PROCESAMIENTO DEL LENGUAJE NATURAL: DE LINGÜISTA A INGENIERO DEL CONOCIMIENTO

Carlos Periñán Pascual

1. LA CUESTIÓN TERMINOLÓGICA

La convergencia entre la lingüística y la informática ha resultado en un nuevo perfil profesional que requiere una formación específica. La evolución experimentada en este campo interdisciplinar durante los últimos años ha dado lugar a etiquetas como lingüística computacional, procesamiento del lenguaje natural (PLN), tecnologías lingüísticas, ingeniería lingüística, industrias de la lengua o lingüística informática. En este apartado, explicamos por qué la lingüística informática y la ingeniería lingüística son campos de investigación propios de la lingüística y la informática respectivamente, mientras que términos como lingüística computacional, PLN y tecnologías lingüísticas suelen referirse a una misma área de conocimiento aunque enfatizando aspectos diferentes dependiendo del punto de vista de la disciplina que la estudie (figura 1).

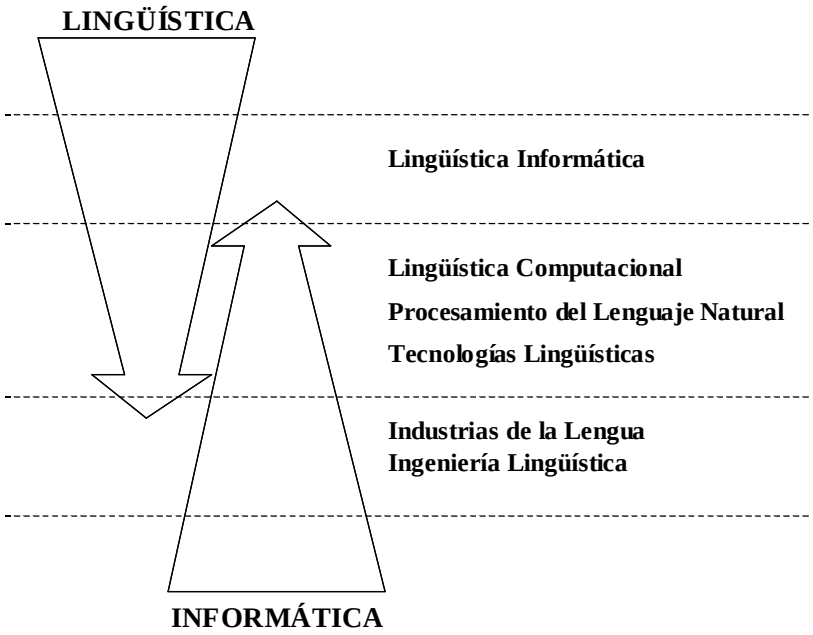

Figura 1. Campos de estudio en la convergencia de la lingüística y la informática.

En un sentido estricto, cualquier actividad que implique un análisis o generación de la

lengua utilizando el ordenador puede considerarse como lingüística computacional. Debido a la finalidad práctica de las investigaciones, los lingüistas prefieren hablar de la lingüística computacional como un área de conocimiento dentro de la lingüística aplicada. En cambio, debido a la posibilidad de desarrollar sistemas de computación que simulen algún aspecto de la capacidad lingüística del ser humano, los informáticos consideran la lingüística computacional como una rama de la inteligencia artificial, en cuyo caso prefieren hablar de PLN. Por tanto, mientras la lingüística computacional se centra más en la modelización del conocimiento lingüístico para posibilitar la construcción de sistemas computacionales que analicen y/o generen textos en lenguaje natural, el PLN hace un mayor énfasis en la búsqueda de soluciones a los problemas que plantea la lingüística computacional pero en el marco de aplicaciones concretas: p.ej. recuperación y extracción de información, resúmenes automáticos, traducción mecánica, etc. (Martí Antonín 2003). Finalmente, se prefiere el término tecnología del lenguaje cuando describimos cómo esas aplicaciones del PLN mejoran la comunicación en la sociedad de la información por encima de las barreras que impone la distancia, el uso de lenguas distintas o el modo en que tiene lugar la comunicación (Martí Antonín y Llisterri 2001). Debido a que la diferencia sólo radica en el prisma desde el que se observe, actualmente los términos lingüística computacional, PLN y tecnologías lingüísticas suelen ser utilizados de forma indiscriminada tanto por lingüistas como informáticos para hacer referencia básicamente a una misma disciplina.

Un tipo de investigación marcadamente diferente se encuentra en la "lingüística informática", la cual estaría orientada hacia el desarrollo de programas de apoyo en los estudios realizados en los diversos campos de la filología (Martí Antonín 2003). La principal finalidad de este tipo de programas es la extracción de datos estadísticos, concordancias, colocaciones, etc a partir de una extensa colección de textos. Por ejemplo, obtener información estadística sobre la aparición de determinadas unidades lingüísticas resulta útil tanto en la descripción de la lengua como en la selección del vocabulario y las construcciones más usuales para la elaboración de programas de enseñanza de lenguas (Moure y Llisterri 1996). Con el fin de que las herramientas informáticas puedan extraer la mayor información léxica posible con el menor esfuerzo humano, es necesario convertir la colección de textos en un corpus. En este sentido, la lingüística informática permite llegar a una serie de conclusiones objetivas que serán contrastadas con nuestras hipótesis iniciales, aplicando así un método hipotético-deductivo: es decir, observamos algún fenómeno concreto de la realidad lingüística, formulamos una hipótesis que se origina de nuestra propia introspección, aplicamos las herramientas informáticas a un inventario de datos léxicos y, por último,

verificamos la hipótesis inicial confrontándola con los resultados obtenidos. La lingüística informática permite igualmente aplicar un método inductivo: es decir, recopilamos un conjunto de datos lingüísticos a través de las herramientas informáticas, aplicamos a esos datos una interpretación y, finalmente, confeccionamos una serie de reglas. En conclusión, el uso de instrumentos informáticos de análisis de córpora permite integrar la objetividad y la subjetividad características de todo método científico en nuestra investigación lingüística (Candalija Reina 1998).

Un área muy diferente donde convergen lingüística e informática la encontramos en la "ingeniería lingüística", también conocida como "industrias de la lengua". Estos términos referencian un campo de investigación todavía en desarrollo que sirve para describir aquellos productos comerciales en los que se aplican técnicas propias del PLN: p.ej. los traductores automáticos, los correctores ortográficos y gramaticales, los programas de enseñanza de lenguas extranjeras asistida por ordenador, etc. Por tanto, este tipo de disciplina se caracteriza por estar estrechamente vinculado al mundo empresarial, el cual desarrolla y comercializa una serie de productos dirigidos a unos usuarios finales no especializados que poseen unas necesidades específicas (Moure y Llisterri 1996). Hasta los años 90, las investigaciones en lingüística computacional implicaban el diseño y la creación de prototipos sofisticados basados en complejos modelos formales teóricos. Cuando comenzó la ingeniería de los productos, algunos de los supuestos teóricos más sólidos se derrumbaron. Por ejemplo, se abandonó la premisa implícita de que cuanto más complejo es un problema, más teóricamente sofisticada debe ser su solución, lo cual favoreció la aparición de numerosos estudios empíricos dentro de investigaciones menos complejas desde un punto de vista lingüístico. El desplazamiento de la figura del lingüista computacional en este tipo de proyectos permite a Ferrari (2004) afirmar que los lingüistas computacionales no deben participar en la ingeniería de los productos, sino dedicarse exclusivamente a la producción de investigación avanzada.

