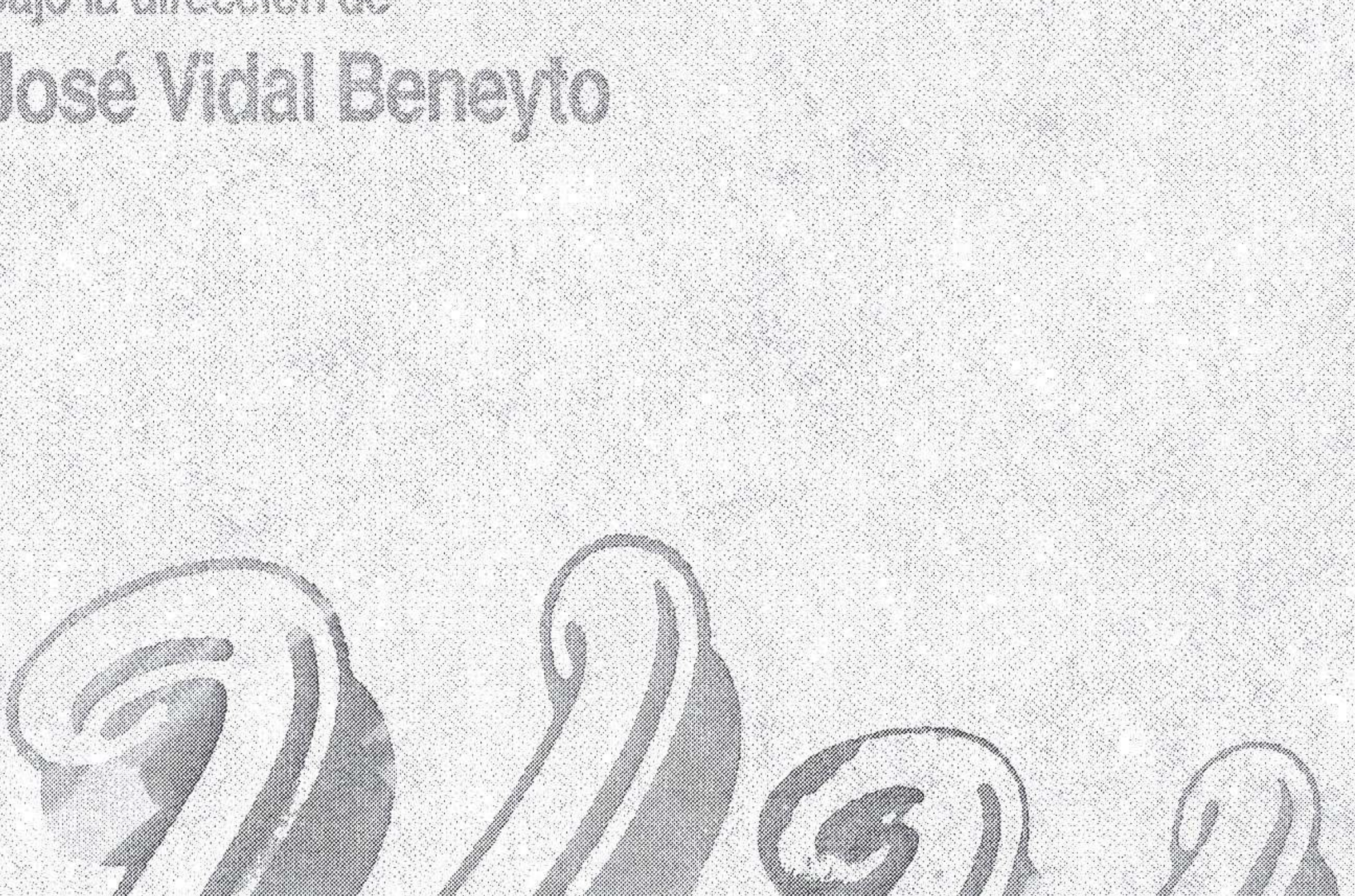
FUNDACIÓN GERMÁN SÁNCHEZ RUIPÉREZ

Biblioteca del Libro

Las industrias de la lengua

Bajo la dirección de

José Vidal Beneyto



Los bancos de datos léxicos: Bases multifuncionales de datos léxicos*

ANTONIO ZAMPOLLI

Trad.: E. Lavín

La automatización ha revolucionado el concepto de lo que es un diccionario, tanto en su aspecto formal como en su uso potencial. Hoy día los diccionarios computarizados están reconocidos como complejas bases de datos, dotadas de múltiples puntos de acceso, que pueden tener gran variedad de estructuras de datos, cuya capacidad de alcance potencial no se ha llegado a explorar. De manera simultánea, los posibles usos de los diccionarios se han multiplicado, por lo que la producción de diccionarios impresos ha llegado a ser un problema de desarrollo más que de investigación. Los diccionarios están considerados como fuentes valiosas, capaces de suministrar información compleja para usos diferentes en el tratamiento de textos, edición de textos, traducción asistida por ordenador, sistemas inteligentes basados en el conocimiento e investigación en inteligencia artificial.

Ya que es importante crear un clima favorable para la innovación al favorecer la experimentación, es también importante la consolidación de forma regular de una experiencia colectiva y una buena práctica con normas y líneas maestras para métodos, procedimientos y estructuras de datos. Sólo de esta manera la creciente industria de la información puede hacer frente a las expectativas del usuario y a la vez a la investigación con cargo a los fondos públicos, y tener una indicación clara de los campos en los que la aplicación de la investigación es más necesaria.

