VOLUMEN MONOGRÁFICO (1999), 11-24

PANORAMA DE LA LINGÜÍSTICA COMPUTACIONAL EN EUROPA

Mª ANTONIA MARTÍ ANTONÍN
Universidad de Barcelona

RESUMEN. *El panorama de la Lingüística Computacional (LC) en Europa en la última década se ha orientado claramente hacia el desarrollo de recursos básicos y aplicaciones, y ha dejado en un segundo término la investigación básica. El problema de si era posible o no el procesamiento del lenguaje natural, que tanto preocupó en los años 70 y 80, parece haber quedado resuelto, y lo que ahora interesa es ver en qué medida y de qué modo la LC puede contribuir a resolver los problemas que tiene planteados la Sociedad de la Información.*

PALABRAS CLAVE. *Lingüística computacional, recursos lingüísticos, programas marco de la Unión Europea, sociedad de la información, ingeniería lingüística.*

ABSTRACT. *Current trends of Computational Linguistics in Europe during this last decade are clearly oriented to develop basic resources and applications, leaving basic research in a second term. It seems that the problem of demonstrating the feasibility of natural language processing, which has occupied researchers during the 70s and 80s, has been solved positively, and now they are faced to find in which way CL can contribute to solve the challenges of the Information Society.*

KEYWORDS. *Computational linguistics, language resources, European Union framework programmes, information society, language engineering.*

1. INTRODUCCIÓN

Antes de adentrarnos en el tema que nos ocupa, puede ser de interés referirnos, ni que sea brevemente, a algunas de las características de los recursos lingüísticos computacionales. El desarrollo de recursos lingüísticos constituye un objetivo que debe plantearse a medio y largo plazo: la confección de un léxico computacional o de una gramática de amplia cobertura representa una inversión importante de recursos humanos y de tiempo. Algunos sistemas de procesamiento del lenguaje, como el analizador de Naomi Sager[1] (Sager *et al.* 1987), o de traducción automática, como

Systran y Metal, tardaron más de diez años en ser operativos. En estas circunstancias, se da la paradoja de que cuando un sistema está disponible para ser usado, su tecnología ha envejecido notablemente respecto de lo que es operativo en aquel momento y en muchos casos debe reprogramarse. En estas condiciones, es razonable que existan reticencias por parte de la inversión privada sobre este tipo de recursos.

Por otro lado, los proyectos subvencionados por la Unión Europea (UE) raramente dan como resultado productos "acabados". Los grupos de investigación universitarios que trabajan en el área de la LC, si pretenden elaborar recursos lingüísticos que sean útiles tanto para su propia investigación como para el desarrollo de aplicaciones —es decir, si pretenden situarse en el ámbito de la I+D—, deben plantearse líneas de trabajo a medio y largo plazo e intentar que todos los recursos que puedan obtener confluyan en unos objetivos previamente determinados. De nada sirve un analizador morfológico que sólo contenga la interpretación de "entre" como preposición o de una gramática del castellano que sólo trate oraciones con una estructura rígida de sujeto-verbo-objeto, aunque el léxico contenga 60.000 entradas y la gramática 200 reglas: un analizador de estas características, probablemente resultado del esfuerzo de un grupo durante unos años, no sirve para nada si no se completa.

Es por ello que el desarrollo de recursos lingüísticos es una tarea que debe abordar la sociedad en su conjunto y requiere el esfuerzo coordinado tanto de la iniciativa privada como de las instituciones y de los centros universitarios: la universidad, como motor en el terreno de la investigación y formación de estos nuevos profesionales; la iniciativa privada, en el desarrollo de aplicaciones que incorporen los nuevos logros; y las instituciones, mediante inversiones que permitan poner a disposición de unos y otros los recursos lingüísticos informatizados básicos. En este sentido es fundamental la colaboración de las universidades con las empresas: el desarrollo de recursos lingüisticos debe ser un objetivo que se emprenda en estrecha colaboración, tratando de conseguir la máxima sinergia entre los equipos humanos y los logros obtenidos por ambos hasta el momento.

En el IV Programa Marco de la UE (y también en lo que se desprende de las directrices que han aparecido ya sobre el V Programa Marco) se hace especial hincapié en este tipo de colaboración, habitual en otras áreas científicas, pero relativamente nueva en lo que se refiere al lenguaje. Aunque todavía no se ha publicado oficialmente, las líneas que marcarán el plan de acción del V Programa Marco están claramente orientadas al desarrollo de proyectos de investigación de carácter aplicado y orientados a satisfacer las necesidades sociales que se consideran prioritarias. Destacan cuatro programas temáticos: el primero hace referencia a la mejora de la calidad de vida, el segundo está orientado a hacer más amigable la Sociedad de la Información, el tercero favorece el crecimiento competitivo mediante la mejora de las comunicaciones y el desarrollo de productos innovadores y,

finalmente, el cuarto concierne a la protección del medio ambiente y al desarrollo de fuentes de energía.