Existen numerosos libros, dirigidos a investigadores, profesores y estudiantes de lingüística o informática, que proporcionan una visión general de la lingüística computacional y el PLN[1]. Estos libros suelen describir los principales métodos o procesos implicados en los sistemas del PLN y los escenarios típicos en que se implementan las aplicaciones de las tecnologías lingüísticas. Además, el lector suele encontrar allí numerosas referencias

[1] Por ejemplo, Grishman (1986) es uno de los primeros libros de texto que describe el campo de la lingüística computacional desde un enfoque eminentemente lingüístico. Por otra parte, si deseamos una presentación del las principales líneas de investigación en el panorama actual del PLN y las tecnologías lingüísticas, nuestras recomendaciones son Varile y Zampolli (1995) y Mitkov (2003), en los cuales se hace un mayor énfasis en los aspectos computacionales.

bibliográficas que le permitirán ampliar sus conocimientos en temas más específicos de esta disciplina, además de una selección comentada de recursos electrónicos (p.ej. software, córpora, etc) disponibles en Internet. Desmarcándonos de la tendencia habitual al presentar esta disciplina, el objetivo de este capítulo gira en torno a la tesis de que es preciso que el lingüista se convierta en ingeniero del conocimiento si deseamos participar activamente en el desarrollo de un proyecto del PLN que resulte finalmente en un producto de la ingeniería lingüística. Como telón de fondo, el apartado 2 resalta la relevancia de las tecnologías lingüísticas en la sociedad actual, mientras el apartado 3 describe a modo de ejemplo el funcionamiento de una de las tareas del PLN que más investigadores ha atraído en los últimos años, i.e. la recuperación de información. El apartado 4 justifica el papel decisivo que desempeña la base de conocimiento para el éxito de un sistema del PLN. Finalmente, el apartado 5 describe las principales directrices que posibilitan el diseño de una buena base de conocimiento.

2. LA SOCIEDAD DE LA INFORMACIÓN Y LAS TECNOLOGÍAS LINGÜÍSTICAS

Todos coincidimos en que la revolución tecnológica que vivimos actualmente ha posibilitado la creación de una "sociedad de la información", pero esta etiqueta extraordinariamente popular adquiere diferentes significados dependiendo del contexto en la que se utilice. El término no sólo se asocia a la capacidad de almacenamiento, manipulación y transmisión de grandes cantidades de datos gracias a la convergencia de los ordenadores y las telecomunicaciones, sino también suele vincularse con el desarrollo de la cohesión cultural y la integración de las comunidades en la Unión Europea (Browne 1997). En esta sociedad de la información, el lenguaje desempeña un papel primordial, ya que es el medio principal a través del cual las personas representamos e intercambiamos ideas, transmitimos conocimientos y, finalmente, creamos cultura. En Europa, disfrutamos de una gran diversidad lingüística, por tanto debemos explorar formas en las que superar las barreras hacia la comunicación y el entendimiento. La solución más inmediata implicaría la utilización de una sola *lingua franca* para las actividades internacionales en el ámbito de los negocios, la administración y la política. Aunque a primera vista podría parecer una medida satisfactoria, el dominio de unas pocas lenguas desembocaría en la ausencia de una verdadera sociedad multicultural, el desequilibrio del poder y el uso empobrecido de los recursos (Comisión Europea 1997).

En realidad, una solución más factible se encuentra en las tecnologías lingüísticas, las cuales posibilitan la comunicación entre personas y máquinas utilizando las destrezas de la comunicación natural. El campo de las tecnologías lingüísticas cubre un amplio abanico de tareas: p.ej. recuperación y extracción de información, resúmenes automáticos, traducción mecánica, sistemas de diálogo persona-máquina, etc. A modo de ejemplo, el siguiente apartado describe el funcionamiento y las principales aplicaciones de los sistemas de recuperación de información.

3. UN EJEMPLO DE SISTEMA DEL PROCESAMIENTO DEL LENGUAJE NATURAL

La recuperación de información consiste en seleccionar automáticamente, a partir de una colección muy extensa de documentos (i.e. textos, imágenes, audio, etc), aquéllos que sirvan como respuesta a la consulta textual del usuario[2]. Hasta hace poco, los sistemas de recuperación de información se centraban en la gestión de bases documentales de contenido muy especializado: p.ej. ISBN, BOE, MEDLINE o WESTLAW. El perfil característico de los usuarios de estos sistemas era un profesional familiarizado no sólo con el contenido de la base de datos sino también con la sintaxis específica en la que formulaban las consultas. En la actualidad, la base de datos documental por excelencia es Internet y los sistemas de recuperación de información adoptan la forma de buscadores como *Google, Yahoo, Altavista*, etc.

La creación de Internet ha permitido pasar de una civilización industrial local a una globalizada. En consecuencia, el acceso a la información se ha democratizado gracias a Internet (Martí Antonín y Llisterri 2001). No obstante, el primer problema que se plantea es el cambio de significado del concepto "información"; ahora no está mejor informado quien tiene más datos, sino quien dispone de los mejores medios para obtener únicamente los que necesita (*ídem*). En este sentido, los sistemas actuales de recuperación de información fracasan en su cometido cuando el resultado de la búsqueda es una lista de documentos excesivamente larga, o bien cuando no se encuentra ningún documento relevante de acuerdo con la consulta del usuario. Ambos casos son lamentablemente bastante frecuentes, debido a que los usuarios formulan sus consultas de forma poco precisa, además del hecho de que

[2] En realidad, un término más apropiado para referirnos a este campo del conocimiento sería "recuperación de documentos" (Harman, Schäuble *et al.* 1995).

carecemos de métodos de almacenamiento y codificación que permitan convertir este acopio de datos en información organizada y estructurada (i.e. conocimiento). Nos encontramos ante la situación paradójica de que tenemos al alcance de nuestra mano cantidades ingentes de información pero somos incapaces de obtener la que realmente nos interesa. La principal consecuencia de este problema es la "sobrecarga informativa" que sufre nuestra sociedad, lo cual implica que tener demasiada información puede ser tan peligroso como tener demasiado poca. Ante este problema, los usuarios suelen sufrir el "síndrome de la fatiga informativa", cuyos síntomas más frecuentes son tensión, irritabilidad y frecuentes sensaciones de desamparo (Reuters 1996). Un segundo problema asociado a Internet es la dimensión multilingüística en la que se enmarca. Mientras el inglés es todavía la lengua predominante en Internet, el número de páginas *web* escritas en otros idiomas es cada vez mayor. La necesidad de formular consultas en una lengua diferente a la utilizada en la redacción de los documentos es cada vez más acuciante. Ante estos dos problemas, el reto actual de las tecnologías lingüísticas es hacer que Internet pueda funcionar realmente como una gran base de conocimiento a la cual podamos acceder de manera eficaz y sin limitaciones idiomáticas.

En un sistema típico de recuperación de información, podemos distinguir las siguientes fases: el procesamiento de los documentos, el procesamiento de la consulta y la presentación de los resultados (Gonzalo Arroyo y Verdejo Maíllo 2003; Tzoukermann, Klavans *et al.* 2003). En primer lugar, es necesario identificar los índices, o palabras claves, que describan de forma óptima el contenido del texto y representar estos índices de forma computacionalmente eficiente. La indización de textos puede realizarse manualmente a partir de índices predefinidos o automáticamente a través de técnicas propias de los sistemas de extracción de información. Uno de los aspectos más complejos del proceso de indización automática es encontrar la correspondencia entre las unidades léxicas y los índices, presentándose casos problemáticos como las variantes flexivas y derivativas, las expresiones multipalabra o los sinónimos de una unidad léxica. En estos casos, recursos lingüísticos como lexicones y ontologías pueden resultar muy valiosos. Una vez construida una lista preliminar de índices, se asigna un peso a cada uno de ellos, donde los valores más altos identifican las palabras claves más relevantes. La asignación de los pesos se realiza de acuerdo a una serie de criterios de relevancia. Por ejemplo, si el índice aparece en todo los documentos de la base de datos, este índice no servirá para discriminar un documento de otro. En cambio, un índice es mucho más significativo si sólo aparece en unos pocos documentos de la colección. En segundo lugar, el sistema debe relacionar la lista de índices con la representación de la consulta, para lo cual se utilizan generalmente técnicas estadísticas o probabilísticas.

