sobre la productividad morfológica de un determinado sufijo en una determinada etapa del inglés.

Este tipo de estudios, caracterizados por utilizar como base de la investigación compilaciones de textos reales en soporte informático (denominadas técnicamente *corpus*), se enmarcan en el ámbito de trabajo conocido como *lingüística de corpus* (Badia, 1997; McEnery y Wilson, 1996), una disciplina de gran desarrollo en los últimos diez años cuyos resultados se han aplicado a ámbitos tan diversos como la lexicografía (Fillmore y Atkins, 1994; Rafel, 1996, 1997; Rojo, 1992; Santamarina, 1996; Sinclair, 1987), la construcción de gramáticas (Hallebeek, 1992; Halliday, 1991; Souter y O'Donoghue, 1991) y la traducción automática (Abaitua *et al.*, 1997; Brown *et al.*, 1990 y 1993; Gale y Church, 1993; Kay y Röscheisen, 1993).

Dentro del campo de la lingüística de corpus, posee una gran vitalidad la investigación sobre el *etiquetado*. El etiquetado morfológico de un corpus implica asignar a cada palabra del corpus un código o etiqueta (*tag*, en inglés) con información relevante a su categoría morfológica. Un programa informático de etiquetado morfológico automático recibe como entrada una secuencia de palabras (por ejemplo, «el niño come peras») y produce como salida una secuencia de etiquetas morfológicas asociada a la secuencia de palabras (por ejemplo, «el_AMS niño_NMS come_VP3 peras_NFP», donde AMS significa «artículo masculino singular»; NMS, «nombre masculino singular»; VP3, «tercera persona singular de verbo en presente de indicativo»; y AMS, «adjetivo masculino singular»). La lista de etiquetas utilizadas por el programa dependerá de las características lingüísticas del corpus, de los objetivos del etiquetado, de los límites de la implementación informática y de los presupuestos teóricos empleados en su desarrollo.

Los corpus etiquetados morfológicamente pueden ser de gran utilidad para la investigación lingüística y para el desarrollo de aplicaciones de procesamiento del lenguaje natural, ya que permiten trabajar directamente con palabras no ambiguas respecto a su categoría gramatical y con patrones morfosintácticos superficiales (no analizados sintácticamente). Así, en un corpus del castellano etiquetado morfológicamente, no sería demasiado difícil detectar automáticamente los casos de pasiva, los pronombres preverbales o los casos de *lo* seguido de adjetivo, por poner sólo tres ejemplos ilustrativos de su uso como fuente en la investigación lingüística.

Ciertamente, un corpus etiquetado con información sintáctica (es decir, un corpus analizado sintácticamente y codificado con la información relevante al respecto) puede resultar aun más valioso que un corpus etiquetado únicamente con información morfológica (García-Miguel, 1994; García-Miguel y Vázquez, 1994; Garside, 1993;

Marcus *et al.*, 1993; Rojo, 1993). Sin embargo, debido a la *enorme complejidad de la tarea y a la dificultad de su* automatización, los corpus amplios etiquetados sintácticamente son muy escasos (Arrarte y Llisterri, 1994; Fernández y Llisterri, 1996; Marcos Marín, 1994: 79-178).

Por el contrario, el número de corpus etiquetados morfológicamente es cada vez mayor, principalmente –aunque no exclusivamente– en lengua inglesa. Su incremento es debido tanto al interés que suscitan, como a la automatización (parcial o completa) del proceso de etiquetado ofrecida por los diversos etiquetadores morfológicos automáticos desarrollados hasta la fecha (Aduriz *et al.*, 1994; Farwell *et al.*, 1995; Pérez *et al.*, 1994; Sánchez y Nieto, 1995).

3. La implementación de teorías lingüísticas

La segunda línea de trabajo, orientada a la implementación de teorías lingüísticas, acostumbra a denominarse *lingüística computacional* (calco del inglés *computational linguistics*), en sentido estricto, y posee un triple objetivo:

a. La elaboración de teorías lingüísticas (o mejor, de modelos lingüísticos) en términos formales e implementables. Dentro de esta línea de investigación, se han desarrollado modelos lingüísticos computacionales como la *gramática léxica funcional* o LFG (Bresnan, 1982; Sells, 1985), la *gramática sintagmática generalizada* o GPSG (Gazdar *et al.*, 1985; Sells, 1985; Borsley, 1996), y *la gramática sintagmática dirigida por el núcleo* o HPSG (Pollard y Sag, 1994; Borsley, 1996), modelos agrupados genéricamente en la categoría de las *gramáticas de unificación* (Shieber, 1986).

b. La descripción de fenómenos lingüísticos concretos en el marco de alguno de estos modelos, y en cualquiera de los niveles de descripción lingüística: semántica (Badia y Colominas, 1995; Climent, 1995), morfología (Agirre *et al.* 1989; Carulla y Oosterhoff, 1996), sintaxis (Balari, 1992; Castellón *et al.*, 1997; Palomar *et al.*, 1995; Taulé y Castellón, 1994), etc.

c. La comprobación automatizada de la consistencia de una teoría lingüística o de sus predicciones (Climent y Farreres, 1996; Covington, 1990; Ruiz *et al.*, 1991; Ruiz y Gómez Guinovart, 1990). Por ejemplo, a partir de una gramática formada por reglas que describan la estructura interna de los constituyentes oracionales (reglas sintácticas del tipo: *SN (Det N*, parafraseable como «un sintagma nominal está formado por un determinante seguido de un nombre») y por reglas que introduzcan los elementos léxicos (reglas de «inserción léxica» como: *N (perro y Det (el)*, resulta relativamente sencillo crear un programa informático capaz de decidir si una determinada frase (por ejemplo, *el perro*) es descrita o no por la gramática.