¿Qué lugar ocupa la Lingüística Computacional en estos proyectos europeos para los próximos cuatro años? Veamos brevemente cuál ha sido el desarrollo de esta disciplina y qué problemas tiene planteados en el presente y de cara al futuro.

2. EL PASADO

Las primeras investigaciones en el procesamiento del lenguaje, que podríamos situar en los años 50, no estaban fundamentadas en ningún modelo lingüístico teórico ni existía una conciencia clara de la dificultad de su tratamiento computacional: se partía de la base de que el lenguaje humano era como los lenguajes artificiales, aunque algo más complejo. Los primeros intentos de desarrollar sistemas de traducción automática pusieron de manifiesto tanto la complejidad del tratamiento automático de las lenguas naturales, como lo limitado de los recursos informáticos de los que en aquel momento se disponían para procesarlas.

A lo largo de los años 70 la investigación en Lingüística Computacional se orientó fundamentalmente a demostrar que el procesamiento del lenguaje era posible: desde la Inteligencia Artificial se diseñan modelos del conocimiento como los *frames* y las redes semánticas; la teoría de computación propone nuevos lenguajes de programación, como Lisp y Prolog, que tratan objetos complejos; se empiezan a desarrollar nuevas técnicas de análisis, y crece el interés por el diseño de gramáticas computacionales basadas en modelos lingüísticos (Sager *et al.* 1987). las aplicaciones desarrolladas durante estos años son los llamados sistemas de "juguete" o *toy systems*, sin interés aplicado y cuyo objetivo es demostrar que un determinado problema de análisis del lenguaje o de ingeniería del conocimiento podía ser resuelto. Estos sistemas se caracterizaban por una serie de limitaciones que en muchos casos se habían introducido deliberadamente. Tanto su tratamiento de dominios lingüísticos y conceptuales restringidos, como su procesamiento integrado de los programas y los datos lingüísticos, los hacían difícilmente transportables a otros dominios y a otras lenguas, y poco resistentes ante el análisis de textos incorrectos o incompletos.

La década de los 80 se caracteriza por la penetración de las teorías lingüísticas en el ámbito de la computación, surgen los primeros formalismos gramaticales con una clara voluntad de poder ser tratados computacionalmente (Gazdar *et al.* 1985; Pollard y Sag 1987) y se empiezan a desarrollar recursos lingüísticos para el procesamiento del lenguaje natural a gran escala. En esos años, el objetivo que se persigue es el desarrollo de aplicaciones como la traducción automática o las interfaces en lenguaje natural, sin las restricciones que caracterizaron a los *sistemas de juguete* de la etapa anterior. Para ello, estas aplicaciones tienen que cumplir dos nuevas condiciones: ser

independientes del dominio, es decir, transportables a otro dominio lingüístico y conceptual; y ser capaces de procesar cualquier tipo de texto.

Un ejemplo paradigmático de este nuevo enfoque lo encontramos en el proyecto Alvey (Oakley y Owen 1990), financiado por el gobierno inglés, cuyo objetivo fue el desarrollo de una plataforma de recursos lingüísticos para la lengua inglesa que sirvieran de base a la investigación posterior. Se trataba de evitar las repeticiones y solapamientos de recursos humanos y económicos tan frecuentes en esta área. En el proyecto participaron los centros de investigación universitarios más relevantes y empresas privadas.

Otro problema que se plantea en estos años es la resolución del desequilibro existente entre los recursos disponibles para el inglés y el resto de lenguas. Desde la Unión Europea se lanzan programas para el desarrollo de recursos lingüísticos en los que se prioriza la participación de los países más deficitarios en este tipo de infraestructuras. El programa Esprit y el proyecto de traducción automática Eurotra incidieron de manera muy positiva en la constitución de equipos de investigación y en el desarrollo de recursos lingüísticos para las lenguas menos favorecidas.

Hacia el final de la década de los años 80 se disponía ya de herramientas de análisis y de formalismos adecuados para el procesamiento y representación del lenguaje. Faltaba obtener los datos lingüísticos a gran escala para que, una vez formalizados e incorporados en las gramáticas y léxicos computacionales, pudieran ser utilizados en las diferentes aplicaciones. El inglés se encontraba en clara ventaja respecto de las otras lenguas.