---

* Este artículo ha sido redactado a partir de los resúmenes de diversas contribuciones escritas por Ingria, Cumming, Slocum, Amsler, Schreuder, Bendow, Byrd, Boguraev y Calzolari.

Es, por tanto, deseable un diálogo continuo entre todas las partes implicadas en el diseño, producción y uso de diccionarios con objeto de valorar la progresión, identificando áreas en donde los esfuerzos que se dirijan a la armonización y regulación logren incentivar la cooperación y la difusión de ideas y técnicas nuevas.

En este marco sería útil considerar la promoción de actividades en parcelas especializadas de este campo, con algunos de los siguientes cometidos:

— Definir los diferentes tipos de fuentes léxicas y terminológicas informatizadas.
— Establecer la situación real de los diccionarios automatizados.
— Investigar «las entradas de diccionario» desde diferentes puntos de vista de forma individual, las fuentes tanto en textos como en bancos de datos, los enlaces entre los elementos estructurales de las entradas dentro de un diccionario y las implicaciones de multilingüismo en el diseño de las entradas.
— Considerar la posibilidad de intercambios de datos léxicos.
— Evaluar los posibles usos de bases de datos léxicos en aplicaciones diversas, entre las que se pueden incluir la traducción asistida por ordenador.
— Considerar el impacto de la tecnología moderna en el diseño de diccionarios.
— Identificar áreas que sean susceptibles de armonización.
— Identificar corrientes nuevas en la investigación de diccionarios.
— Hacer recomendaciones sobre las prioridades en investigación y desarrollo.

El concepto de sistema de base multifuncional de datos léxicos (BMDLs):

De los diccionarios automatizados a bases multifuncionales de datos

La característica de multifuncionalismo le viene dada por la capacidad de la base de datos lingüísticos (BDL) de poder ser

utilizada con finalidades varias tanto en diferentes aplicaciones como por diversos tipos de usuarios. Al ampliar la noción de «auxiliar del diccionario» hay que inclinarse por un núcleo único de datos «neutros» de diccionario, al que pueden acceder muchos interfaces diferentes según las necesidades de todo un abanico de posibles aplicaciones. Diferentes procedimientos externos deben ser capaces de utilizar segmentos diferentes del contenido de un diccionario para determinadas aplicaciones, y cada usuario (sea humano o mecánico) percibirá sólo los datos que le son esenciales en cada caso, ignorando, casi por completo, todo lo relativo a su organización física interna.

La característica de «multifuncionalismo» le viene como resultado de su capacidad de acceso múltiple. Con la implementación de múltiples puntos de acceso se pueden buscar los aspectos diferentes de una palabra simplemente siguiendo caminos diferentes en la BDL. Cuando los datos originales se puedan considerar bajo perspectivas diversas, se podrá conseguir el efecto evidente de «multiplicar» la información ofrecida por los propios datos fuente.

Sin embargo, es posible crear, como derivados, muchos subléxicos potencialmente secundarios, que consistirían en partes del diccionario específicamente seleccionadas, como apéndices, diccionarios, sinónimos, tesoros, etc. De hecho, en una BDL amplia y estructurada, estos subléxicos se diferencian sólo en la forma como se seleccionan, clasifican y se relacionan entre sí los datos originales.

Bases multifuncionales de datos lingüísticos en la investigación teórica y aplicada

Las BMDL han de tender a conseguir información estructurada y acabada en muchos niveles descriptivos. Hacen un uso real de la cada vez más perfeccionada tecnología computacional, y parecen constituir uno de los campos de investigación más prometedores, ya que en gran parte tienden a ampliar y desarrollar los logros de proyectos anteriores de diccionarios automatizados, debido a su amplio abanico de áreas de aplicación.

Las BMDL tienen una gran gama de usuarios en potencia:

— Personas (especialistas y lexicólogos, lexicógrafos, lingüistas o usuarios corrientes de diccionarios de consulta).
— «Usuarios procesadores» (es decir, otros programas o sistemas complejos de los que las BMDL son parte integrante).

Una BMDL debe ser diseñada de la manera más flexible, tanto desde una perspectiva computacional como lingüística. Su implementación debe facilitar el uso en muchas ramas del campo de las industrias de la lengua que requieran hacer uso de un lexicón: es decir, prácticamente en todos los casos, porque, sea cual fuere la tarea que realicemos en la que participe el lenguaje natural, en algún momento tendremos que manejar palabras y, por tanto, enfrentarnos a los problemas del acceso léxico.

Las posibles áreas de aplicación de las BMDL van desde la simple verificación ortográfica a la lematización, o a la investigación lexicológica y la práctica lexicográfica (por ejemplo, mejorar la coherencia y consistencia en la realización de diccionarios), o a numerosas aplicaciones de lingüística computacional, como los analizadores, los sistemas de pregunta y respuesta, la comunicación entre hombre y máquina, la traducción asistida por ordenador, la enseñanza de idiomas, etc.

Las BMDL de hecho pueden ser consideradas como punto de conjunción de los diferentes tipos de información al que debe tener acceso todo sistema de procesamiento de lenguaje natural: morfológico, sintáctico, semántico, pragmático o conceptual.

Criterios y métodos para acceder a material léxico ya existente

La implementación de múltiples puntos de acceso en la reestructuración de diccionarios sencillos computarizados como las BMDL nos permite aprovechar al máximo la información suministrada por las definiciones en lenguaje natural contenidas en los diccionarios normales.