Actualmente la estrategia más frecuente para representar la consulta del usuario es una combinación de palabras clave y operadores booleanos. No obstante, podemos mejorar estos sistemas expresando las consultas en lenguaje natural sin la necesidad de limitarnos a una sintaxis restringida y un vocabulario controlado que el usuario deba aprender. Finalmente, el resultado de la búsqueda se presenta como una lista de documentos ordenada decrecientemente de acuerdo con su relevancia.

En aquellos entornos de búsqueda de información en los que la lengua en que está formulada la consulta no coincide con la lengua en la que están redactados los documentos, hablamos de "recuperación de información multilingüe". Esta tarea es crucial en un escenario como Internet, enmarcándose más concretamente en lo que se conoce como "*web* semántica" (Berners-Lee 1998), donde los sistemas que procesan conocimiento requieren la creación de una semántica comprensible por la máquina que permita representar toda la información que contienen los documentos de la *web*[3].

Un sistema de recuperación de información multilingüe puede adoptar uno de los siguientes enfoques: traducción de las consultas, traducción de los documentos o uso de una interlingua (Gonzalo Arroyo y Verdejo Maíllo 2003). La mayor parte de la investigación se ha centrado en la traducción de la consulta a cada una de las lenguas utilizadas en la base documental. No obstante, este enfoque presenta dos inconvenientes principales. Por una parte, una consulta no suele proporcionar suficiente contexto como para asignar automáticamente el sentido en que el usuario utiliza cada término. Como consecuencia, los errores de traducción de los términos que forman la consulta afectan considerablemente a la efectividad del sistema. Por otra parte, la consulta debe traducirse a cada una de las lenguas contempladas en la base documental. El método basado en la traducción de los índices de cada documento a la lengua en que se realiza la consulta presenta la ventaja de que dichas traducciones son más precisas debido al mayor contexto que proporcionan los documentos, además del hecho de que si se produce algún error en la traducción, éste no es tan crucial, ya que el documento dispone de muchos más términos de indización. Sin embargo, teóricamente mejor es el enfoque que permite representar de forma interlingüística, i.e. sin mediación de una lengua natural, tanto los índices documentales como la consulta del usuario. Por ejemplo, al no manejar unidades léxicas sino conceptuales, se proporciona un tratamiento más adecuado a los fenómenos de sinonimia y polisemia. Utilizar conceptos como descriptores de los documentos solventa el

[3] *ContentWeb* (Aguado de Cea, Álvarez de Mon *et al.* 2002), una plataforma que permite la recuperación automática de información por medio del lenguaje natural en aplicaciones de comercio electrónico, propone una anotación de los documentos de la *web* con información sintáctica, semántica y discursiva.

"problema del vocabulario" (Gordon 1999), i.e. diferentes personas pueden referirse al mismo documento de muchas formas diferentes[4]. A pesar de las ventajas evidentes de este enfoque, todavía se está investigando cómo realizar una perfecta transducción de la consulta y los índices documentales a una representación interlingüística.

En general, la mayoría de los investigadores en recuperación de información creen que es más fácil mejorar los resultados de los sistemas de búsqueda por medio de métodos estadísticos que con técnicas propias del PLN (Harman, Schäuble *et al.* 1995; Tzoukermann, Klavans *et al.* 2003). Sin embargo, también coinciden en que los recursos lingüísticos (p.ej. lexicones, ontologías, etc.) están teniendo cada vez un mayor impacto en la efectividad de la recuperación de los documentos. Por ejemplo, durante las fases de procesamiento de los documentos y la consulta, el lexicón permite agrupar los alomorfos bajo un mismo lema y la ontología proporciona un tratamiento adecuado de las relaciones sinonímicas y taxonómicas.

4. EL PROCESAMIENTO DEL LENGUAJE NATURAL Y LOS SISTEMAS EXPERTOS

Antes de describir las tareas que puede desempeñar un lingüista en un proyecto del PLN, es necesario presentar las características principales de este tipo de sistemas. Tradicionalmente el PLN se estudia como una rama de la inteligencia artificial, donde sus otros dos campos principales de investigación son la construcción de sistemas expertos y la robótica. En realidad, podemos concebir un sistema del PLN como un sistema experto, ya que se trata en definitiva de un sistema basado en el conocimiento, integrando así estas dos ramas de la inteligencia artificial bajo el nombre de "ingeniería del conocimiento". La finalidad de los sistemas expertos es producir programas "inteligentes" que permitan resolver problemas complejos en contextos profesionales e industriales. En el caso del PLN, estos problemas implican la implementación computacional de alguna destreza lingüística. Evidentemente cualquier programa de ordenador debe poseer conocimiento en mayor o menor grado para que pueda desempeñar el cometido para el que ha sido diseñado. No obstante, las características esenciales de los sistemas expertos son la representación explícita del conocimiento, a través

[4] Tradicionalmente este problema se ha solventado expandiendo la consulta a través de un tesauro; en otras palabras, el sistema identifica los términos de la consulta, busca en un tesauro términos adicionales que estén directamente vinculados con los términos de la consulta, los términos adicionales se añaden a la consulta original y, finalmente, la nueva consulta expandida se compara con el material de la base de datos (Gordon 1999).

de formalismos como reglas de producción, redes semánticas, marcos o guiones, y la separación modular entre el conocimiento y el resto del sistema (Adarraga 1994).

En el desarrollo de un sistema experto, diferenciamos tres grandes fases: la extracción, el análisis y la codificación del conocimiento. El proceso comienza con la extracción del conocimiento a partir de las conversaciones con el experto humano. Es necesario una descripción en lenguaje natural de la tarea que se pretende modelar. En segundo lugar, construimos un modelo conceptual de las entidades y relaciones que supuestamente usa el experto humano en sus razonamientos. Finalmente, tiene lugar la implementación computacional del modelo resultante en la fase anterior. Con el fin de poder llevar a cabo estas tres fases, un sistema experto se apoya en una base de conocimiento, un motor de inferencia y una interfaz. La base de conocimiento almacena un conjunto de hechos que se suponen que son verdad acerca de un determinado dominio, además de un conjunto de reglas de inferencia capaces de generar nuevos hechos a partir de los existentes. Por otra parte, el motor de inferencia integra e interpreta la información representada en la base de conocimiento y los datos proporcionados por el usuario a través de la interfaz. El comportamiento "inteligente" de un sistema experto depende tanto de la calidad de los datos almacenados en su base de conocimiento como de la eficacia del motor de inferencia (Floridi 1999).

Por consiguiente, si equiparamos un sistema del PLN con un sistema experto podemos deducir que uno de los componentes centrales de nuestra aplicación es la base de conocimiento, por lo cual el lingüista que desee implicarse activamente en un proyecto de este tipo debe convertirse en un ingeniero del conocimiento. Este nuevo perfil profesional no implica que el lingüista se convierta en un informático, pero al menos exige, como apuntan Cassen, Dégremont *et al.* (1991), la adquisición de conocimientos básicos sobre matemáticas y lógica formal (p.ej. cálculos formales, estructura de conjuntos, teoría de grafos...), ingeniería del software (p.ej. diagramas de flujos, estructuración de subrutinas...) y lenguajes de programación con el fin de que los lingüistas e informáticos puedan presentar sus respectivos conocimientos del tal forma que se logre una comunicación interdisciplinar que posibilite un tratamiento oportuno a los problemas que el PLN plantee. En el siguiente apartado, describimos el tipo de conocimiento que podemos utilizar en un sistema del PLN.