Sin embargo, cuando el número de reglas y su complejidad aumentan de manera considerable (como inevitablemente sucede al intentar describir con cierta amplitud cualquier idioma), su implementación informática resulta de gran ayuda para comprobar la buena formación de los enunciados propuestos, o para comprobar los efectos de la incorporación de una nueva regla o de la modificación de una regla ya existente (Ruiz, 1996).

Los *formalismos lingüísticos* (o *sistemas de programación lingüística*) son lenguajes artificiales diseñados específicamente para representar conocimientos lingüísticos. Algunos formalismos lingüísticos –como DCG (Pereira y Warren, 1980), FUG (Kay, 1982), PATR (Shieber, 1986), DATR (Evans y Gazdar, 1996), la morfología de dos niveles (Koskenniemi, 1983), GFU (Ruiz, 1993) o ALE (Carpenter y Penn, 1997)-también son entendidos (o, mejor dicho, interpretados) directamente por los ordenadores, por lo que son especialmente adecuados para la implementación informática y la comprobación automática de las teorías lingüísticas. Para llevar a cabo estas tareas, como complemento o sustituto de los formalismos lingüísticos, se emplean lenguajes de programación declarativos y, en particular, el lenguaje de programación Prolog (Covington, 1994; Dik, 1992; Gazdar y Mellish, 1989; Pereira y Shieber, 1987).

4. Las aplicaciones lingüísticas de la informática

La tercera línea de trabajo de la lingüística computacional (entendida ahora nuevamente en sentido amplio) consiste en el diseño y elaboración de sistemas informáticos encaminados a la comprensión y generación de lenguas naturales. Este campo recibe las denominaciones de *procesamiento del lenguaje natural, tecnologías de la lengua* o *ingeniería lingüística*, dependiendo del aspecto de esta actividad donde se desee poner el énfasis (Abaitua, 1996: 244).

Algunas de las aplicaciones lingüísticas de la informática más populares en el mundo de los ordenadores personales son las tecnologías del habla (en particular, los sistemas de dictado) y la traducción automática. Junto a estas dos, expondré a continuación los fundamentos de otra aplicación lingüística de la informática de particular importancia para los lectores de este Anuario: los sistemas de extracción de información. Otras aplicaciones lin-

güísticas de la informática relevantes, no tratadas en este trabajo por motivos de espacio, son: la verificación lingüística automática (por ejemplo, los correctores ortográficos, sintácticos y estilísticos incorporados en los procesadores de textos) (Gojenola y Sarasola, 1994; Gómez Guinovart, 1996b; Mitton, 1996; Ramírez y Sánchez, 1996; Robertson y Willett, 1993; Rodríguez Magro, 1993); los diccionarios electrónicos de consulta (Jucker, 1994; Lorenzo y Gómez Guinovart, 1996; Rafel, 1996, 1997); y los sistemas de diálogo persona-máquina en lenguaje natural (por ejemplo, para formular consultas en castellano a una base de datos, para hacer reservas de vuelo por teléfono hablando en castellano con un ordenador, o para que el ordenador le plantee problemas de matemáticas a un estudiante y lo asesore en su solución) (Allen, 1995: 541-576; Appelt, 1985; Ferrari, 1991; Hovy, 1988).

4.1. Las tecnologías del habla

El objetivo de las tecnologías del habla (Keller, 1994; Llisterri, 1991; Moure y Llisterri, 1996: 153-171) es permitir la comunicación oral entre las personas y los ordenadores. Un enunciado oral de habla humana es una señal sonora continua que varía a lo largo del tiempo, es decir, es una señal analógica. Por contra, los ordenadores trabajan con señales digitales, es decir, con cadenas de símbolos discretos (o sea, cadenas de símbolos distintos no conectados entre sí de manera continua). En función de la dirección del mensaje en la situación comunicativa, el procesamiento del habla se enfrenta con dos tareas bien diferenciadas:

a. El reconocimiento del habla

b. La síntesis del habla

4.1.1. El reconocimiento del habla

El *reconocimiento del habla* (Allen, 1995: 611-628) consiste en convertir un enunciado oral (una señal sonora continua) en su representación simbólica discreta (por ejemplo, en el caso de los sistemas de dictado, en una cadena de letras agrupadas en palabras ortográficamente correctas). La popularidad del reconocimiento del habla se debe en gran medida a los sistemas de dictado para procesamiento de texto en ordenadores personales. Estos programas de dictado, comercializados por empresas como IBM y Dragon Systems (véase, más abajo, el apartado 5.1 de este trabajo), ofrecen versiones para habla fragmentada, en las que el usuario debe hacer una pausa entre las palabras, y versiones para habla continua, que permiten dictarle texto al ordenador sin necesidad de hacer pausas entre las palabras.

Una de las características más deseables en un sistema de reconocimiento del habla es su resistencia al ruido ambiente, especialmente para poder trabajar en entornos ruidosos (por ejemplo, en una fábrica, para controlar vocalmente el brazo de un robot), o cuando se desea poder operar a través del teléfono (por ejemplo, para dictarle a una centralita automatizada el número del abonado con el que se pretende comunicar). Por el momento, a pesar del interés evidente que esta cuestión suscita entre los proveedores de servicios de telecomunicaciones, y aunque se han producido avances innegables (Hernando *et al.*, 1997), no se ha encontrado aún la solución definitiva a la baja fiabilidad del reconocimiento del habla en entornos ruidosos.

Otro de los grandes retos del reconocimiento irrestricto del habla continua es la independencia del locutor. En primer lugar, un reconocedor es irrestricto si es capaz de reconocer el vocabulario general de una lengua. Esta característica es imprescindible, por ejemplo, en un sistema de dictado; sin embargo, otras aplicaciones del reconocimiento, como las centralitas automatizadas, únicamente necesitan reconocer unas pocas palabras. En segundo lugar, un reconocedor es independiente del locutor si está concebido para reconocer el habla de cualquier persona, mientras que, por el contrario, se dice que un reconocedor es dependiente del locutor si está concebido para reconocer el habla de una única persona.