Los productos de la tradición lexicográfica, o sea, los diccionarios impresos, son reconocidos ahora no solamente como una de las principales fuentes de datos y de información en torno al

lenguaje, sino también como sólidos bancos de conocimiento general, con un papel cognitivo de importancia. Al utilizar métodos y técnicas de BD, podemos conseguir de los diccionarios ya existentes en forma de lectura automatizada (aunque sólo sea con el simple propósito de la fotocomposición) información utilizable y explícita del lexicón, y podemos descubrir e implementar numerosas relaciones léxicas, morfológicas, semánticas y sintácticas entre las entradas de un diccionario.

Es posible, por ejemplo, a través de análisis y procedimientos apropiados, llevar a cabo una versión sistemática de definiciones en estructuras formalizadas que resuman su contenido informativo. Es posible utilizar un método inductivo a través de la generalización progresiva dada por elementos comunes.

Los problemas que habrá que resolver son los siguientes:

— ¿Es posible integrar recursos léxicos diferentes?
— ¿Es posible concebir BMDL que sean por lo menos parcialmente independientes de las diferentes teorías lingüísticas?
— ¿Cuál es la mejor manera de aprovechar al máximo los diccionarios existentes?

Usos de las BMDL: ejemplos, problemas, posibles acciones

El interés por las BMDL se muestra hoy en diferentes sectores de la investigación y de las aplicaciones. Indicaremos unos cuantos ejemplos y mencionaremos algunos de los problemas conectados con las posibles actividades de investigación y desarrollo.

Análisis computacional y sistemas de generación

Un gran problema de por sí es hasta qué punto y bajo qué condiciones es posible construir una BMDL de modo que la información lingüística pueda ser usada directa o indirectamente mediante sistemas computacionales (para análisis, generación y,

en general, procesamiento de lenguaje natural), que utilicen diferentes estructuras teóricas y computacionales. Este problema no tiene que ser subestimado, ya que la realización del componente léxico —en un sistema computacional de palabras reales— es lo que más tiempo y dinero cuesta.

Se han diseñado analizadores en numerosas concepciones teóricas y computacionales. Puesto que esas concepciones proponen con frecuencia elementos y operaciones lingüísticas diferentes, las consecuencias que se derivan de las diferencias en la construcción de gramáticas para el análisis son inmediatamente evidentes. Menos evidentes, aunque no por ello menos reales, son las consecuencias de naturaleza léxica. Aun así, la información que es, en cierto sentido, idéntica, está representada en lexicones que difieren considerablemente. Parece existir una sucesión de posibilidades para compartir información léxica entre dos sistemas:

1. La información está representada de forma idéntica en los sistemas; por ejemplo, los dos utilizan características con nombres y significados idénticos.
2. Se representa la información de la misma forma en los sistemas; por ejemplo, los dos utilizan características con significados iguales, pero con nombres diferentes.
3. Los fenómenos lingüísticos se analizan de la misma forma en los sistemas, pero las representaciones léxicas no son comparables; por ejemplo, una determinada construcción sintáctica está representada por una única característica en uno de los sistemas, y por la concatenación de muchas en el otro, como se ve en el ejemplo presentado de formulación del control objeto/objeto.
4. Uno de los sistemas hace distinciones más matizadas que el otro en sus análisis; por ejemplo, uno distingue la formulación del objeto desde el control objeto, mientras que el otro no lo hace.
5. Los fenómenos lingüísticos se analizan de forma no comparable en los sistemas; por ejemplo, uno utiliza una base sintáctica en su análisis, mientras que el otro lo hace desde una semántica.

A menudo dos lexicones no difieren en modo alguno de los caminos indicados, más bien en una mezcla de los mismos, o sea,

algunos de los fenómenos se analizan y representan de forma idéntica, otros muestran el mismo análisis y representaciones diferentes, mientras que otros poseen análisis y representaciones diferentes. Las diferencias de 1 a 3 pueden ser sometidas a traducción no asistida por ordenador, mientras que 4 requeriría, como mínimo, la intervención humana en la traducción desde el sistema menos perfeccionado al más perfecto.

Desde un punto de vista preteórico, es posible imaginar varios aspectos de especificación léxica que sean cruciales en el sistema ideal de generación de textos, pero no tanto en el análisis (al menos en el caso de un analizador que admite entrada de datos elaborados, cooperativos e idiomáticos). La selección léxica necesita ser manejada tanto en el eje paradigmático como en el sintagmático.

En el nivel paradigmático, para que un término sea seleccionado, al menos ha de ser apropiado tanto morfológica como semánticamente. Este último requisito implica que el término debe ser no sólo denotacionalmente apropiado, sino también contener la cantidad apropiada de información —esto es, debe poseer un nivel preciso de generalidad y no contener presuposiciones o implicaciones que no sean consistentes con las finalidades del sistema—. De manera ideal, consideraciones tales como el registro de habla tendrían que ser respetadas también en este nivel.

En el nivel sintagmático, se deben respetar las restricciones de coocurrencia de los diferentes niveles de precisión. En el nivel menos elaborado, el sistema debe producir estructuras complementarias sintácticamente gramaticales, mientras que en los niveles de mayor elaboración el sistema tendría que imponer limitaciones semánticas a estos complementos, utilizar frases idiomáticas (en el sentido de frases no composicionales), mantener un registro de habla consistente, e incluso (en un sistema ideal) respetar los tipos de preferencias colocacionales que algunos elementos léxicos poseen para determinados componentes y patrones sintácticos.