5. EL DISEÑO DE UNA BASE DE CONOCIMIENTO PARA EL PROCESAMIENTO DEL LENGUAJE NATURAL

5.1 Introducción

Una base de conocimiento para el PLN debe estar formada al menos por dos componentes: el lexicón y la ontología. El lexicón contiene la información morfosintáctica, semántica y pragmática sobre el comportamiento de las unidades léxicas de una determinada lengua, pero la representación del significado de una unidad léxica debe almacenarse en la ontología, o modelo del mundo. En otras palabras, mientras el lexicón contiene el conocimiento lingüístico, la ontología estructura jerárquicamente el conocimiento del mundo compartido por un hablante medio. En este enfoque, los significados de las unidades léxicas en el lexicón están codificados como "unidades semánticas" a las que se les asigna un concepto ontológico provisto de una rica estructura interna, el cual debe contener un postulado de significado entre sus propiedades. No obstante, la frontera entre conocimiento léxico y ontológico resulta a veces bastante difusa, lo que implica un constante proceso de negociación entre los dos equipos de investigadores (Nirenburg, Beale ***et al.*** 1996).

Una base de conocimiento para el PLN debe ser multipropósito, en el sentido de que sea multilingüística y multifuncional (*ídem*). En otras palabras, no sólo debe estar diseñada para poder trabajar con diversas lenguas naturales, sino además posibilitar su reutilización en diversas tareas del PLN (p.ej. traducción automática, recuperación y extracción de la información, etc). A nivel de diseño, el multilingüismo y la multifuncionalidad influyen enormemente en algunos aspectos del diseño del sistema. Por una parte, construiremos un lexicón para cada una de las lenguas que intervengan en nuestro sistema del PLN, donde la ontología servirá de puente entre las diversas lenguas. Por otra parte, debemos incorporar en nuestra base de conocimiento toda aquella información que pudiera resultar potencialmente útil en una aplicación del PLN. En realidad, el tipo de conocimiento que requiere un sistema del PLN depende del tipo de aplicación en que se implemente. Por ejemplo, los correctores ortográficos requieren muy poca información léxica, mientras que un sistema de comprensión textual necesita información morfológica, sintáctica, semántica y pragmática de las unidades léxicas, además de conocimiento no lingüístico de diversa índole (Nirenburg y Raskin 2004). Por tanto, la estrategia más razonable en la utilización de una base de conocimiento multifuncional en una tarea del PLN es ajustar la base de conocimiento a las necesidades de la aplicación que diseñemos, pero permitiendo al sistema tener acceso a información adicional

en cualquier momento. En los subapartados 5.2 y 5.3 describimos el conocimiento específico que podemos almacenar en el lexicón y la ontología respectivamente.

5.2 El lexicón

5.2.1 La entrada léxica

En esta última década, la lingüística computacional se ha ido impregnando de un marcado lexicalismo. En la lingüística teórica de los años 50 y 60, las regularidades del lenguaje eran explicadas por el componente sintáctico, dejando al lexicón como un depósito de idiosincrasias (Atkins, Levin *et al.* 1994). A partir de los años 80, el panorama lingüístico se caracterizó por un profundo interés en las relaciones que se establecen entre las propiedades semánticas y sintácticas de las unidades léxicas, como observamos en la Gramática Léxico-funcional (Kaplan y Bresnan 1982), la Gramática de Estructura Sintagmática Generalizada (Gazdar, Klein *et al.* 1985) o la Gramática de Estructura Sintagmática Nuclear (Pollard y Sag 1994) entre otros muchos modelos lingüísticos. De hecho, la mayoría de teorías lingüísticas contemporáneas siguen siendo lexicalistas, en el sentido de que las unidades léxicas no son consideradas como átomos de la gramática, sino como objetos complejos generalmente representados por estructuras de rasgos (Gibbon 2000). Independientemente del sistema de notación utilizado para representar las propiedades léxicas, p.ej. la matriz de rasgos y atributos de la Gramática Léxico-funcional, las estructuras conceptuales de Jackendoff (1983, 1990) o los marcos predicativos de la Gramática Funcional de Dik (1997), muchos investigadores coinciden en que toda la información sobre el comportamiento lingüístico de las unidades léxicas debe integrarse bajo una misma interfaz en el componente léxico. Con el fin de diseñar un lexicón lo más completo posible, recomendamos seguir las directrices de EAGLES[5] (1993, 1996a, 1996b, 1999; Underwood y Navarretta 1997), las cuales establecen unos estándares descriptivos a modo de recomendaciones en la construcción de un lexicón para el PLN. Estas recomendaciones nos permiten realizar un trabajo homogéneo y consistente, facilitando así la integración de nuestro

[5] EAGLES (*The Expert Advisory Group on Language Engineering Standards*) es un proyecto financiado por la Comisión Europea con el fin de proporcionar unas directrices para la estandarización de las tecnologías lingüísticas. Más concretamente, el *Computational Lexicons Interest Group* se encargó de analizar las principales prácticas de codificación lexicográfica, comparando recursos léxicos computacionales ya existentes en lenguas europeas como el alemán, catalán, danés, español, francés, griego, holandés, inglés, irlandés, italiano, portugués y sueco.

componente léxico con otros recursos lexicográficos. A modo de ejemplo, describimos a continuación tres rasgos que comúnmente se especifican en el lexicón: la categoría gramatical, el género y el marco de subcategorización.

Un rasgo que siempre se incorpora en el lexicón es la categoría gramatical, o parte del discurso, de las unidades léxicas almacenadas en la base de conocimiento, ya que a veces esta información es suficiente para discriminar los sentidos de las palabras durante el proceso de desambiguación léxica. Por tanto, el etiquetado gramatical de las unidades léxicas que configuran el texto de entrada resulta una práctica habitual antes del análisis del aducto.

Con respecto al rasgo del género, si una lengua sólo posee género semántico (p.ej. el inglés) [6], éste suele deducirse a partir del significado asociado a las unidades léxicas, por lo que sería redundante especificar este rasgo en el lexicón. En cambio, el género formal debe ser especificado explícitamente en una entrada léxica sólo si no puede deducirse a partir de las desinencias de las unidades léxicas. Por ejemplo, en el caso de los sustantivos españoles, la distinción entre masculino y femenino no siempre se reconoce en el significante por la dicotomía flexiva *-o*/*-a* respectivamente[7]. La frecuente arbitrariedad de la asignación del género formal a los significados de los sustantivos hace que el género se considere como un accidente gramatical (Alarcos Llorach 1994) y por esta razón es necesario explicitarlo en las entradas léxicas.

Por último, los verbos suelen estar asociados a un marco de subcategorización, el cual está constituido por una lista de argumentos a los que se les asigna individualmente una serie de rasgos y un índice que sirva de identificador al mismo tiempo que especifique el orden canónico. Entre los rasgos obligatorios, debemos especificar el tipo de sintagma en que cada argumento se materializa a nivel textual (p.ej. sintagma nominal, preposicional, adverbial, etc) y la función sintáctica que desempeña cada uno de estos sintagmas. Además, debemos integrar en el marco de subcategorización las preferencias de selección de los argumentos. A pesar de que casi siempre se habla de "restricciones" de selección, éstas no deben considerarse como prohibiciones en la inserción de términos, sino más bien indican la prototipicidad de la expresión que formalice el argumento. En cuanto a su naturaleza, la lingüística teórica suele servirse de unidades léxicas para describir las preferencias de selección de una palabra. En

---

[6] De hecho, el género en inglés desempeña un papel relativamente pequeño en relación con el valor que presenta en muchas otras lenguas (Lyons 1968), no sólo porque la única presencia del género ocurre con unos pocos sustantivos que denotan el sexo del referente (p.ej. *actor-actress*, *brother-sister*), sino además porque los adjetivos ingleses son invariables en género y número, por lo que no existe el mecanismo sintáctico de la concordancia con los sustantivos.