Por ejemplo, una centralita automatizada de un sistema público de consulta telefónica ha de ser independiente del usuario obligatoriamente, ya que debe ser capaz de reconocer el habla de cualquier persona que pueda llamar. En cambio, un sistema de dictado para procesamiento de texto en ordenadores personales puede permitirse ser dependiente del locutor, ya que, en principio, va a ser utilizado por una única persona (la dueña del programa) en un entorno único (su propio ordenador). Como, además, en el estado actual de la cuestión, el reconocimiento irrestricto del habla continua con independencia del locutor no ha alcanzado todavía un grado de fiabilidad aceptable para su comercialización, los sistemas de dictado para habla continua exigen una adaptación a su usuario, adaptación que se logra sometiendo al sistema a una fase de entrenamiento. Durante esta fase, que en la práctica puede suponer media hora de lectura de un texto preparado, el usuario le proporcionará al sistema los datos necesarios sobre las características de su voz y sobre las características de su pronunciación particular de los sonidos de la lengua. Una vez realizado este entrenamiento, el porcentaje de acierto de la conversión de voz a texto de un sistema personal de dictado en condiciones óptimas puede resultar bastante elevado: de un 95% en los sistemas para habla fragmentada, y algo inferior en los sistemas para habla continua.

4.1.2. La síntesis del habla

La *síntesis del habla* consiste en convertir un conjunto de símbolos discretos (por ejemplo, en el caso de los sistemas de síntesis para usuarios invidentes de ordenadores personales, una cadena de letras agrupadas en palabras y posiblemente acompañadas por signos de puntuación) en una señal sonora continua de habla.

Aunque el problema de la inteligibilidad de la voz sintetizada se resolvió hace ya mucho tiempo, queda por solucionar la cuestión de su naturalidad, es decir, conseguir que la voz generada por el ordenador no suene «robótica» (Aguilar *et al.*, 1994). Una de las claves para que la voz sintetizada parezca más natural es la curva de entonación adoptada en la generación de los enunciados, y es en este terreno donde más se está investigando en la actualidad (Bullón y Pérez, 1994; Garrido, 1991; Hernáez *et al.*, 1995; López Gonzalo *et al.* 1994; Martí y Gudayol, 1994; entre otros).

4.2. La traducción automática

La traducción automática por ordenador (Abaitua, 1997; Aguilar-Amat, 1996; Church y Hovy, 1993; Hutchins y Somers, 1992; Jones, 1996; Whitelock y Kilby, 1995) constituye una de las líneas de investigación de la lingüística computacional de mayor complejidad intrínseca y, al mismo tiempo, uno de los desarrollos de mayor interés para el público no especialista. Sin embargo, muchas de las personas interesadas en este campo contemplan con cierto escepticismo las posibilidades de la traducción automática y de la traducción asistida por ordenador. Entre las causas más probables de las suspicacias actuales hacia esta tecnología hay que señalar los pobres resultados ofrecidos hasta ahora por los programas informáticos comerciales autodenominados de «traducción automática» y la generalización injustificada de esta percepción negativa al conjunto de aplicaciones informáticas específicamente diseñadas para su incorporación en el proceso de la traducción humana.

Con el fin de no caer en las confusiones derivadas de la ambigüedad semántica del término *traducción automática*, conviene establecer, en primer lugar, una distinción terminológica necesaria entre los diversos términos utilizados para referirse a las distintas modalidades de traducción que resultan de los diferentes grados de colaboración entre las personas y los ordenadores (Gómez Guinovart, 1996a; Hutchins y Somers, 1992). Así, es preciso distinguir cuidadosamente entre los siguientes dos conceptos:

a. La traducción totalmente automática (de gran calidad).

b. La traducción asistida por ordenador.

4.2.1. La traducción totalmente automática

El término *traducción totalmente automática de gran calidad* fue acuñado en 1960 por Yehoshua Bar-Hillel para referirse al objetivo final e ideal de la investigación sobre la traducción por ordenador. Con este término se suele hacer referencia a un programa informático, aún inexistente, capaz de traducir cualquier texto de cualquier género textual entre dos lenguas, sin que importe ni la dificultad del texto original, ni la distancia cultural entre las lenguas implicadas. En este sentido particular, la traducción totalmente automática no existe, ni es probable que vaya a existir en un futuro más o menos próximo.

Como única excepción posible a esta afirmación, la traducción totalmente automática ha alcanzado un nivel de fiabilidad semejante al (profesional) humano en dominios muy específicos. Por ejemplo, el sistema Taum-Météo es capaz de traducir los partes meteorológicos del inglés al francés sin apenas intervención humana (se calcula que los textos que necesitan ser revisados no llegan al 5% del total de textos traducidos por este sistema). El Environment Department de Canadá utiliza Taum-Météo desde 1977, y ha pasado de traducir 8,5 millones de palabras al año en 1984, a traducir 17 millones de palabras al año en la actualidad (Isabelle y Bourbeau, 1985; Vasconcellos, 1993).

4.2.2. La traducción asistida por ordenador

Dentro de esta categoría de programas, suele distinguirse entre:

a. La traducción semiautomática (con intervención humana).

b. La traducción (humana) con ayuda del ordenador.

Los programas informáticos de *traducción semiautomática* son programas capaces de ofrecer una traducción del texto original que debe ser controlada por la persona que supervisa su funcionamiento para conseguir una calidad de traducción similar a la profesional humana (Gómez Guinovart, 1997). En general, estos programas producen de manera automática una primera versión en borrador del texto, que debe ser corregida a conciencia para alcanzar una calidad estándar en el mercado de la traducción. En el mundo de la informática personal, los programas de este tipo más populares son los comercializados con diversas denominaciones por la empresa Globalink, mientras que en el ámbito de las estaciones de trabajo dos de los programas mejor conocidos son Systran y Metal.