Mientras que podemos llamar de manera apropiada a todo lo anterior «consideraciones léxicas», la información utilizada para hacer estas opciones frecuentemente está contenida en los componentes de un generador de textos más que en los del lexicón. Los sistemas existentes difieren considerablemente, tanto en lo referente a los fenómenos que se manejan, como en el lugar en que tratan

la gramática. Esto no debe sorprender, ya que se han creado los sistemas para realizar diversas tareas no similares y que se basan en varias teorías diferentes de comunicación, conocimiento y gramática. En consecuencia, la información que almacena un sistema en el lexicón puede que no esté representada en modo alguno en el otro sistema, o puede que esté distribuida (explícita o implícitamente) a través de cualquiera de los componentes que se ocupan de la semántica, conocimiento cultural, planos textuales y sintaxis.

Después de haber valorado lo que ha sido la información léxica y lo que podría redistribuirse en sistemas computacionales de generación de textos, habría que hacer un esfuerzo ahora para examinar las posibilidades de armonización y reutilización de las fuentes.

Sistema de traducción, diccionarios bilingües

El relanzamiento de proyectos de investigación en los campos de la traducción automática y de la traducción asistida por ordenador ha despertado la atención de investigadores y de instituciones financieras afines por el problema de la construcción de grandes diccionarios multilingües. La necesidad de la investigación se ha hecho evidente en lingüística contrastiva, necesidad claramente relacionada con la de poseer *corpora* técnicos y de referencia de naturaleza multilingüe. El coste y trabajo de la realización de diccionarios bilingües adecuados y de su conexión con BD terminológicas (existentes) son tales que la Comunidad Económica Europea ha financiado un estudio sobre la reutilización de fuentes léxicas.

Para garantizar la corrección en la traducción y la coherencia terminológica, se emplean los glosarios de ordenadores y diccionarios de lectura automatizada. La utilización diferencial de tales materiales, en formas diferentes de traducción, tanto por medio del hombre como de máquinas, es un tema muy importante. La cuestión central del debate serán los diferentes tipos de información requerida por el hombre frente a la máquina, y cómo se relaciona eso con la forma de traducción. Por ejemplo, el hombre tiene acceso directo a los bancos de datos terminológicos, estén

on-line o no, y contienen información útil sólo durante la traducción, y ésta, en general, apenas puede ser empleada por los ordenadores.

En el otro extremo, los diccionarios a los que se tiene acceso por medio de sistemas de traducción totalmente automatizados poseen información útil durante la traducción, pero normalmente no la puede emplear el hombre para muchas otras cosas. En medio están los diccionarios utilizados en actividades humanas asistidas por ordenador; de manera óptima, deberían poseer las dos clases de información. Entonces surge la pregunta: ¿existe una síntesis de las dos clases de información, válida para las dos formas de traducción?

Las cuestiones que deberían ser examinadas son las siguientes:

— ¿Son los diccionarios bilingües impresos una buena o válida fuente de información para la implementación de BDL bilingües, como lo son de forma comprobada los diccionarios impresos monolingües?
— ¿Hay necesidad de otras fuentes de información, como, por ejemplo, las de datos textuales (bi o plurilingües)?
— ¿Es posible ampliar los métodos computacionales usados en la implementación de BD monolingües para las bilingües?
— ¿Son los diccionarios bilingües útiles para conectar dos bases de datos léxicos monolingües?
— ¿Cuál es la mejor estructura para generar vínculos correctos y válidos no simplemente entre palabras léxicas, sino también entre conceptos léxicos?

Bases de conocimiento

Una base léxica de conocimiento (BLC) es una representación computarizada de la información de que disponemos sobre significados de conceptos y de sus relaciones. Proporciona sistemas informatizados con su contrapartida para la comprensión humana del léxico. Para la actividad operativa de una BLC es fundamental que autocomprenda sus propios datos. Las BLC son más que bases de datos en virtud de que sus datos no son almacenados de forma redundante, de que sus valores de datos son comprobados

para ser corregidos, tanto en la forma como en el contenido, y por lo general porque no graban solamente la información nueva que reciben, sino que también la asimilan, y porque en lugar de dar sin más respuestas a consultas de información, lo que hacen es generarlas.

Una BLC no contiene información de cara al público. Contiene información para uso de otros programas; sólo se almacenan los programas que espera recibir y aquellos otros que asumen valores correctos. Una base de conocimiento sabe la clase conceptual o las relaciones ISA que corresponden a todos los valores correctos de todos sus datos.

Los textos existentes de lectura automatizada, ya sean las entradas muy estructuradas de diccionarios u otros libros de referencia, ya sean textos descriptivos de redacción telegráfica, pueden servir como fuentes de información para las bases léxicas de conocimiento.

Para extraer la información de estas fuentes es necesario procesarlas en varias etapas: a partir del texto de fotocomposición se puede transliterar a caracteres generales de ordenador, y entonces se anulan las instrucciones para formatear el texto con objeto de obtener una lista de valores y atributos de la información sintáctica y semántica que contienen, y a partir de aquí se pueden ejecutar etapas adicionales de procesamiento para que se creen bases de datos desde las cuales sea posible agrupar las bases léxicas de conocimiento.