[7] Por ejemplo, *mapa* y *árbol* son masculinos, mientras que *mano* y *nariz* son femeninos.

cambio, en el PLN el lexicón debe almacenar las preferencias de selección en forma de unidades conceptuales de la ontología, con el fin de solventar el problema de la ambigüedad semántica propia de las unidades léxicas. La incorporación de preferencias de selección en el marco de subcategorización ayuda a delimitar la estructura semántica de las unidades léxicas, realizando un papel bastante decisivo en la resolución de la ambigüedad sintáctico-semántica y la recuperación de información.

5.2.2 La construcción del lexicón

La construcción de un lexicón computacional puede llevarse a cabo a través de dos posibles métodos: la creación y la adquisición (Calzolari y Picchi 1994). Mientras el primer método implica la construcción manual del lexicón a partir de la introspección del lexicógrafo, el segundo método permite su elaboración de forma automática o interactiva a partir del conocimiento proveniente de los recursos lingüísticos en formato electrónico (p.ej. diccionarios o córpora textuales).

Velardi, Pazienza ***et al.*** (1991) advierten que la construcción manual de un lexicón computacional, donde el lexicógrafo manipula directamente las representaciones formales que deben almacenarse en el lexicón, presenta problemas tanto teóricos como prácticos. Por una parte, los investigadores que poseen diferentes posturas teóricas tenderán a observar fenómenos distintos ante una misma realidad lingüística. Por otra parte, la consistencia del lexicón se convierte en un problema acuciante cuando su tamaño excede en unos cientos de entradas. A toda esta falta de sistematicidad y consistencia, debemos añadir el gran coste humano en horas de trabajo que implica la creación a partir de cero de un lexicón computacional a gran escala.

Como alternativa a la estrategia de creación, existe la posibilidad de "reutilizar" recursos lingüísticos existentes con el fin de adquirir conocimiento léxico de forma automática o semiautomática. Actualmente existe un gran número de herramientas lexicográficas para la adquisición semiautomática de conocimiento léxico. No obstante, ya que los recursos lingüísticos legibles por la máquina son cada vez más numerosos y de fácil acceso, la adquisición automática se vuelve más atractiva.

Uno de los recursos lingüísticos más utilizados es el diccionario electrónico, donde podemos aprovechar la gran cantidad de información sobre el conocimiento lingüístico y del mundo que contienen las entradas léxicas. Los diccionarios que más se han empleado en su formato electrónico para tareas del PLN han sido ***Webster's Seventh New Collegiate***

Dictionary (Gove 1972), *Oxford Advanced Learner's Dictionary of Current English* (Hornby 1974), *Longman Dictionary of Contemporary English* (Procter 1978) y *Collins COBUILD English Language Dictionary* (Sinclair 1987). En caso de decantarnos por este método, la siguiente cuestión que debemos plantearnos es si el proceso de extracción de esa información debe ser automático, por medio de la aplicación de un algoritmo, o más bien interactivo, con la ayuda de alguna herramienta lexicográfica. Para tomar esta decisión es necesario conocer previamente en qué condiciones nos van a llegar estos recursos lingüísticos. Una de las principales limitaciones en la explotación de los diccionarios electrónicos durante el proceso de construcción de una base de conocimiento radica en la esencia misma del arte de la lexicografía. En primer lugar, los diccionarios están diseñados para ser usados por humanos. Los lexicógrafos explotan el conocimiento lingüístico de sus usuarios potenciales, de tal modo que las entradas contienen sólo la información necesaria para que el hablante de una lengua sea capaz de conectarla con su conocimiento lingüístico general. Por ello, se ignoran hechos básicos sobre los significados de las unidades léxicas, pero imprescindibles en una base de conocimiento que sirva para la comprensión de un texto de entrada. A pesar de todo, muchos investigadores creen que los diccionarios contienen suficiente conocimiento para un sistema del PLN, ya que este conocimiento se presenta de forma implícita pero fácilmente accesible a través de la información de otras entradas léxicas (Dolan, Vanderwende *et al.* 1993; Wilks, Slator *et al.* 1996). En segundo lugar, los lexicógrafos trabajan bajo grandes presiones de tiempo y espacio, lo cual propicia la inconsistencia en las entradas. Por ejemplo, las unidades léxicas que tienen un comportamiento similar morfológica, sintáctica y/o semánticamente no reciben a veces un tratamiento homogéneo en los diccionarios. Como consecuencia, el trabajo necesario para determinar las posibles variantes metatextuales podría ser superior al trabajo implicado en la construcción manual de la misma base de conocimiento (Ide y Véronis 1994). En tercer lugar, no hay un tratamiento adecuado de la polisemia, ya que los diversos significados de una unidad léxica son tratados de forma aislada. Finalmente, los diccionarios electrónicos suelen contener errores, los cuales se cometen en el proceso de introducción al ordenador de la información contenida en el diccionario en papel. La corrección de estos errores es igualmente muy costosa tanto en tiempo como recursos humanos. Por ejemplo, se tardó casi un año en comprobar y corregir la versión electrónica del *Oxford Advanced Learner's Dictionary of Current English* (Moreno Ortiz 1998). A pesar de todas estas limitaciones, no todas las investigaciones realizadas con diccionarios electrónicos han sido infructuosas. No obstante, la cantidad de información semántica útil para el PLN que ha sido extraída automáticamente a partir de las definiciones lexicográficas es muy reducida

(Ide y Véronis 1994). Con todos estos argumentos, llegamos a la conclusión de que el tiempo empleado en la adquisición automática y revisión de un lexicón puede ser muy similar al tiempo que se emplearía en la construcción del mismo lexicón a través de un método interactivo, razón por la cual recomendamos este último método.

Otro recurso lingüístico que se utiliza en el proceso de adquisición de conocimiento lingüístico es el corpus textual. Los córpora permiten disponer de forma cómoda y completa del uso de la lengua de la manera más objetiva posible, apoyando así nuestros análisis en la observación ordenada y sistemática de la realidad, lejos de especulaciones hipotético-deductivas (Candalija Reina 1998). La mayoría de los experimentos llevados a cabo para la adquisición de información léxica a través de córpora se hallan aún en fase experimental (Moreno Ortiz 1998). Los datos obtenidos de los córpora incluyen principalmente información sobre la frecuencia de uso de las unidades léxicas y sus colocaciones. No obstante, los córpora presentan el inconveniente de que la mayoría sólo contiene textos a un nivel léxico superficial, unos pocos proporcionan anotaciones sintácticas, pero casi ninguno presenta información sobre los significados léxicos. Por tanto, con el propósito de convertir los córpora en una herramienta útil para la adquisición léxica, es necesario aplicar un profundo preprocesamiento lingüístico que implique el etiquetado concerniente a la información semántico-conceptual de las unidades léxicas además de la identificación de las estructuras sintácticas en las que intervienen.

Actualmente, se recomienda la construcción de bases de conocimiento a través de herramientas lexicográficas que permitan al ingeniero lingüista combinar múltiples recursos lingüísticos, ya que interconectando diccionarios electrónicos y córpora informatizados podemos obtener una gran riqueza de información léxica que nos permitirá construir un componente léxico más robusto (Guthrie, Pustejovsky *et al.* 1996).