Con todo, para poder evaluar adecuadamente la calidad de estos programas de traducción semiautomática, no hay que perder de vista el destino que se le vaya a dar a la traducción obtenida. Evidentemente, el grado de exigencia será distinto para una traducción de uso interno en una empresa, que para una traducción de gran tirada que se deba vender en los quioscos. Así, la Comisión de las Comunidades Europeas traduce alrededor de 30 millones de palabras al ano con el programa Systran, lo que representa un 15% del total de las traducciones realizadas por este organismo comunitario (Vasconcellos, 1993). Los documentos traducidos por Systran para la CCE no son objeto de ninguna revisión, ya que son de uso interno y tienen una finalidad meramente informativa.

Además de los programas de traducción semiautomática, existe una amplia gama de aplicaciones informáticas que, a pesar de non estar concebidas específicamente para la labor de la traducción, ocupan un lugar privilegiado entre las herramientas utilizadas por las personas que se dedican a esta actividad. Estas aplicaciones, que se suelen englobar en el término *traducción con ayuda del ordenador*, pueden provenir de tecnologías tales como la ofimática, la telemática o la gestión documental, e incluyen diversas aplicaciones lingüísticas de la informática como el reconocimiento y síntesis del habla, la verificación lingüística automática, los diccionarios electrónicos o los sistemas de gestión de terminología.

Sin embargo, la utilidad informática más característica de la traducción con ayuda del ordenador son los entornos de trabajo con memoria de traducción, representadas en el mundo de la informática personal por TranslationManager de IBM y Translator's Workbench de Trados (Berry, 1992). Estas aplicaciones integran en un único entorno de trabajo herramientas como un procesador de textos especialmente diseñado para traducir, un sistema de administración de proyectos de traducción, un conjunto de diccionarios bilingües acompañados de herramientas de gestión de las bases de datos léxicos y una *memoria de traducción*. La memoria de traducción es una base de datos en la que se almacenan la versión original y traducida de cada una de las frases que traduce el usuario. Cuando el usuario está traduciendo una frase, el programa detecta automáticamente si esa misma frase u otra frase similar ya fue traducida con anterioridad, de manera que se pueda reutilizar la traducción sin necesidad de reescribirla completamente, haciendo las modificaciones que se consideren más oportunas.

4.3. Extracción de información

En primer lugar, conviene distinguir entre los *sistemas de extracción de la información*, cuyo objetivo consiste en descubrir la información importante de un texto, de los *sistemas de catalogación documental automatizada* y de los *sistemas de recuperación de la información textual*.

En los sistemas de catalogación documental automatizada (Moya e Hípola, 1987), el ordenador típicamente trata de determinar de manera general el contenido de los textos, con vistas a su clasificación dentro de una determinada tipología semántica preestablecida. Por ejemplo, un texto bancario puede ser catalogado como «adeudo por domiciliación» y otro como «contrato de cuenta». Las categorías así obtenidas pueden emplearse para la posterior recuperación de los textos, a partir de las consultas formuladas a la base de datos documental.

Los sistemas de recuperación de la información textual (también denominados sistemas de gestión documental) (Codina, 1994a, 1994b; Sosa, 1997) son programas informáticos que permiten automatizar la creación, el mantenimiento y la consulta de bases de datos documentales. Estos sistemas comparan los documentos de la base de datos con las necesidades de información expresadas en las consultas, con el objetivo de seleccionar los documentos relevantes para el usuario del sistema. La búsqueda se agiliza mediante un índice de la base de datos documental (exactamente, un fichero invertido) y una lista de palabras vacías. El lenguaje de interrogación se basa en operadores booleanos y en operadores de proximidad. Estos criterios pueden complementarse con el uso de algoritmos para calcular el índice de relevancia de los documentos recuperados, es decir, el grado en que es pertinente un documento para la necesidad de información que se especifica en la consulta.

Los sistemas de extracción de la información (Hayes, 1994: 2753-2756; Moreno *et al.*, 1993) son sistemas que convierten la información textual de los documentos analizados en información estructurada (por ejemplo, en registros de una base de datos) o en resúmenes muy concisos. Por ejemplo, la información extraída de un artículo periodístico sobre una acción terrorista podría consistir en el número de personas que resultaron afectadas, el nombre de la organización terrorista que perpetró el acto, el lugar donde sucedió, el momento (día y hora) en que ocurrió el suceso, el tipo de acción terrorista llevado a cabo, etc.

El ejemplo no es casual, sino que muestra el tipo de textos y el tipo de información con que se trabajó en la tercera edición de la Message Understanding Conference (MUC-3) (Chinchor *et al.*, 1993). El MUC reúne cada dos años a diversas empresas y universidades que compiten por presentar el sistema de extracción de información de mayor rendimiento. Los resultados globales de los programas presentados a concurso en el MUC-3, en lo que respecta a la fiabilidad de los datos extraídos automáticamente, se pueden cuantificar en torno al 50%.

A pesar de que estos resultados, desde el punto de vista operativo, pueden considerarse insuficientes, el futuro de la extracción de información resulta bastante prometedor en aplicaciones diseñadas para procesar documentos que respondan a tipos textuales muy específicos, como los pronósticos del tiempo para la navegación marítima o los mensajes cursados por télex en las transferencias bancarias internacionales.

5. Fuentes de información

5.1. Revistas

La publicación científica más importante y de mayor prestigio internacional sobre lingüística computacional es la revista trimestral *Computational Linguistics* (Cambridge: The MIT Press). Otras publicaciones internacionales de gran alcance son: *Literary and Linguistic Computing* (Oxford: Oxford University Press), enfocada hacia la informática aplicada a la investigación lingüística y literaria; *Computers and the Humanities* (Dordrecht: Kluwer), sobre las aplicaciones de la informática a las humanidades en general y también a la lingüística; *Machine Translation* (Dordrecht: Kluwer), dedicada específicamente al campo de la traducción automática; *Journal of Logic, Language and Information*, sobre los aspectos lógicos y computacionales de los lenguajes naturales y los lenguajes formales (Dordrecht: Kluwer); y *Natural Language Engineering* (Cambridge: Cambridge University Press), sobre aplicaciones prácticas de la lingüística computacional.