Usando estas técnicas es posible derivar jerarquías ISA y categorizaciones temáticas a partir de las entradas del diccionario, reglas gramaticales y semánticas a partir de códigos de diccionario, al igual que cualquier otra información útil sobre las propiedades de las palabras, como su grado de ambigüedad y capacidad de redefinición gramatical. Los anuarios pueden proporcionar información sobre las diferentes clases de nombres propios al igual que relaciones críticas entre los miembros de estas clases. Las enciclopedias aumentan estos datos aún más y proporcionan conocimientos básicos de vocabulario sobre conceptos elevados. De las entradas de las enciclopedias se pueden sacar definiciones de los conceptos. Por último, a partir de textos descriptivos de redacción telegráfica se pueden obtener actualizaciones y añadidos para todas las fuentes de referencia anteriores, modificando así los

valores cambiantes en anuarios y enciclopedias, al mismo tiempo que se aumenta el lexicón de los diccionarios.

Las técnicas de extracción de información de las fuentes textuales incluyen los tratamientos convencionales de textos y técnicas informáticas más avanzadas de comprensión lingüística de textos. Dado que es imposible diseñar analizadores generales para la entrada ilimitada de texto, es relativamente fácil hacerlo con analizadores expertos de textos en determinados modelos sintácticos. Tales herramientas reconocen fácilmente determinadas clases de información en el texto, y con el tiempo, y a través del procesado de millones de palabras-texto, se puede derivar una cantidad considerable de información cultural. Estos expertos en textos se han diseñado para obtener datos geográficos y biográficos, para el reconocimiento de neologismos, y por lo general como apoyo para la recogida de nombres propios y compuestos que hay que añadir a una base léxica de conocimiento.

Investigación psicolingüística

La palabra es una unidad muy importante en el sistema del lenguaje natural. No solamente las palabras son unidades que portan significado, sino que también funcionan como interfaz entre nuestro sistema cognitivo y el mundo. Por esta y otras razones, las palabras desempeñan un papel esencial en la investigacion lingüística.

Las palabras se utilizan en la investigación psicolingüística como unidades de investigación en sí mismas, o como subunidades dentro de unidades mayores de interés. Si se estudian ciertas propiedades de estas últimas (las subunidades), entonces se definirán clases diferentes de unidades, reflejando cada una de forma diferente la propiedad que se estudia.

La lógica de la investigación experimental determina, por tanto, que estos tipos diferentes de unidades no tengan que diferir en ningún otro aspecto.

Por tanto, es necesario comparar estas clases en muchas de las propiedades que tiene la palabra y que pueden desempeñar un papel en ese experimento determinado.

La razón para la necesidad de conocer las propiedades de las palabras es doble: una es el interés por las propiedades en sí, y la otra tiene que ver con los principios de diseño experimental.

Una base de datos léxicos es extremadamente útil en psicolingüística experimental como modo de acceso a las muchas propiedades de las palabras de una lengua. La estructura e información de un BDL (propiedades de la palabra tales como categoría sintáctica, composición ortográfica, estructura morfológica, composición fonológica, estructura silábica, rasgos de acento, contorno de la palabra, frecuencia) tienen que investigarse en función de las consultas que debe formular un psicolingüista experimental a una base de datos léxicos.

Las aplicaciones prácticas (por ejemplo, en el aprendizaje de idiomas) deben ser sometidas a debate.

En la consecución de objetivos psicolingüísticos, algunos investigadores han usado BDL para comprobar la plausibilidad y aplicabilidad de algunos conceptos en las teorías de reconocimiento auditivo de la palabra y, en general, para responder a ciertas preguntas cruciales sobre el proceso del lenguaje humano.

De una vez por todas, no se puede facilitar una «lista de la compra». Hay, sin duda, propiedades esenciales que una BDL debería poseer. Son bastante comunes a las exigidas por lingüistas y lingüistas informáticos computacionales. Particularmente, los investigadores en psicolingüística piden ya recuentos adecuados y periódicamente actualizados de frecuencia.

Sistemas de diccionario para el mercado de masas

El mercado de diccionarios de lectura automatizada (DLA) abarcaría tres tipos principales de aplicación:

— Donde se utiliza el DLA como herramienta en el tratamiento de textos.
— Como herramienta de trabajo intelectual en la investigación literaria, lingüística o cualquier otra académica.
— Referencias generales —una alternativa más flexible al uso del diccionario impreso.

Los DLA en la oficina: Dos tendencias parecen influir en el uso de los DLA en la oficina: la primera parece ser el trasvase del trabajo de escribir a máquina, desde auxiliares y oficinistas a jefes y ejecutivos. Es el resultado natural del número de estaciones de trabajo que hay en mesas de ejecutivos, y el desarrollo de redes de trabajo. El ejecutivo que tiene que mecanografiar el correo electrónico va a necesitar más que un procesador de textos, quizá algo parecido al Proyecto Epístola de IBM. La segunda tendencia —y relacionada con la anterior— es el avance espectacular de la tecnología del ordenador, lo que significa que la capacidad de procesamiento y de almacenamiento no constituye un factor de limitación.

Éste será quizá el primer mercado para los DLA. Sin embargo, las exigencias de utilización puede que no sean homogéneas, al menos inicialmente. Sin duda alguna, habrá un desarrollo de estos mercados de exigencias muy específicas y de requisitos especializados. Pero esto será ya en la segunda etapa. Lo que se necesitará inicialmente será un DLA que ofrezca una gama completa de información léxica dotado de software que permita al usuario conseguir el tipo adecuado y el nivel de información con pocas dificultades.