5.3 La ontología

La ontología es uno de los componentes centrales en un base de conocimiento para el PLN. En el campo de la informática, la ontología se define como un catálogo del tipo de cosas que existen en un dominio de interés desde la perspectiva de una persona que utiliza una lengua con el propósito de hablar sobre ese dominio (Sowa 2000). Por tanto, las ontologías tienen como objetivo presentar el conocimiento compartido por una comunidad acerca de un dominio. Con este fin, diseñar una ontología implica determinar el conjunto de categorías semánticas que refleje adecuadamente la organización conceptual del dominio sobre el que el

sistema debe trabajar, optimizando la cantidad y calidad de la información almacenada (Lenci 2000). Las ontologías difieren básicamente en tres aspectos (Vossen, Bloksma *et al.* 1998; Lenci 2000): el alcance (i.e. general o terminológica), la granularidad de las unidades (i.e. simples etiquetas o unidades provistas de estructura interna) y el uso que se vaya a hacer de la ontología (i.e. multipropósito o específica). Las principales razones para utilizar una ontología en un sistema del PLN son las siguientes (Bateman 1991; Nirenburg, Beale *et al.* 1996):

- almacenar el conocimiento del mundo y permitir que los lexicones de diferentes lenguas compartan ese mismo conocimiento
- realizar inferencias sobre el conocimiento del mundo a partir de los significados de las unidades léxicas
- proporcionar una base para la construcción de una interlingua, la cual se utiliza para la representación del significado de un texto de entrada o salida

Los proyectos del PLN que han desarrollado algunas de las ontologías más extensas son CyC (Lenat 1995), EDR (Yokoi 1995) y EuroWordNet (Vossen 1998), pero todavía queda mucho trabajo por hacer en el ámbito ontológico. Una ontología para el PLN puede ser más o menos robusta dependiendo del tratamiento que reciban las dimensiones de cobertura, acceso y semántica (Barker, Porter *et al.* 2001). En otras palabras, una ontología robusta debe tener una gran variedad de componentes (cobertura) que puedan ser encontrados fácilmente (acceso) y sean suficientemente generales como para ser utilizados en diversos contextos al mismo tiempo que sean suficientemente específicos como para expresar el conocimiento relevante (semántica). A este respecto, el objetivo es crear una base de conocimiento cuyo modelo ontológico reciba una buena evaluación en cobertura, acceso y semántica, a diferencia de lo que ocurre en la práctica ontológica actual, donde generalmente se hace más énfasis en la cobertura y el acceso en detrimento de la semántica. En este sentido, los conceptos ontológicos no deben ser símbolos atómicos sino poseer una rica representación conceptual que refleje la estructura del sistema cognitivo del ser humano. Por tanto, una de las tareas en el campo de la ontología que mayor esfuerzo requerirá por parte de los lingüistas computacionales será la asignación de postulados de significado a las unidades conceptuales.

Velardi, Pazienza *et al.* (1991) diferencian dos posturas claras a la hora de enfrentarse a la descripción del significado en una base de conocimiento: podemos describir el contenido cognitivo de la unidad léxica por medio de rasgos o primitivos semánticos (i.e. significado conceptual) o bien indicar las asociaciones que se establecen entre las unidades léxicas del

lexicón (i.e. significado relacional). Mientras el estudio del significado conceptual pertenece a la semántica léxica profunda, el significado relacional se vincula a una semántica léxica más superficial. En sentido estricto, la semántica léxica superficial no proporciona una verdadera definición de la unidad léxica, sino más bien describe su uso en la lengua a través de las "relaciones de significado" con otras unidades léxicas. A pesar de las limitaciones inherentes al enfoque relacional, todavía se trabaja en proyectos[8] donde los lexicones contienen entradas léxicas provistas de un *genus* y una lista de "palabras clave" que entran en algún tipo de relación con el *definiendum*. Evidentemente resulta más fácil explicitar las asociaciones que se establecen entre las unidades léxicas de la lengua en forma de relaciones de significado que tratar de describir el contenido cognitivo de las palabras, pero el poder inferencial de la representación del significado relacional es drásticamente menor que el del significado conceptual, ya que a menudo es necesario aplicar mecanismos de razonamiento con el fin de obtener una representación semántica formal del texto de entrada. La semántica superficial puede ser suficiente en algunas tareas del PLN, pero si construimos una base de conocimiento robusta, tendremos la oportunidad de poder incorporarlo en una gran variedad de aplicaciones de la ingeniería lingüística, fomentando así el concepto de reutilización de los recursos.

Una vez que se adopta el enfoque de la semántica profunda, debemos aplicar una sólida metodología que nos permita describir de forma consistente el significado de una unidad conceptual por medio de una representación interlingüística del conocimiento. En este contexto, utilizamos el término interlingua como se hace habitualmente en la traducción automática, i.e. una representación universal e independiente de la lengua generada a partir del texto origen y desde la cual se genera el texto destino (Cerdà 1995). Por tanto, la interlingua debe ser capaz de describir directamente la realidad del mundo sin mediación de una lengua natural. Aunque nadie ha conseguido especificar las propiedades y estructuras conceptuales de una interlingua ideal, su realidad psicológica parece incuestionable, dado que un hablante es capaz de traducir entre lenguas tipológicamente diferentes. Actualmente existen diversos sistemas del PLN que utilizan un enfoque basado en la interlingua. Por ejemplo, en el caso de UNITRAN (Dorr 1993), un sistema de traducción automática capaz de traducir textos en inglés, español y alemán bidireccionalmente, y MILT (Dorr, Hendler *et al.* 1995), un sistema tutorial para el aprendizaje de lenguas extranjeras, la representación interlingüística se construye a partir de una versión de la "estructura conceptual léxica" propuesta por Jackendoff (1983, 1990). A diferencia del lexicón, EAGLES (1999) no presenta

[8] Por ejemplo, este es el caso de SIMPLE (Lenci, Bel *et al.* 2000).

unas claras directrices de cómo representar el significado en una ontología, sino más bien sugiere una serie de preguntas a partir de las cuales podemos tomar nuestras primeras decisiones metodológicas:

- ¿Es posible determinar un procedimiento de identificación de los componentes de un postulado de significado?
- ¿Trataremos esos componentes como "primitivos semánticos" y, en caso afirmativo, como "universales"?
- ¿Tendremos un conjunto "finito" y "completo" de componentes de significado?

Todas estas preguntas plantean líneas de investigación en las que actualmente trabajan los ingenieros del conocimiento.

5.4 Las macroestructuras cognitivas

Una tarea del PLN que requiera la comprensión del lenguaje natural necesita estructuras cognitivas de alto nivel que contengan conocimiento prototípico que facilite las inferencias y predicciones al igual que la selección y el control de la información. Los postulados de significado de las unidades léxicas no son suficientes para describir el conocimiento del sentido común de las personas, pero sí contribuyen activamente a construir las "macroestructuras cognitivas". En otras palabras, una base de conocimiento debe integrar la información léxico-conceptual con el conocimiento episódico, correlación que casi ningún sistema del PLN ha conseguido todavía. Definimos a estos esquemas como “macroestructuras” porque son construcciones más amplias que los postulados de significado. Mientras el postulado de significado es una representación del conocimiento orientada a la unidad conceptual, la macroestructura cognitiva organiza el conocimiento en escenas y episodios bajo los parámetros de la temporalidad y la causalidad. Por otra parte, calificamos a estas estructuras como “cognitivas” porque se construyen con entidades del modelo ontológico. La principal ventaja de este enfoque cognitivo se encuentra en el hecho de que las expectativas sobre lo que va a ocurrir en una determinada situación no son léxicas sino conceptuales, por lo que diversas expresiones léxicas sinonímicas pueden ajustarse perfectamente a la realización lingüística de una misma expectativa.