En España, la única publicación científica regular dedicada íntegramente a esta disciplina es la *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*, de periodicidad semestral y distribución limitada a los miembros de esta asociación.

La revista mensual *BYTE España* (Barcelona: MC Ediciones) publica con cierta asiduidad reseñas sobre aplicaciones lingüísticas de la informática, y en ella he publicado evaluaciones de los sistemas de dictado (mencionados con anterioridad en el apartado 4.1.1) *Vocal Works* de Dragon Systems (Nº 32, septiembre 1997, p. 155), *Simply Speaking Gold* de IBM (Nº 34, noviembre 1997, p. 36), *ViaVoice* de IBM (Nº 35, diciembre 1997, p. 24) y *Speak Naturally* de Dragon Systems (Nº 36, enero 1998); de los programas de traducción semiautomática *Telegraph* de Globalink (Nº 22, octubre 1996, p. 37) y *T1* de GMS (ídem, p. 52); del entorno de trabajo con memoria de traducción *TranslationManager* de IBM (Nº 31, julio-agosto 1997, p. 41); y de diversos diccionarios electrónicos de consulta, como el de la Real Academia Española (Nº 13, diciembre 1995, p. 174), *Le grand Robert électronique* (Nº 17, abril 1996, p. 77), el *Diccionario de uso* de María Moliner (Nº 21, septiembre 1996, p. 63), el *Gran diccionario de la lengua española* de Larousse-Planeta (Nº 25, enero 1997, p. 49), el diccionario del inglés de Merriam-Webster (núm. 27, marzo 1997, p. 43) o el *Diccionari de freqüències* del Institut d'Estudis Catalans (Nº 36, enero 1998).

5.2. Asociaciones e instituciones

A nivel internacional, la asociación profesional más importante es la *Association for Computational Linguistics* (ACL) <http://www.aclweb.org>. Otras asociaciones de ámbito internacional son: *Association for Literary and Linguistic Computing* (ALLC), *Association for the Computers and the Humanities* (ACH); *International Association for Machine Translation* (IAMT) y *European Association for Logic, Language and Information*.

En España, la *Sociedad Española para el Procesamiento del Lenguaje Natural*, constituida en 1983, agrupa a unos 300 profesionales y estudiosos de todas las vertientes de la lingüística computacional. Sus actividades se centran en la organización de un congreso de periodicidad anual, la edición semestral de una revista de carácter científico, la gestión de una lista electrónica <sidsepln@si.ehu.es> y el mantenimiento de un servidor de información a través de la Web <http://gplsi.dlsi.ua.es/SEPLN>.

En el ámbito institucional, el *Observatorio Español de Industrias de la Lengua*, <http://www.cervantes.es/oeil/Oeiles.htm> creado por el *Instituto Cervantes*, se encarga de promover la ingeniería lingüística en España. Sus principales actividades son la difusión de información a la comunidad investigadora y el establecimiento de contactos entre el mundo académico y el empresarial.

5.3. Catálogos de recursos en Internet

La página Web más adecuada para iniciar una búsqueda en este campo de investigación y desarrollo es, sin lugar a dudas, *The ACL NLP/CL Universe* <http://www.cs.columbia.edu/~radev/cgi-bin/universe.cgi>. Se trata de una página creada por el profesor Dragomir R. Radev de la Universidad de Columbia, con índices a cientos de recursos relacionados con la lingüística computacional y el procesamiento del lenguaje natural.

En el campo de las tecnologías del habla, la *página WWW Information for Speech/Acoustics Research* <http://www.aist-nara.ac.jp/IS/Shikano-lab/database/internet-resource/e-www-site.html> contiene una colección muy completa de enlaces Web relacionados con las tecnologías de la voz, elaborada por el profesor Kiyohiro Shikano (del instituto NAIST, en Japón). Contiene información sobre universidades, centros de investigación públicos y privados, bases de datos de habla, recursos informáticos, revistas y congresos, en relación con el reconocimiento y generación de los sonidos del lenguaje.

Bibliografía

ABAITUA, Joseba. (1996). «Ingeniería de la lengua y normalización lingüística». En: FORCADA, Vicente; *et. al.* (eds.). *Estudios computacionales del español y el inglés*. Madrid: Instituto Cervantes, p. 243-259.

ABAITUA, Joseba. (1997). *La traducción automática: presente y futuro*. URL: <http://www.uvigo.es/webs/h06/webh06/sli/paxinas/abaitua.html>.

ABAITUA, Joseba; CASILLAS, Arantza; MARTÍNEZ, Raquel. (1997). «Segmentación de corpus paralelos para memorias de traducción». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XXI, p. 17-30.

ADURIZ, I.; *et al.* (1994). «Euslem: Un lematizador/etiquetador de textos en euskara». En: *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Córdoba: Universidad de Córdoba.

AGIRRE, E.; *et al.* (1989). «Aplicación de la morfología de dos niveles al euskara». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. VIII, p. 87-102.

AGUILAR, Lourdes; *et al.*(1994). «Diseño de pruebas para la evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla». En: *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Córdoba: Universidad de Córdoba.

AGUILAR-AMAT, Ana. (1996). «Los problemas lingüísticos de la traducción automática». En: GÓMEZ GUINOVART, Javier; LORENZO, Anxo (eds.). *Lingüística e informática*. Santiago de Compostela: Tórculo Edicións, p. 187-248.

ALLEN, James. (1995). *Natural Language Understanding*. Redwood: Benjamin/Cummings.