El hecho de que un DLA contenga cantidades de datos innecesarios para el usuario no será una barrera de la manera que lo sería un diccionario convencional.

Sin duda alguna será preciso adaptar tanto el contenido como el software a las necesidades del ejecutivo. Los tipos de información que sin duda se necesitarán serán, por ejemplo, vocabulario especializado relacionado con los campos técnicos/profesionales, nuevos términos y acceso fácil a un tesoro.

En la casa: El ritmo acelerado del cambio tecnológico es posible que fomente también un mercado de DLA de uso doméstico.

Como consecuencia de la gran capacidad cada vez mayor de almacenaje y procesamiento, es probable que la informática doméstica ofrezca una completa gama de funciones que incluya no sólo las aplicaciones individuales (como juegos y tratamiento de textos), sino también medios de comunicación (como el correo electrónico y la banca desde casa) y de información (bolsa de valores e información enciclopédica y léxica).

Estos medios serán no sólo un instrumento recreativo, como lo ha sido hasta ahora el ordenador doméstico. También será una herramienta para la escritura creativa, para el diseño, para la composición musical y literaria, etc. El valor extra que se añade por la disponibilidad de información léxica puede justificar en muchos casos el desembolso exigido para la compra de un sistema de este tipo.

En la enseñanza: Este sector, sobre todo la utilización del ordenador para un mejor y más conveniente uso del material de consulta, se está empezando a estudiar.

Ahora se comienza a pensar en la utilización de DLA en la adquisición de un segundo idioma para perfeccionamiento de la expresión escrita. Cito de un artículo de Benbow. El desarrollo de un mercado para el nativo de lengua inglesa estará, como en otros campos, dictado por la disponibilidad de una tecnología apropiada.

Se pueden prever importantes y amplios usos de los DLA (o una serie determinada de ellos) en centros de enseñanza y formación, en todos los niveles educativos. No es probable, sin embargo, que la disponibilidad sólo de los DLA sea suficiente para generar un mercado importante.

Si los centros educativos y de formación están utilizando equipos sobre los que se podría instalar los DLA, por un modesto desembolso para instalar DLA (o una serie de ellos) se podría dotar al sistema de valor significativo.

En cuanto a la enseñanza del inglés como lengua extranjera, se podría ver más fácilmente el desarrollo de un mercado autónomo de DLA. La enseñanza de la lengua inglesa siempre ha sido rápida en adoptar la tecnología apropiada, como el uso del magnetófono y el casete en los laboratorios de idiomas y el desarrollo del vídeo como instrumento de enseñanza.

Obviamente, el tipo de información requerida por un hablante de lengua inglesa no nativo tendrá que diferir bastante del que necesita el nativo: la información sobre la tipología verbal y el uso contextual serán prioritarios, por citar un caso. Por tanto, como este tipo de información aparece ya en diccionarios de enseñanza de lengua inglesa, este requisito de especialización no supondrá un problema.

Consecuencias para la lexicografía

1. Los usos de un DLA serán diferentes de los de un diccionario impreso. Una encuesta reciente del OED (Old English Dictionary) indica que esto será lo que ocurrirá por lo menos en lo que se refiere a este diccionario. La diferencia en cuanto al uso, sin embargo, no estriba simplemente en diferentes medios de información sobre los que el texto se apoya. El usuario de DLA en muchos casos estará buscando información no fácilmente conseguible en un diccionario impreso, y las vías de acceso a la información serán infinitamente más variadas que las que puede ofrecer un diccionario impreso. Un ejemplo fundamental sería una consulta a la inversa (hallar un término a partir de su significado) y otras exploraciones semánticas.

2. Surgirán nuevos tipos de usuarios al disponer de DLA. El uso del diccionario impreso se limita en la práctica a conseguir información sobre lemas conocidos. Con los DLA será posible conseguir información sobre grupos de palabras relacionados por uno o varios factores; por clasificación temática, clase gramatical, pronunciación, lengua de origen, etc. La lista no tiene límites. Por ello, las implicaciones completas son imposibles de predecir.

3. Los diccionarios impresos y sus homólogos de lectura automatizada tomarán rumbos opuestos. Los usos varios, los grupos diferentes de usuarios, la distinta naturaleza harán que la correspondencia inicial entre diccionarios impresos y sus homólogos de lectura automatizada desaparezca pronto. Si, como parece probable, los DLA proliferan, surgirá el problema de mantener los normales. La cantidad de material puede que sustituya a la calidad como principal preocupación de los fabricantes de DLA; el planteamiento de un equilibrio aceptable debería ser la preocupación principal de todos aquellos que trabajen en el desarrollo de DLA.

4. El software será una clave determinante en la utilidad de un DLA. Se necesitará mucha investigación y desarrollo para proporcionar un software lo suficientemente aceptable que haga que los DLA sean realmente valiosos.

5. La influencia del lexicógrafo puede afectar al desarrollo de un DLA. En la preparación de un diccionario impreso el papel

que desempeña un lexicógrafo es capital. Puede que no sea ése el caso en el desarrollo de los DLA. La influencia del usuario, del especialista en ciencias de la información y del especialista en marketing puede que sea significativamente más importante en la fabricación de un DLA. Esto nos lleva al problema de la calidad: el lexicógrafo es el factor más importante para el control de calidad. Si su papel queda degradado, la calidad del producto estará en peligro.