El proceso de comprensión de un texto implica reconocer las unidades léxicas del texto, decidir cómo se estructuran sintácticamente dentro de cada oración, determinar el significado explícito de cada oración y, finalmente, realizar inferencias intraoracional e interoracionalmente con el fin de determinar el significado implícito en cada oración (Norvig

1987). Ahora bien, la comprensión de un texto no se limita a la comprensión del conjunto de oraciones individuales que lo forman; implica necesariamente la integración de toda la información proveniente de estas oraciones en un "modelo de situación" (Zwaan y Radvansky 1998) con el fin de reconstruir el mundo textual que subyace al sentido literal de las realizaciones lingüísticas que aparecen en la superficie textual. Un texto es un conjunto de oraciones que están conectadas de alguna forma significativa a un tema o conjunto de temas, los cuales se describen a través de una serie de escenas y episodios ordenados temporal y causalmente (Moorman y Ram 1994). Según Zwaan y Radvansky (1998), si queremos comprender adecuadamente la situación descrita en un texto, es necesario conocer:

- las entidades (i.e. protagonistas y objetos) que forman el modelo del mundo
- la sucesión temporal de eventos
- la organización espacial donde se sitúan las entidades
- las relaciones causales entre los eventos
- los objetivos que tienen los protagonistas y los planes de acción que sirven para alcanzar dichos objetivos

Desde la invención de los guiones por Schank y Abelson[9] (1977), se han realizado muy pocos esfuerzos para construir una extensa base de datos de macroestructuras cognitivas. Por ejemplo, los paquetes de expectativas (Gordon 1999) y *ThoughtTreasure* (Mueller 1999) contienen una serie de hechos y reglas que representan el conocimiento estereotípico más relevante sobre situaciones cotidianas, pero este conocimiento no se presenta de forma episódica, por lo cual no pueden aplicarse mecanismos de razonamiento analógico[10].

6. CONCLUSIONES

En este capítulo hemos descrito las líneas generales que deben motivar el diseño y la elaboración de una base de conocimiento léxico-conceptual de propósito general para su implementación en un sistema del PLN. En realidad, recomendamos la estrategia de crear una extensa base de conocimiento nuclear junto con una serie de bases de conocimiento satélites

[9] Estos autores introdujeron la noción de “guión” dentro del marco de la psicología y la inteligencia artificial, empezando a implementarlo en aplicaciones informáticas para la comprensión de historias y sirviendo luego como punto de arranque para el desarrollo de un modelo de organización de la memoria.

[10] La gran ventaja de este tipo de razonamiento frente al razonamiento lógico, como el utilizado por CyC (Lenat, Guha *et al.* 1990), radica en la innecesidad de una axiomatización precisa y completa antes de que se active el proceso de razonamiento (Singh y Barry 2003).

vinculadas a dominios especializados, cada una de las cuales estará formada por su propio lexicón y ontología. De esta forma, reducimos el número de sentidos que el sistema debe considerar para cada unidad léxica tanto en los procesos de análisis como generación. Cada una de las bases de conocimiento especializado estará asociada a un contexto particular, por lo que el sistema sólo tendrá en cuenta la información almacenada en su lexicón y ontología cuando se active el contexto en cuestión. Ante esta gran cantidad de información que debemos gestionar, resulta imprescindible diseñar una serie de herramientas semiautomáticas que nos ayuden en la edición, revisión y consulta de la base de conocimiento.

7. BIBLIOGRAFÍA

Adarraga, P., 1994, "Sistemas basados en conocimiento: conceptos básicos". *Psicología e Inteligencia Artificial.* Eds. P. Adarraga y J.L. Zaccagnini. Trotta, Madrid. 141-186.

Aguado de Cea, G., I. Álvarez de Mon *et al.*, 2002, "Primeras aproximaciones a la anotación lingüístico-ontológica de documentos de la Web Semántica: OntoTag". *Inteligencia Artificial* 17: 37-49.

Alarcos Llorach, E., 1994, *Gramática de la Lengua Española*. Espasa Calpe, Madrid.

Atkins, B.T.S., B. Levin *et al.*, 1994, "Computational approaches to the lexicon: an overview". *Computational Approaches to the Lexicon.* Eds. B.T.S. Atkins y A. Zampolli. Oxford University Press, Oxford. 17-45.

Barker, K., B. Porter *et al.*, 2001, "A library of generic concepts for composing knowledge bases". *1st International Conference on Knowledge Capture, Victoria BC, 2001*. 14-21.

Bateman, J.A., 1991, "The theoretical status of ontologies in natural language processing". *Text Representation and Domain Modelling - Ideas from Linguistics and AI.* Eds. S. Preuss y B. Schmitz. Technische Universitaet Berlin, Berlín. 50-99.

Berners-Lee, T., 1998, "Semantic Web road map".
[http://www.w3.org/DesignIssues/Semantic.html]

Browne, M., 1997, “Information policy for an information society”. *Cultural Crossroads: Ownership, Access and Identity, Sydney, 1997*.
[http://sirio.deusto.es/abaitua/konzeptu/nlp/Browne_M.html]

Calzolari, N. y E. Picchi, 1994, “A lexical workstation: from textual data to structured database”. *Computational Approaches to the Lexicon*. Eds. B.T.S. Atkins y A. Zampolli. Oxford University Press, Oxford. 439-467.

Candalija Reina, J.A., 1998, “Sobre la cientificidad de la gramática: el uso de corpora informatizados como método de análisis lingüístico”. *Estudios de Lingüística Cognitiva*. Ed. J.L. Cifuentes Honrubia. Universidad de Alicante, Alicante. 295-307.

Cassen, B., J.F. Dégremont *et al.*, 1991, "Formación del personal investigador y estudios de doctorado en lingüística computacional". *Las Industrias de la Lengua.* Ed. J. Vidal Beneyto. Pirámide-Fundación Germán Sánchez Ruipérez, Madrid. 411-415.

Cerdà, R., 1995, "Perspectivas en traducción automática". *LynX Working Papers* 2: 1-22.

Comisión Europea, 1997, "Language Engineering: Harnessing the Power of Language". HLT Central.
[http://www.hltcentral.org/usr_docs/Harness/harness-en.htm]

Dik, S.C., 1997, *The Theory of Functional Grammar*. Mouton de Gruyter, Berlín/Nueva York.

Dolan, W., L. Vanderwende *et al.*, 1993, “Automatically deriving structured knowledge bases from on-line dictionaries”. *1st Conference of the Pacific Association for Computational Linguistics, Vancouver, 1993*. 5-14.

Dorr, B.J., 1993, *Machine Translation: A View from the Lexicon*. MIT Press, Cambridge (Mass.).

Dorr, B.J., J. Hendler *et al.*, 1995, "Use of LCS and discourse for intelligent tutoring: On beyond syntax". *Intelligent Language Tutors: Balancing Theory and Technology*. Eds. M. Holland, J. Kaplan *et al.* Lawrence Erlbaum Associates, Hillsdale. 289-309.

EAGLES, 1993, "EAGLES: Computational lexicons methodology task". Informe técnico EAG-CLWG-METHOD/B.
[http://www.ilc.cnr.it/EAGLES96/method/method.html]

EAGLES, 1996a, "EAGLES: Synopsis and comparison of morphosyntactic phenomena encoded in lexicons and corpora. A common proposal and applications to European languages". Informe técnico EAG-CLWG-MORPHSYN/R.
[http://www.ilc.cnr.it/EAGLES96/morphsyn/morphsyn.html]

EAGLES, 1996b, "EAGLES: Preliminary recommendations on subcategorisation". Informe técnico EAG-CLWG-SYNLEX/P.
[http://www.ilc.cnr.it/EAGLES96/synlex/synlex.html]

EAGLES, 1999, "EAGLES: Preliminary recommendations on lexical semantic encoding". Informe técnico LE3-4244.
[http://www.ilc.cnr.it/EAGLES96/EAGLESLE.PDF]

Ferrari, G., 2004, "State of the art in Computational Linguistics". *Linguistics Today - Facing a Greater Challenge.* Ed. P. Sterkenburg. John Benjamins, Amsterdam/Filadelfia. 163-186.