ARRARTE, Gerardo; LLISTERRI, Joaquim. (1994). *Informe sobre recursos lingüísticos para el español (I): Corpus escritos y orales disponibles y en desarrollo en España*. Madrid: Instituto Cervantes.

APPELT, Douglas. (1985). *Planning English Sentences*. Cambridge: Cambridge University Press.

BADIA, Toni. (1997). «El processament computacional de corpus. Tècniques automàtiques d'anàlisi morfològica i sintàctica». En: PAYRATÓ, Lluís; *et al.* (eds.). *Corpus, corpora*. Barcelona: PPU, p. 217-254.

BADIA, Toni; COLOMINAS, Carme. (1995). «Propuesta para una representación semántica de la estructura de predicado y argumentos». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVII, p. 210-224.

BALARI, Sergi. (1992). «Sujetos nulos en HPSG». En: MARTÍN VIDE, Carlos (ed.). *Lenguajes naturales y lenguajes formales VII*. Barcelona: PPU, p. 279-286.

BERRY, Mark. (1992). «The Trados Translator's Workbench II». En: *Proceedings of the 33rd Annual Conference of the American Translators Association*. San Diego: ATA, p. 285-292.

BORSLEY, Robert. (1996). *Modern Phrase Structure Grammar*. Cambridge: Blackwell.

BRESNAN, Joan (ed.). (1982). *The Mental Representation of Grammatical Relations*. Cambridge: MIT Press.

BROWN, Peter F.; *et al.* (1990). «A Statistical Approach to Machine Translation». *Computational Linguistics*. Vol. XVI, núm. 2, p. 79-85.

BROWN, Peter F.; *et al.* (1993). «The Mathematics of Statistical Machine Translation: Parameter Estimation». *Computational Linguistics*. Vol. XIX, núm. 2, p. 263-311.

BULLÓN, José L.; PÉREZ, Juan C. (1994). «Conversión de texto a voz en castellano aplicando el algoritmo PSOLA». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIV, p. 217-231.

CARPENTER, Bob; PENN, Gerald. (1997). *ALE: The Attribute Logic Engine*. Pittsburgh: Universidad Carnegie Mellon.

CARULLA, Marta; OOSTERHOFF, Auke. (1996). «El tratamiento de la morfología flexiva del castellano mediante reglas de dos niveles en una gramática de unificación». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIX, p. 72-80.

CASTELLÓN, Irene; *et al.* (1997). «Propuesta de alternancias de diátesis verbales para el español y el catalán». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol XXI, p. 31-48.

CHINCHOR, Nancy; HIRSHMAN, Lynette; LEWIS, David. (1993). «Evaluating Message Understanding Systems: An Analysis of the Third Message Understanding Conference (MUC-3)». *Computational Linguistics*. Vol. XIX, núm. 3, p. 409-449.

CHURCH, Kenneth; HOVY, Eduard. (1993). «Good Applications for Crummy Machine Translation». *Machine Translation*. Vol. VIII, p. 239-253.

CLIMENT, Salvador. (1995). «La semántica del adjetivo y su representación mediante estructuras de rasgos». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVI, p. 1-14.

CLIMENT, Salvador; FARRERES, Xavier. (1996). «Implementando HPSG en ALE». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVIII, p. 27-42.

CODINA, Lluís. (1994a). «Sistemas automáticos de recuperación de información textual». En: GÓMEZ GUINOVART, Javier (ed.). *Aplicaciones lingüísticas de la informática*. Santiago de Compostela: Tórculo Edicións, p. 63-80.

CODINA, Lluís. (1994b). «Sistemas de gestión documentales: estado del arte y estrategias de utilización». *Binary*. Vol. LXII, p. 114-119 (I); vol. LXIII, p. 92-100 (II); y vol. LXIV, p. 106-112 (III).

COVINGTON, Michael. (1990). «Parsing Discontinuous Constituents in Dependency Grammar». *Computational Linguistics*. Vol. XVI, núm. 4, p. 237-240.

COVINGTON, Michael. (1994). *Natural Language Processing for Prolog Programmers*. Englewood Cliffs: Prentice-Hall.

DIK, Simon. (1992). *Functional Grammar in Prolog*. Berlín: Mouton de Gruyter.

EVANS, Roger; GAZDAR, Gerald. (1996). «DATR: A Language for Lexical Knowledge Representation». *Computational Linguistics*. Vol. XXII, nº 2, p. 167-216.

FARWELL, David; HELMREICH, Stephen; CASPER, Mark. (1995). «SPOST: a Spanish Part-of-Speech Tagger». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVII, p. 42-53.

FERNÁNDEZ, Adelaida; LLISTERRI, Joaquim. (1996). *Informe sobre recursos lingüísticos para el español (II): Corpus escritos y orales disponibles y en desarrollo en España*. Madrid: Instituto Cervantes.

FERRARI, Giacomo. (1991). «Towards a Realistic Dialogue Model». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XI, p. 11-22.

FILLMORE, Charles J.; ATKINS, B. T. S. (1994). «Starting where the Dictionary Stop: The Challenge of Corpus Lexicography». En: ATKINS, B. T. S.; ZAMPOLLI, Antonio (eds.). *Computational Approaches to the Lexicon*. Oxford: Oxford University Press, p. 349-393.

GALE, William A.; CHURCH, Kenneth W. (1993). «A Program for Aligning Sentences in Bilingual Corpora». *Computational Linguistics*. Vol. XIX, nº 1, p. 75-102.

GARCÍA-MIGUEL, José M. (1994). «Corpus de textos analizados sintácticamente». En: GÓMEZ GUINOVART, Javier (ed.). *Aplicaciones lingüísticas de la informática*. Santiago de Compostela: Tórculo Edicións, p. 19-33.

GARCÍA-MIGUEL, José M.; VÁZQUEZ, Victoria. (1994). «Lingüística de corpus y lingüística descriptiva: el caso de la duplicación de objetos». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIV, p. 47-62.