3. LA SITUACIÓN ACTUAL: LA SOCIEDAD DE LA INFORMACIÓN

Veamos cuáles son las características que configuran la situación actual de la Lingüística Computacional. En 1992 la Dirección General XIII de la UE publica el informe Danzin, donde se recogen las directrices que deberá seguir la LC en los años siguientes. Se trata de un informe estratégico donde se señalan las líneas de trabajo, objetivos, tipo de resultados, prioridades, condiciones de organización y condiciones financieras para el desarrollo de las Industrias de la Lengua en Europa. En líneas generales el informe antepone el desarrollo de recursos básicos a la investigación, y aboga por la reutilización de los recursos ya existentes. Durante los años 80 se había puesto especial interés en favorecer el desarrollo de la investigación básica en lingüística computacional para las diferentes lenguas de la UE y era necesario rentabilizar estas inversiones en la línea de crear recursos lingüísticos a partir de lo que se había desarrollado en los años anteriores.

Cabe, pues, señalar las siguientes cinco características como distintivas de esta nueva etapa:

(1) Abandono de las especulaciones lingüísticas, en favor del desarrollo de recursos a gran escala —en especial, léxicos y gramáticas— con información básica, independiente de cualquier aplicación, y cuyo contenido responda a unos criterios estándar comunes para todas las lenguas.
(2) Desarrollo de metodologías basadas en el análisis de corpus lingüísticos, que pasan a ocupar un lugar privilegiado por diversas razones: (a) constituyen una fuente de información valiosa para la creación de diccionarios, léxicos computacionales y gramáticas; (b) debidamente etiquetados, se utilizan como recurso en los sistemas de anotación automática; (c) se utilizan como banco de pruebas para los sistemas de procesamiento del lenguaje; y (d) el estudio del lenguaje puede enriquecerse gracias al análisis de datos textuales. Como resultado, aparece una nueva disciplina, la Lingüística de Corpus, orientada al procesamiento y explotación de este nuevo tipo de recurso lingüístico.
(3) Aplicación de criterios de estandarización de los datos lingüísticos que faciliten su reutilización e integración. En esta línea se promueven diversas acciones como la TEI (Text Encoding Initiative), en colaboración con los Estados Unidos; y el programa EAGLES (European Advisory Group for Language Engineering Standards), que define códigos estándar para el etiquetado de corpus textuales y de voz, léxicos y formalismos para el análisis del lenguaje.
(4) Desarrollo de aplicaciones con un claro interés práctico.
(5) Se promueven proyectos para la integración de los recursos desarrollados en los años anteriores y que den como resultado los recursos básicos que se requieren. Algunos de estos proyectos son los siguientes:
   (a) *Interval*, orientado a la validación de recursos terminológicos multilingües. El hecho de que la mayoría de textos sobre los que se aplican los programas de traducción automática y de recuperación de información sean de áreas temáticas especializadas ha despertado un gran interés por la recopilación y estandarización de terminologías específicas. Los objetivos del proyecto Interval tratan diferentes aspectos: el estudio de las necesidades de los posibles usuarios; la cooperación de expertos especialmente para el proceso de validación; la recopilación de terminología del área de la economía y telecomunicaciones que se ha utilizado como banco de prueba para el proyecto, y la difusión de los resultados.
   (b) *Parole*, dedicado a la recopilación y codificación de corpus y léxicos para 14 lenguas: alemán, catalán, danés, finlandés, griego, español (sólo el léxico), francés, holandés, inglés, italiano y portugués. El objetivo del proyecto es la codificación de estos recursos según unos estándares comunes de modo que puedan servir de referencia a la comunidad que

trabaja en procesamiento del lenguaje. Los léxicos deben contener 20.000 entradas con información morfológica y sintáctica. La información que contienen está formalizada en un lenguaje declarativo, independiente de cualquier teoría, y deben ser multifuncionales —es decir, aptos para ser usados en cualquier tipo de aplicación— y flexibles —es decir, que permitan la incorporación de otros niveles de información—. Los corpus deben contener 20 millones de formas, y un subconjunto debe estar analizado y desambiguado morfológicamente. Los resultados de ambos proyectos deberán ser accesibles vía Internet, o a través de otros organismos de distribución de recursos como ELRA (European Language Resources Association).

(c) *Speechdat*, orientado al tratamiento de voz.