6. Los DLA no podrán reemplazar por completo a los diccionarios. No debe pasarse por alto el hecho de que para las cuestiones más simples (por ejemplo, el significado, grafía, pronunciación de una palabra determinada) el diccionario impreso constituye la mejor y más conveniente manera de conseguir la información solicitada (a menos que, evidentemente, en el momento que se necesite la información se esté trabajando en una estación de trabajo con acceso a un DLA).

7. Es posible que para muchos usuarios de diccionario la existencia de grabados e ilustraciones sea un buen complemento. Lo mismo puede decirse del vídeo y del material audiovisual que puede aparecer como material de ayuda de las definiciones verbales. Ciertas características técnicas adicionales —como la reproducción sintética de la pronunciación de una forma determinada— aumentarán el valor del DLA.

Reutilización de diccionarios de lectura automatizada

El número de diccionarios de lectura automatizada aumenta debido a la difusión de la fotocomposición. La posibilidad que ofrecen las modernas tecnologías en la distribución de productos lexicográficos tradicionales o nuevos constituirá otro factor para disponer cada vez más de recursos léxicos en forma de lectura automatizada.

Sin duda, los diccionarios tradicionales no contienen toda la información necesaria para los sistemas de procesamiento del lenguaje natural.

Por otra parte, poseen, sin embargo, bastante información más

o menos organizada y pertinente con respecto a los sistemas computacionales. No obstante, parece conveniente evaluar, en el nivel metodológico y de organización, las posibilidades existentes para emplear la información disponible en DLA en diferentes aplicaciones informáticas.

La reorganización de DLA en forma de bases de datos con una metodología propia de integración puede que abra, al usuario especialista común, el acceso a información alfabetizada de recuperación difícil.

Los DLA y la investigación en lingüística computacional

Considerando el gran número de sistemas que hay dentro del paradigma de la lingüística computacional, sorprende que exista un número tan reducido de entradas de diccionario disponibles en estos sistemas.

Hay que admitir que la mayor parte de estos sistemas son experimentales; sin embargo, para las aplicaciones con palabras reales, teniendo en cuenta los avances recientes en la tecnología computacional que hacen que estas aplicaciones sean factibles, se necesitan vocabularios de mucha mayor extensión. Dado el número de cortapisas reales que encierran muchos de los sistemas, siendo los más críticos las formalizaciones idiosincráticas para redactar entradas de diccionario, parecen ser un reto, y de hecho lo son, en la labor de crear un vocabulario amplio y sólido para cualquier aplicación, los diferentes formatos de lexicón, y las diferentes consideraciones sobre lo que constituye la información lingüísticamente relevante (sintáctica, semántica y pragmática), cuyo lugar apropiado es el lexicón. Y apenas sorprende, por tanto, que un número de investigadores esté tras la consecución de diccionarios de lectura automatizada (DLA). Esperan que la información ya ordenada, categorizada, indexada y, sobre todo, disponible en forma de lectura automatizada, sea convenientemente utilizada, si no para tener un lexicón considerable «de la nada», sí por lo menos para crear automáticamente una parte sustancial del mismo, consiguiendo de la nada lo que se tiene la esperanza que sea

un objeto internamente consistente (y coherente). Un lexicón logrado de esta forma ahorrará esfuerzos y creará el volumen de vocabulario terminal que podría ser por consiguiente ampliado, si fuera necesario, y adaptado a las tareas y aplicaciones pertinentes.

Existe un trabajo paralelo afín, en línea con la corriente actual de «programación basada en el conocimiento», que plantea el problema de la adquisición de segmentos fundamentales de conocimiento estructurado sobre la propia palabra. La esperanza gira aquí de manera similar en torno a la creencia de que se pueden encontrar caminos para localizar y extraer parte del conocimiento requerido para realizar un tipo de funciones (semi)inteligentes de fuentes de lectura automatizada, por ejemplo, de diccionarios y enciclopedias.

Será útil analizar los caminos que recorren muchos investigadores en busca de esos fines, usando diccionarios diferentes y utilizando técnicas diversas en apoyo de aplicaciones varias. Se pondrá énfasis especial en determinar la cantidad de información disponible en forma de lectura automatizada y cuánta podrá utilizarse en la investigación lingüística computacional.

Dado que es cierto que los sistemas difieren en organización, estructura y contenido de sus lexicones, aún es posible aislar ciertos tipos de información, determinados por la función y las aplicaciones particulares, que deberían estar disponibles a nivel léxico. Una categorización de los requisitos léxicos, a medida que varían en una gama de aplicaciones dentro del mismo marco de la lingüística computacional, aportará una nueva dimensión para evaluar la utilidad de la información léxica que pueda conseguirse en forma de lectura mecanizada.

Una fuente de lectura automatizada puede ser percibida como una base de datos léxicos que ofrece información, por ejemplo, sobre los núcleos silábicos de una palabra o sobre un comportamiento sintáctico idiosincrático. Dicha información puede que sea esencial para un sistema de síntesis del lenguaje o para un programa de análisis, pero es de poca utilidad, por ejemplo, para un componente de interpretación de un sistema finito de lenguaje natural. En el último caso, la intención es la de considerar el diccionario como una especie de base de conocimiento, donde la información de naturaleza semántica más general estaría codificada, o en un formato de texto libre, o en alguna forma de etiqueta-

ción semántica. De interés general en ambos casos son cuestiones tales como los tipos de datos que estarán disponibles en forma de lectura mecanizada, y el tipo de sistema computacional a los que le puedan ser útiles los datos.