Floridi, L., 1999, *Philosophy and Computing: An Introduction*. Routledge, Londres/Nueva York.

Gazdar, G., E. Klein *et al.*, 1985, *Generalized Phrase Structure Grammar*. Harvard University Press, Cambridge (Mass.).

Gibbon, D., 2000, "Computational lexicography". *Lexicon Development for Speech and Language Processing*. Eds. F. Eynde y D. Gibbon. Kluwer, Dordrecht. 1-42.

Gonzalo Arroyo, J. y M.F. Verdejo Maíllo, 2003, “Recuperación y extracción de información”. *Tecnologías del Lenguaje*. Ed. M.A. Martí Antonín. Universitat Oberta de Catalunya, Barcelona. 157-192.

Gordon, A.S., 1999, *The Design of Knowledge-rich Browsing Interfaces for Retrieval in Digital Libraries*. Tesis doctoral. Northwestern University.

Gove, P., ed., 1972, *Webster's Seventh New Collegiate Dictionary*. G. & C. Merriam, Nueva York.

Grishman, R., 1986, *Computational Linguistics: An Introduction*, Cambridge University Press, Cambridge.

Guthrie, L., J. Pustejovsky *et al.*, 1996, “The role of lexicons in natural language processing”. *Communications of the ACM* 39, 1: 63-72.

Harman, D., P. Schäuble *et al.*, 1995, “Document retrieval”. *Survey of the State of the Art in Human Language Technology*. Eds. G. Varile y A. Zampolli. Cambridge University Press, Cambridge. 259-262.

Hornby, A.S., 1974, *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford.

Ide, N. y J. Véronis, 1994, “Knowledge extraction from machine-readable dictionaries: an evaluation”. *Machine Translation and the Lexicon*. Ed. P. Steffens. Springer Verlag, Berlín. 19-34.

Jackendoff, R.S., 1983, *Semantics and Cognition*. MIT Press, Cambridge (Mass.).

Jackendoff, R.S., 1990, *Semantic Structures*. MIT Press, Cambridge (Mass.).

Kaplan, R.M. y J. Bresnan, 1982, "Lexical-Functional Grammar: A formal system for grammatical representation". *The Mental Representation of Grammatical Relations*. Ed. J. Bresnan. MIT Press, Cambridge (Mass.). 173-281.

Lenat, D.B., 1995, "CyC: a large-scale investment in knowledge infrastructure". *Communications of the ACM* 38, 11: 33-38.

Lenat, D.B., R.V. Guha *et al.*, 1990, "CyC: toward programs with common sense". En *Communications of the ACM* 33, 8: 30-49.

Lenci, A., 2000, "Building an ontology for the lexicon: semantic types and word meaning". *Workshop on Ontology-Based Interpretation of Noun Phrases, Kolding, 2000.*

Lenci, A., N. Bel *et al.*, 2000, "SIMPLE: A general framework for the development of multilingual lexicons". *International Journal of Lexicography* 13, 4: 249-263.

Lyons, J., 1968, *Introduction to Theoretical Linguistics.* Cambridge University Press, Cambridge.

Martí Antonín, M.A., ed., 2003, *Tecnologías del Lenguaje*. Universitat Oberta de Catalunya, Barcelona.

Martí Antonín, M.A. y J. Llisterri, 2001, "La ingeniería lingüística en la sociedad de la información". *Digithum, Revista digital d'humanitats* 3.
[http://www.uoc.edu/humfil/articles/esp/llisterri-marti/llisterri-marti_imp.html]

Mitkov, R., ed., 2003, *The Oxford Handbook of Computational Linguistics.* Oxford University Press, Oxford.

Moorman, K. y A. Ram, 1994, "A functional theory of creative reading". Informe técnico GIT-CC-94/01. Georgia Institute of Technology.
[http://citeseer.ist.psu.edu/moorman-functional.html]

Moreno Ortiz, A., 1998, "El lexicón en la lexicografía computacional: adquisición y representación de información léxica". *Alfinge* 10: 249-272.

Moure, T. y J. Llisterri, 1996, “Lenguaje y nuevas tecnologías: el campo de la lingüística computacional”. *Avances en Lingüística Aplicada.* Ed. M. Fernández Pérez. Universidad de Santiago de Compostela, Santiago de Compostela. 147-227.

Mueller, E.T., 1999, "A database and lexicon of scripts for ThoughtTreasure". [http://cogprints.ecs.soton.ac.uk/archive/00000555/]

Nirenburg, S., S. Beale *et al.*, 1996, “Lexicons in the MikroKosmos Project”. *AISB’96 Workshop on Multilinguality in the Lexicon, Brighton, 1996.*

Nirenburg, S. y V. Raskin, 2004, *Ontological Semantics*. MIT Press, Cambridge (Mass.).

Norvig, P., 1987, *A Unified Theory of Inference for Text Understanding*. Tesis doctoral. University of California, Berkeley.

Pollard, C e I. Sag, 1994, *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago.

Procter, P., ed., 1978, *Longman Dictionary of Contemporary English*. Longman, Harlow (Essex).

Reuters, 1996, *Dying for Information: An Investigation Into the Effects of Information Overload in the USA and Worldwide*. Reuters, Londres.

Schank, R. y R.P. Abelson, 1977, *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum, Hillsdale.

Sinclair, John, ed., 1987, *Collins COBUILD English Language Dictionary*. Collins, Londres.

Singh, P. y B. Barry, 2003, "Collecting commonsense experiences". *2nd International Conference on Knowledge Capture, Florida, 2003*.

Sowa, J.F., 2000, “Ontology, metadata, and semiotics”. *Conceptual Structures : Logical, Linguistics, and Computational Issues.* Eds. B. Ganter y G.W. Mineau. Springer Verlag, Berlín. 55-81.

Tzoukermann, E., J.L. Klavans *et al.*, 2003, "Information retrieval". *The Oxford Handbook of Computational Linguistics.* Ed. R. Mitkov. Oxford University Press, Oxford. 529-544.

Underwood, N. y C. Navarretta, 1997, “Towards a standard for the creation of lexica”. Informe técnico. Center for Sprogteknologi. Copenague.

Varile, G. y A. Zampolli, eds., 1995, *Survey of the State of the Art in Human Language Technology*. Cambridge University Press, Cambridge.

Velardi, P., M.T. Pazienza *et al.*, 1991, “How to encode semantic knowledge: a method for meaning representation and computer-aided acquisition”. *Computational Linguistics* 17, 2: 153-170.

Vossen, P., 1998, “Introduction to EuroWordNet”. *Computers and the Humanities* 32, 2-3: 73-89.

Vossen, P., L. Bloksma *et al.*, 1998, *The EuroWordNet Base Concepts and Top Ontology*. Informe técnico. University of Amsterdam.
[http://www.illc.uva.nl/EuroWordNet/docs.html]

Wilks, Y.A., B.M. Slator *et al.*, eds., 1996, *Electric Words. Dictionaries, Computers and Meanings*. MIT Press, Cambridge (Mass.).

Yokoi, T., 1995, “The EDR Electronic Dictionary”. *Communications of the ACM* 38, 11: 42-44.

Zwaan, R.A. y G.A. Radvansky, 1998, "Situation models in language comprehension and memory". *Psychological Bulletin* 123, 2: 162-185.