(d) *Euro WordNet* (Vossen 1998), cuyo objetivo es la construcción de redes léxico-semánticas de diversas lenguas europeas, integradas e interconectadas. El proyecto se basa en WordNet, una ontología léxico-conceptual para la lengua inglesa desarrollada en la Universidad de Princeton y de amplia utilización en diversos sistemas de tratamiento de la lengua. EuroWordNet puede servir como fuente de conocimiento para el desarrollo de aplicaciones y líneas de investigación, especialmente para la elaboración de sistemas de extracción y recuperación de información, extracción de terminología y traducción automática.

La conjunción de estos cinco factores incide directamente tanto en el rumbo que tomarán a partir de este momento las aplicaciones tradicionales de la LC (véase 3.1), como en la aparición de otras nuevas aplicaciones lingüísticas (véase 3.2). La investigación básica se desplaza hacia el desarrollo de métodos de adquisición de datos y conocimiento lingüístico que hagan posible el desarrollo de los recursos básicos necesarios en la Sociedad de la Información.

3.1. *Las aplicaciones tradicionales*

En primer lugar, las aplicaciones tradicionales del área, como son las interfaces en lenguaje natural y la traducción automática, deberán tratar dominios irrestrictos. Actualmente se exige, además, que estas aplicaciones sean competitivas, es decir, que funcionen en tiempo real, que sean multilingües, asequibles a un gran número de usuarios y que integren, en el caso de las interfaces en lenguaje natural, otros sistemas de comunicación como la voz y la imagen.

Los sistemas tradicionales de traducción automática tenían como usuarios grandes empresas e instituciones que necesitaban traducir documentación técnica, manuales o documentos administrativos. Estos sistemas requerían una infraestructura sofisticada tanto de hardware como de software. Actualmente, sólo en los Estados Unidos, existen más de 1.000 productos de software para la traducción que funcionan

sobre ordenadores personales y son asequibles a cualquier usuario. Se trata de sistemas que ofrecen una traducción aproximada que requiere una intervención importante de postedición. Los sistemas desarrollados en los años 70 y 80 se han adaptado a entornos PC, se han simplificado y, además de ser asequibles, han diversificado su oferta mediante traductores de páginas Web, servicios de traducción para el correo electrónico y la creación de estaciones de trabajo para la traducción automática. Estas estaciones de trabajo integran procesadores de textos multilingües, sistemas de extracción, gestión y reconocimiento de terminología, y memoria de traducción a partir de textos ya traducidos.

En esta línea de trabajo tenemos el proyecto *Euramis* (European Advanced Multilingual Information System) cuyo objetivo es la creación de una plataforma para la traducción automática basada en memorias de traducción. El sistema permite el tratamiento de 16 pares de lenguas con el inglés, español, francés y alemán como lenguas fuente y está integrado por cuatro módulos: (1) un módulo de memoria de traducción que contiene frases paralelas en cada uno de los pares de lenguas tratados; (2) un módulo de análisis de textos para la extracción de terminología y frases de alta frecuencia; (3) un módulo de terminología procedente de Eurodicautom, la base de datos terminológicos central de la Comisión de la UE; y (4) un módulo de traducción, que se aplica cuando se han agotado los recursos de memoria de traducción.

También en el ámbito de la traducción automática tenemos el proyecto *Multimeteo* (Coch 1996), un generador de informes metereológicos en holandés, inglés, francés, alemán y español. Con este proyecto se trata de dar respuesta a la necesidad que tienen diferentes tipos de usuarios, como los profesionales de la agricultura, turismo, transporte, comercio, construcción y de los medios de comunicación, de disponer de información meteorológica actualizada. El *input* del sistema son tablas con información referente a parámetros meteorológicos (temperatura, humedad, etc.) procedentes de diferentes oficinas meteorológicas. Un *planificador* lee el *input* numérico y genera una fórmula conceptual. A partir de esta representación, el módulo de *realización* genera el texto. El sistema de traducción está basado, desde un punto de vista teórico en la teoría "sentido-texto" de Igor Mel'cuk. Figuran como usuarios del proyecto, Meteo France (Francia), el Instituto Nacional de Meteorología (España) y el Royal Meteorological Institute (Bélgica).

3.2. *Nuevas aplicaciones*

La aparición de Internet como red de telecomunicaciones y de la WWW está cambiando el concepto de lo que se considera información y el modo de tratarla: no está mejor informado quien más datos tiene, sino quien dispone de los mejores medios para obtener sólo y exclusivamente aquellos que necesita.