GARRIDO, Juan M. (1991). «Estilización de patrones melódicos del español para sistemas de conversión texto-habla». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XI, p. 209-219.

GARSIDE, Roger. (1993). «The Large-Scale Production of Syntactically Analysed Corpora». *Literary and Linguistic Computing*. Vol. VIII, N° 1, p. 39-46.

GAZDAR, Gerald; *et al.* (1985). *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

GAZDAR, Gerald; MELLISH, Chris. (1989). *Natural Language Processing in Prolog: An Introduction to Computational Linguistics*, Wokingham: Addison-Wesley.

GOJENOLA, Koldo; SARASOLA, Kepa. (1994). «Aplicación de la relajación gradual de restricciones para la detección y corrección de errores sintácticos». En: *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Córdoba: Universidad de Córdoba.

GÓMEZ GUINOVART, Javier. (1996a). «Traducción automática e traducción asistida por ordenador: aspectos terminolóxicos e tipoloxía». *Viceversa: Revista Galega de Traducción*. Vol. II, p. 99-103.

GÓMEZ GUINOVART. Javier. (1996b). *Fundamentos y límites de los sistemas de verificación automática de la sintaxis y el estilo*. Santiago de Compostela: Universidad de Santiago de Compostela.

GÓMEZ GUINOVART, Javier. (1997). «Traducción automática inglés-español: estado del arte». En: FERNÁNDEZ-CORUGEDO, Santiago (ed.). *Some Sundry Wits Gathered Together*. A Coruña: Universidade da Coruña, p. 31-40.

HALLEBEEK, Jos. (1992). *A Formal Approach to Spanish Syntax*. Amsterdam: Rodopi.

HALLIDAY, M. A. K. (1991). «Corpus Studies and Probabilistic Grammar». En: AIJMER, Karin; ALTENBERG, Bengt (eds.). *English Corpus Linguistics. Studies in Honour of Jan Svartvik*. Londres: Longman, p. 30-43.

HAYES, Philip. (1994). «Natural Language Processing: Applications». En: ASHER, R. E. (ed.). *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon Press, p. 2.748-2.757.

HERNÁEZ, I. *et al.* (1995). «Curvas de F0 en euskara: Primera aproximación a la obtención de modelos para conversión de texto a voz». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVII, p. 272-286.

HERNANDO, Javier; NADEU, Climent; MARIÑO, José. (1997). «Técnicas robustas de reconocimiento del habla en ambientes adversos». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XX, p. 27-43.

HOVY, Eduard. (1988). *Generating Natural Language under Pragmatic Constraints*. Hillsdale: Lawrence Erlbaum.

HUTCHINS, John; SOMERS, Harold. (1992). *An Introduction to Machine Translation*. Londres: Academic Press.

ISABELLE, Pierre; BOURBEAU, Laurent. (1985). «Taum-Aviation: Its Technical Features and Some Experimental Results». *Computational Linguistics*. Vol. XI, n° 1, p. 18-27.

JONES, Daniel. (1996). *Analogical Natural Language Processing*. Londres: UCL.

JUCKER, Andreas. (1994). «New Dimensions in Vocabulary Studies: Review article of the Oxford English Dictionary (2nd edition) on CD-ROM». *Literary and Linguistic Computing*. Vol. IX, n° 2, p. 149-154

KAY, Martin. (1982). «Parsing in Functional Unification Grammar». En: DOWTY, David; KARTUNNEN, Lauri; y ZWICKY, Arnold (eds.). *Natural Language Parsing*. Cambridge: Cambridge University Press, p. 251-278.

KAY, Martin; RÖSCHEISEN, Martin. (1993). «Text-Translation Alignement». *Computational Linguistics*. Vol. XIX, n° 1, p. 121-142.

KELLER, E. (ed.). (1994). *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art and Future Challenges*, Chichester: John Wiley & Sons.

KOSKENNIEMI, Kimmo. (1983). *Two-level morphology: A general computational model for word-form recognition and production*. Helsinki: Universidad de Helsinki.

LEZCANO, Emma; PÉREZ GUERRA, Javier; SEOANE, Elena. (1997). «English Corpus Linguistics and Historical Research». En: FERNÁNDEZ-CORUGEDO, Santiago (ed.). *Some Sundry Wits Gathered Together*. A Coruña: Universidade da Coruña, p. 73-98.

LLISTERRI, Joaquim. (1991). *Introducción a la fonética: el método experimental*, Barcelona: Anthropos.

LÓPEZ GONZALO, E.; *et al.* (1994). «Modelado lingüístico y acústico para un sistema de conversión de texto a habla». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIV, p. 257-272.

LORENZO, Anxo; GÓMEZ GUINOVART, Javier. (1996). «Terminoloxía, informática e lingua galega». *Cadernos de lingua*. Vol. 13, p. 5-33.

MARCOS MARÍN, Francisco A. (1994). *Informática y humanidades*. Madrid: Gredos.

MARCUS, Mitchell P.; SANTORINI, Beatrice; MARCINKIEWICZ, Mary Ann. (1993). «Building a Large Annotated Corpus of English: the Penn Treebank». *Computational Linguistics*. Vol. XIX, n 2, p. 313-330.

MARTÍ, Josep; GUDAYOL, Francesc. (1994). «El ritmo y la entonación en la lectura del castellano». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIV, p. 273-287

McENERY, Tony; WILSON, Andrew. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

MINSKY, Marvin (ed.). (1967). *Semantic Information Processing*. Cambridge: Cambridge: The MIT Press.

MITTON, Roger. (1996). *English Spelling and the Computer*. Londres: Longman.

MOURE, Teresa; LLISTERRI, Joaquim. (1996). «Lenguaje y nuevas tecnologías: el campo de la lingüística computacional». En: FERNÁNDEZ, Milagros (coord.). *Avances en lingüística aplicada*. Santiago de Compostela: Universidad de Santiago de Compostela, p. 147-227.