La recuperación de información, una cuestión que hasta hace poco afectaba a colectivos de profesionales muy concretos, ha pasado a ser uno de los problemas clave

que deberá hacer frente la Sociedad de la Información. Internet constituye un gran banco de datos que puede devenir inservible si no se dota de sistemas de recuperación de información que satisfagan las demandas de los usuarios. Por otra parte, tanto las instituciones públicas, como los organismos y empresas de todo tipo, en la medida en que disponen de gran cantidad y variedad de información disponible, precisan también de este tipo de recursos.

El problema de la recuperación de información se centra en dos aspectos fundamentales: el filtrado de los datos y el multilingüismo. En el primer caso se trata de diseñar sistemas de recuperación de información que recuperen sólo los documentos que interesan al usuario, que eviten el exceso de "ruido" (o recuperación de textos irrelevantes) y que no ignoren ningún texto que pueda ser de interés. En lo que se refiere al segundo aspecto, se hace necesario dotar a estos sistemas de recursos que permitan el acceso multilingüe a los datos (a su vez, también multilingües).

Estas necesidades de la Sociedad de la Información están incidiendo de manera muy directa en las directrices que está tomando la Lingüística Computacional en el cambio de siglo. Podríamos señalar el tratamiento de textos lingüísticos no restringidos como el aspecto clave del cambio. Las repercusiones de este nuevo enfoque son muchas, y afectan en profundidad a las líneas de investigación tradicionales en el área de la LC para el desarrollo de recursos básicos.

La confección de resúmenes es una aplicación estrechamente relacionada con la recuperación de información. En este caso, se trata de producir una representación concisa para el lector que contenga lo esencial del documento. Inicialmente este proceso se realizaba mediante la técnica del reconocimiento de esquemas (*pattern-matching*), que identificaba los fragmentos relevantes del texto. Más tarde comenzaron a aplicarse técnicas procedentes de la IA basadas en el procesamiento del documento.

Otra línea de investigación consiste en la confección de textos de documentación técnica mediante un lenguaje controlado (para el inglés, *Simplified English* o SE), un subconjunto de una lengua natural resultado de aplicarle una serie de restricciones, por ejemplo, usar un vocabulario básico no superior a 1000 palabras, prohibir determinadas construcciones, emplear frases de no más de 20 palabras, etc. Para la confección de textos mediante lenguajes controlados se desarrollan sistemas de análisis automático que comprueban que los textos que se producen se atienen a las normas y, al mismo tiempo, ayudan a los redactores a aprender las normas de estos sublenguajes. La motivación que ha impulsado a algunas empresas a utilizar un lenguaje controlado para la confección de los manuales de usuario de sus productos, en lugar de traducirlos a diferentes lenguas, es que, siendo el inglés la lengua más usada en el ámbito tecnológico, se supone que a un hablante de otra lengua familiarizado con el dominio técnico le puede resultar bastante fácil comprender los textos en inglés producidos mediante este tipo de sublenguaje. A pesar de ello, continúa la demanda de traducción automática, y los programas de control de los

textos escritos en lenguajes controlados pueden ser los precursores de sistemas de TA totalmente automáticos.

Las nuevas directrices que han adoptado las aplicaciones tradicionales de la lingüística computacional, así como las nuevas opciones que están emergiendo, requieren disponer de recursos lingüísticos a gran escala y para el mayor número de lenguas posible. Uno de los problemas clave que plantea la elaboración de este tipo de recursos es la adquisición de información lingüística para la elaboración de las gramáticas y léxicos necesarios para su funcionamiento. ¿De dónde se puede obtener información sobre el comportamiento "real" de las lenguas naturales? ¿Cómo introducir esta información en un tiempo razonable y con el mínimo de errores posible? En el siguiente apartado presentaremos diversos sistemas de adquisición de conocimiento lingüístico.

3.3. *Nuevos recursos*

La construcción de gramáticas generales de la lengua requiere disponer de la información necesaria sobre el uso real de la misma por parte de los hablantes: los léxicos y las reglas gramaticales que las componen deberán procesar textos reales, no las frases que aparecen como ejemplos en los tratados de lingüística. Es evidente que el recurso a las intuiciones del hablante, aunque sea un experto en el tema, son del todo insuficientes, en especial si se tiene en cuenta que muchas veces estas intuiciones se contradicen con los datos. Es por ello que el desarrollo de sistemas automáticos y semiautomáticos de adquisición de conocimiento lingüístico constituye una línea de investigación clave en estos momentos si se quieren superar las limitaciones del pasado.

Los corpus textuales y los diccionarios constituyen las fuentes fundamentales de donde se extrae la información sobre las lenguas. Los primeros constituyen muestras efectivas del uso que los hablantes hacen de la lengua (véase 3.3.2); los segundos contienen datos semicodificados sobre el lenguaje que facilitan el proceso de extracción (véase 3.3.1).

3.3.1. *Los diccionarios como fuente de información (MRD)*

Un diccionario puede ser considerado como una determinada variedad de corpus lingüístico, debido a que posee unas características muy específicas que lo distinguen claramente del resto: (a) ha sido elaborado con una clara finalidad: codificar información sobre el léxico de la lengua; (b) tiene una estructura interna predeterminada: un diccionario puede considerarse como una serie de entradas organizadas alfabéticamente donde cada entrada contiene información altamente estructurada; (c) existe un cierto grado de codificación en determinados contenidos: los diccionarios contienen códigos predefinidos sobre la categoría morfológica, información temática, etc.; (d) contiene relaciones léxicas internas (sinonimia,

hiponimia, etc.) de manera implícita o explícita; (e) contiene un vocabulario restringido: aunque no todos los diccionarios disponen de un vocabulario controlado para la elaboración de las definiciones, es cierto que constituyen un subconjunto de una lengua; (f) se sigue una sistemática en la elaboración de las definiciones que permite identificar las diferentes piezas de contenido que las constituyen: el término genérico y los modificadores. Todas estas características han favorecido el desarrollo de una metodología específica para la extracción de información lingüística de los diccionarios. No sería razonable que, ante un texto de estas características, se utilizaran los mismos métodos que con corpus no restringidos.

Aunque existe un cierto escepticismo sobre la utilidad de los diccionarios como fuente de información para la construcción de léxicos computacionales, debido a que su calidad es muy irregular, cabe destacar, en su defensa, la cantidad de conocimiento que éstos han acumulado a lo largo de años de tradición lexicográfica. Por otra parte, hay que tener en cuenta que los métodos de extracción de información permiten mejorar y sistematizar los datos extraídos y combinar diversos diccionarios dando como resultado una estructura de información de calidad superior a la de las fuentes. Tanto los trabajos llevados a cabo utilizando el LDOCE (*Longman Dictionary of Contemporary English*) como fuente de información, como los resultados de los proyectos europeos *Acquilex-I* y *Acquilex-II*, confirman esta idea. Estos proyectos han definido unos procedimientos estandarizados para la extracción de información a partir de diccionarios accesibles por ordenador (*machine readable dictionaries* o MRD) que presentamos a continuación.

Aunque la información presente en un diccionario es variada en cuanto a forma y contenido, responde sin embargo a unos patrones bastante estables. De ahí que haya sido posible definir modelos computacionales para describirla, tanto a nivel de diccionario, como en lo que respecta a las entradas. El esquema básico de tratamiento de un MRD para extraer de él información léxica se establece en cuatro etapas. En primer lugar, el MRD debe sufrir un preproceso que lo transforma en una fuente manejable por el ordenador (*machine tractable dictionaries* o MTD), a partir de la cual se lleva a cabo la extracción. El soporte físico de los MRD suele limitarse a las cintas de fotocomposición de los diccionarios de uso humano, con lo que el texto aparece marcado con una serie de códigos tipográficos que complican el acceso (aunque facilitan la interpretación).

El siguiente paso consiste en construir, a partir del MTD, una Base de Datos Léxica (BDL) que permita el acceso estructurado a la información contenida en el diccionario. Realmente, de lo que se trata en esta fase es de añadir al diccionario capacidades de acceso a la base de datos, no de elaborar su contenido. El último paso es el que permite la extracción de la información léxica deseada. En ciertos casos, la extracción es directa —por ejemplo, en la extracción de la categoría sintáctica de las entradas—, mientras que, en otros casos, el proceso puede ser complejo —por

ejemplo, en la extracción de las relaciones taxonómicas— e implica un análisis lingüístico de las definiciones.

El proyecto *Acquilex* trataba de examinar la viabilidad de la utilización de MRD para la construcción del componente léxico de los sistemas de procesamiento del lenguaje natural (Castellón 1992). Su objetivo último era la construcción de bases de conocimiento léxico multilingüe. El proyecto significó la integración no sólo de los equipos de investigación participantes, sino también de las herramientas y fuentes de información léxica disponibles. Se trabajó en un contexto multilingüe con los diccionarios LDOCE para el inglés, el Garzanti para el italiano, el Van Dale para el holandés, el Vox para el español, y los bilingües Collins inglés-italiano y Van Dale inglés-holandés. El empleo de diccionarios bilingües permitió explorar su utilización como vehículo potencial de transferencia de información léxica.