MOYA, Félix; HÍPOLA, Pedro. (1987). «Problemas lingüísticos en la automatización de los sistemas de clasificación documental». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. V, p. 74-85.

PALOMAR, Manolo; FERRÁNDEZ, Antonio; MORENO, Lidia. (1995). «Aportaciones a la resolución de la elipsis en la coordinación». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVII, p. 101-114.

PEREIRA, Fernando; SHIEBER, Stuart. (1987). *Prolog and Natural-Language Analysis*. Stanford: CSLI.

PEREIRA, Fernando; WARREN, David. (1980). «Definite Clause Grammars for Language Analysis». *Artificial Intelligence*. Vol. XIII, p. 231-278.

PÉREZ, Ricard; TROTZIG, David; LLORÉ, Xavier. (1994). «Morfeo: Analizador morfológico y "tagger" del español». En: *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Córdoba: Universidad de Córdoba.

POLLARD, Carl; SAG, Ivan. (1994). *Head-Driven Phrase Structure Grammar*, Stanford: CSLI.

RAFEL, Joaquim (dir.). (1996). *Diccionari de freqüències. 1: Llengua no literària*. Diccionari del Català Contemporani, Corpus Textual Informatitzat de la Llengua Catalana. Barcelona: Institut d'Estudis Catalans.

RAFEL, Joaquim. (1997). «El Diccionari del català contemporani i el corpus textual informatitzat de la llengua catalana». En: PAYRATÓ, Lluís; *et al.* (eds.). *Corpus, corpora*. Barcelona: PPU, p. 71-92.

RAMÍREZ, Flora; SÁNCHEZ LEÓN, Fernando. (1996). «GramCheck: Un corrector gramatical para español». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIX, p. 30-37.

ROBERTSON, Alexander M.; WILLETT, Peter. (1993). «A Comparison of Spelling-Correction Methods for the Identification of Word Forms in Historical Text Databases». *Literary and Linguistic Computing*. Vol. VIII, nº 3, p. 143-152.

RODRÍGUEZ MAGRO, Consuelo. (1993). *Corrector: un sistema de verificación gramatical y estilística de textos basado en una gramática robusta*. Madrid: Universidad Autónoma de Madrid.

ROJO, Guillermo. (1992). «El futuro "Diccionario de construcciones verbales del español actual"». En: MARTÍN VIDE, Carlos (ed.). *Lenguajes naturales y lenguajes formales VIII*. Barcelona: PPU, p. 41-50.

ROJO, Guillermo. (1993). «La base de datos sintácticos del español actual». *Español Actual*. Vol. LIX, p. 15-20.

RUIZ, Juan Carlos. (1993). «GFU-LAB: Un sistema computacional para la co-descripción de la sintaxis y la semántica». En: MARTÍN VIDE, Carlos (ed.). *Lenguajes naturales y lenguajes formales IX*. Barcelona: PPU, p. 237-248.

RUIZ, Juan Carlos. (1996). «Modelos de análisis sintáctico en el procesamiento del lenguaje natural». En: GÓMEZ GUINOVART, Javier; LORENZO, Anxo (eds.). *Lingüística e informática*. Santiago de Compostela: Tórculo Edicións, p. 31-55.

RUIZ, Juan Carlos; ABAITUA, Joseba; ZUBIZARRETA, Ramón. (1991). «Un compilador de LFG y su aplicación al euskara». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVII, p. 177-187.

RUIZ, Juan Carlos; GÓMEZ GUINOVART, Javier. (1990). «Aproximación al tratamiento computacional del modelo de rección y ligamiento». En: MARTÍN VIDE, Carlos (ed.). *Lenguajes naturales y lenguajes formales V*. Barcelona: Universidad de Barcelona, p. 655-664.

SÁNCHEZ LEÓN, Fernando; NIETO SERRANO, Amalio. (1995). «Desarrollo de un etiquetador morfosintáctico para el español». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XVII, p. 14-28.

SANTAMARINA, Antón. (1996). «Informática e lexicografía». En: GÓMEZ GUINOVART, Javier; LORENZO, Anxo (eds.). *Lingüística e informática*. Santiago de Compostela: Tórculo Edicións, p. 9-29.

SELLS, Peter. (1985). *Lectures on Contemporary Syntactic Theories*. Stanford: CSLI.

SHIEBER, Stuart. (1986). *An Introduction to Unification-Based Approaches to Grammar*. Stanford: CSLI.

SINCLAIR, John (ed.). (1987). *Looking Up: an account of the COBUILD Project in Lexical Computing*, Londres: Collins.

SOUTER, Clive; O'DONOGHUE, Tim F. (1991). «Probabilistic Parsing in the COMMUNAL Project». En: JOHANSSON, Stig; STENSTRÖM, Anna-Brita (eds.). *English Computer Corpora*. Berlín: Mouton de Gruyter, p. 33-48.

SOSA, Eduard. (1997). «Sistemes de recuperació d'informació i processament del llenguatge natural». En: CID, Pilar; BARÓ, Jaume (eds.). *Anuario SOCADI de Documentación e Información 1997*. Barcelona: SOCADI, p. 129-135.

TAULÉ, Mariona; CASTELLÓN, Irene. (1994). «Generación de alternancias de subcategorización mediante reglas léxicas». *Revista de la Sociedad Española para el Procesamiento del Lenguaje Natural*. Vol. XIV, p. 335-352.

VASCONCELLOS, Muriel. (1993). «The Current State of MT Usage». *MT News International*. Vol. VI, p. 12-17.

WHITELOCK, Peter; KILBY, Kieran. (1995). *Linguistic and Computational Techniques in Machine Translation System Design*. Londres: UCL.

WILKS, Yorik, SLATOR, Brian; GUTHRIE, Louise. (1996). *Electric Words: Dictionaries, Computers, and Meanings*. Cambridge: The MIT Press.