El proceso seguido para los diferentes diccionarios, que fueron especificados en un formato común, es similar al descrito anteriormente. El núcleo de Acquilex residía en el proceso posterior de extracción de información semántica a partir de las bases de datos léxicas. El contenido de las BDL es exclusivamente léxico y no se incorporó ningún tipo de información que no estuviera los diccionarios de origen. Como resultado del proyecto, actualmente se dispone de una metodología estándar en relación con las técnicas automáticas y semiautomáticas de extracción de información a partir de diccionarios en soporte electrónico.

3.3.2. *Los corpus como fuente de información*

La información no normalizada, imprecisa o incompleta, como los esquemas de subcategorización, las restricciones selectivas, o el grado de plausibilidad de las diferentes acepciones de una entrada léxica en un contexto dado, debe buscarse en otras fuentes de información. En los últimos años se ha producido un auge considerable en la utilización de corpus textuales como fuente de obtención de tales informaciones.

Un corpus constituye una muestra de una lengua que se ha construido a partir de una selección de textos realizada según un determinado criterio: el análisis de la obra de un autor, el estudio de una lengua o bien la observación de unos determinados aspectos de la misma. Los corpus constituyen una buena fuente de información para el desarrollo de recursos básicos, como gramáticas y léxicos computacionales; para la investigación filológica, lingüística y lexicográfica; y como banco de pruebas para la investigación en lingüística teórica y computacional.

Los corpus no procesados consisten en el texto ASCII sin más, o en el texto completado con marcas SGML sin información lingüística. Un corpus de este tipo permite la extracción de colocaciones, concordancias y lexías. Existen programas como TACT, desarrollado en la Universidad de Toronto, que realizan este tipo de proceso. En los corpus etiquetados, cada unidad (que suele ser la palabra ortográfica o

gramatical) viene acompañada por una lista de las posibles categorías sintácticas o semánticas que se le pueden asociar. A menudo se incluyen subcategorías y otras informaciones léxicas. Es corriente someter un texto en estas condiciones a un proceso de desambiguación (normalmente estadístico) que conduzca a la asignación a cada palabra de la categoría más plausible (o a la ordenación de las etiquetas de acuerdo con su plausibilidad). Finalmente, los corpus parentizados sintácticamente están anotados de manera que se han identificado sus constituyentes. Esta parentización puede considerarse como un esqueleto de análisis sintáctico.

Quizás la aplicación más extendida del uso de corpus sea en el campo de la lexicografía, en la construcción de diccionarios y lexicones. El ejemplo más conocido es el diccionario Collins Cobuild (Sinclair 1987), construido íntegramente a partir de la consulta de los ejemplos existentes en un corpus de 6 millones de palabras creado a tal efecto por la propia editorial y la Universidad de Birmingham. Actualmente, la editorial Collins dispone de un corpus de 350 millones de formas etiquetadas morfosintácticamente para la confección de sus diccionarios. También son dignos de mención los estudios sobre frecuencias de palabras o sobre el régimen preposicional de los verbos frasales en inglés. Cabe destacar el CLS (*Cambridge Language Survey*), de Cambridge University Press, que dispone de un corpus de 200 millones de palabras etiquetadas morfosintácticamente y asociadas a la correspondiente entrada del CIDE (*Cambridge International Dictionary of English*). Por su parte, el Institut d'Estudis Catalans dispone de un corpus de 55 millones de formas etiquetadas morfosintácticamente —el *Corpus Textual Informatitzat de la Llengua Catalana*— que se utilizará para la confección de su *Diccionari del Català Contemporani*. La Real Academia Española está elaborando un corpus de 100 millones de formas —el Corpus de Referencia del Español Actual (CREA)— con unas características similares.

Se han seguido diversas aproximaciones para abordar el problema fundamental del etiquetado, tanto sintáctico como semántico, de las palabras de un corpus. Quizás la más extendida sea la modelización del flujo de palabras en términos de modelos de Markov (en el caso más general, de Modelos Ocultos de Markov) y en la estimación de los parámetros del modelo mediante técnicas estadísticas. Con este tipo de técnicas, se logran resultados por encima del 97% de éxito sin grandes dificultades.

Otro de los aspectos más importantes de la investigación en tratamiento de corpus textuales es el estudio de las coapariciones de palabras u otras unidades léxicas. La elaboración y utilización de medidas de coaparición léxica (basadas en la relación de asociación o frecuencia de la coaparición de dos palabras en un corpus) posee una gran relevancia en campos como la desambiguación de acepciones, la extracción de restricciones selectivas, la selección léxica, o para la asignación correcta de componentes en el análisis sintáctico automático.

4. EL FUTURO

Las líneas de investigación que han marcado la trayectoria de la Lingüística Computacional a lo largo de los años 90 en Europa han sido el desarrollo, estandarización y compatibilización de recursos lingüísticos, con el objetivo de dotar a las lenguas reconocidas en el ámbito de la UE de la infraestructura necesaria para participar en la Sociedad de la Información. La situación resultante de la aplicación de la política investigadora iniciada en 1992 se caracteriza por una clara mejora en la cantidad y calidad de la comunicación interlingüística y está permitiendo la aplicación de programas de procesamiento del lenguaje en los sistemas de telecomunicación más diversos —traductores automáticos en Internet, recuperación de información multilingüe, traductores de páginas Web, etc.—, así como la incorporación de módulos de procesamiento del lenguaje en las aplicaciones más diversas.

Las instituciones de carácter nacional y supranacional están asumiendo el desarrollo de lo que se ha dado en llamar Industrias de la Lengua, cuyo objetivo consiste precisamente en la creación de este tipo de infraestructuras. La normalización lingüística, implicará, a partir de ahora, que las lenguas dispongan de los recursos que les permitan participar en la Sociedad de la Información. Esta nueva situación puede agravar los problemas que tienen planteados las lenguas minoritarias, en la medida en que no dispongan de los recursos lingüísticos necesarios para participar en este nuevo orden de cosas. Probablemente, será necesario promover acciones especiales para garantizar los derechos lingüísticos de los hablantes de las diferentes lenguas de la UE.

En el V Programa Marco se contempla el procesamiento del lenguaje como un componente más de las aplicaciones, con la finalidad de mejorar el nivel de interacción de los posibles usuarios con la aplicación. En esta línea, los módulos lingüísticos deberán garantizar tanto la calidad del proceso comunicativo (introducción de datos, consultas, etc.), como el acceso multilingüe y la compatibilidad con otros sistemas de interacción, como el reconocimiento del habla o los entornos multimedia. Este nuevo planteamiento presupone que ya se han alcanzado los objetivos básicos en el terreno de la infraestructura lingüística, y se puede plantear la integración de los recursos disponibles para el desarrollo de aplicaciones y la mejora de servicios que exige la Sociedad de la Información. El comercio electrónico, el acceso a grandes bancos de datos, los sistemas de recuperación de información, contarán, entre otros, con módulos lingüísticos de interacción.

NOTAS

1. Naomi Sager, de la Universidad de Nueva York, desarrolló a partir de 1968 un analizador del inglés basado en la teoría distribucionalista de Z. H. Harris. En 1980 se disponía ya de un analizador de la lengua inglesa de amplia cobertura, utilizado para el procesamiento de informes médicos.

BIBLIOGRAFÍA

Castellón, I. 1992. *Adquisición automática de conocimiento léxico.* Tesis doctoral, Universitat de Barcelona.

Coch, J. 1996. "Evaluating and comparing three text-production techniques". *Proceedings of the 16th Conference of Computational Linguistics: Coling 96.*

Gazdar, G., E. H. Klein, G. K. Pullum y I. A. Sag. 1985. *Generalized Phrase Structure Grammar.* Cambridge: Harvard University Press.

Martí, M. A., I. Castellón y A. Fernández. 1998. "Extracción de información de corpus diccionariales". *Lengua y Tecnologías de la Información.* Eds. J. Gómez Guinovart y M. Palomar. *Novatica* 133: 4-10.

Oakley, B. y K. Owen. 1990. *"Alvey", Britain's Strategic Computing Initiative.* Cambridge: The MIT Press.

Pollard, K. y I. Sag. 1987. *Information-Based Syntax and Semantics. Volume I: Fundamentals.* Stanford: CSLI.

Sager, N., C. Friedman, y M. S. Lyman. 1987. *Medical Language Processing: Computer Management of Narrative Data.* Reading: Addison-Wesley.

Sinclair, J., ed. 1987. *Looking Up: An Account of the Cobuild Project in Lexical Computing.* Londres: Collins.

Vossen, P., ed. 1998. Monográfico dedicado a EuroWordNet. *Computers and the Humanities.* En preparación.