Una gran parte del trabajo se ha empleado en intentar dar soluciones computacionales a un gran número de problemas, que radican en la información extraíble de un DLA. El análisis sintáctico, el desarrollo gramatical, la selección de sentido de la palabra, la síntesis del lenguaje, la sólida interpretación textual, la adquisición de conocimientos, la organización de la información, el acceso léxico con ayuda fonética, el análisis de la frase, son sólo unas cuantas de las funciones que se han favorecido sustancialmente, ya que se disponía de grandes diccionarios *on-line*.

Una experiencia colectiva hace factible, y por supuesto muy necesario, examinar aspectos más generales del trabajo con diccionarios de lectura automatizada. Ahora ya es posible preguntarse por cuestiones tales como ¿cuál es el costo total (el esfuerzo de mano de obra dedicado al proyecto) de intentar enjaezar lo que con frecuencia es un objetivo tosco y voluminoso? ¿Qué tipo de información se puede esperar razonablemente de un DLA? ¿Hasta dónde es fiable esta información? ¿Cuál sería la mejor manera de utilizarlo en un contexto determinado? El consenso sobre estos temas, junto a la opinión sobre el puesto que debe ocupar en esa escala la gama de diccionarios disponibles, será de gran utilidad para la comunidad investigadora.

Estructura y acceso en los lexicones automatizados

El problema de la estructura y acceso en relación con la organización de un lexicón automatizado se puede ver desde varias perspectivas:

— Como problemas de software y de hardware.
— Como problemas de estructura conceptual o lógica de los datos, es decir, en una base de datos o en una red.
— Como problemas de estructura «léxica» y de relaciones

entre las palabras y, por tanto, de la conexión y concepto lingüísticos.
— Como problemas de naturaleza más psicolingüística, si se estimula el proceso mental de acceso léxico.
— Como problemas relacionados con los fines perseguidos.
— Como problema relacionado con los usuarios (personas o procesos) de los datos.

El tipo de organización y el consiguiente acceso que se requieren son muy diferentes si tenemos que crear un diccionario de simple lectura automatizada, o una base de datos «estructurada» o incluso una base de conocimiento más compleja. En el primer caso, sin duda es suficiente una «cadena» o «formato de textos», mientras que en los casos restantes es posible que una organización más estructurada/codificada proporcione los mejores resultados.

La reestructuración de los datos de un DLA organizado en forma de base de datos puede que haga posible lograr información adicional sobre los datos originales, multiplicados por un factor dado por el número de perspectivas diferentes que se establezcan en los datos, por ejemplo, de las diferentes relaciones de ordenación o de los varios caminos que conecten los datos.

El lexicón aparecerá de hecho dividido en tantos subgrupos como relaciones hayan sido determinadas y formalizadas. Al representar el lexicón por el conjunto de esas relaciones, podemos tener acceso al diccionario a través de las unidades léxicas, o a través de rasgos, o de relaciones, o podemos buscar la red para comprobar dónde se empareja con la pregunta formulada, y recoge las diferentes partes del contenido léxico en relación tanto con el punto de entrada como de las opciones activadas en ese punto.

El concepto del lexicón en sí mismo puede ser ampliamente tratado y cambiado en el marco informático, sobre todo si se considera como un objeto mucho más interrelacionado, donde las palabras están compuestas en muchos niveles, y cada tipo de complejidad tiene que ser analizado y tratado, de manera que cada una de estas complejidades en varios estratos pueda ser captada o por reglas generales o por funciones apropiadas de acceso.

FUNDACIÓN GERMÁN SÁNCHEZ RUIPÉREZ

Biblioteca del Libro

Las industrias de la lengua

Bajo la dirección de

José Vidal Beneyto

Los bancos de datos léxicos: Bases multifuncionales de datos léxicos*

ANTONIO ZAMPOLLI

Trad.: E. Lavín

La automatización ha revolucionado el concepto de lo que es un diccionario, tanto en su aspecto formal como en su uso potencial. Hoy día los diccionarios computarizados están reconocidos como complejas bases de datos, dotadas de múltiples puntos de acceso, que pueden tener gran variedad de estructuras de datos, cuya capacidad de alcance potencial no se ha llegado a explorar. De manera simultánea, los posibles usos de los diccionarios se han multiplicado, por lo que la producción de diccionarios impresos ha llegado a ser un problema de desarrollo más que de investigación. Los diccionarios están considerados como fuentes valiosas, capaces de suministrar información compleja para usos diferentes en el tratamiento de textos, edición de textos, traducción asistida por ordenador, sistemas inteligentes basados en el conocimiento e investigación en inteligencia artificial.

Ya que es importante crear un clima favorable para la innovación al favorecer la experimentación, es también importante la consolidación de forma regular de una experiencia colectiva y una buena práctica con normas y líneas maestras para métodos, procedimientos y estructuras de datos. Sólo de esta manera la creciente industria de la información puede hacer frente a las expectativas del usuario y a la vez a la investigación con cargo a los fondos públicos, y tener una indicación clara de los campos en los que la aplicación de la investigación es más necesaria.

---

* Este artículo ha sido redactado a partir de los resúmenes de diversas contribuciones escritas por Ingria, Cumming, Slocum, Amsler, Schreuder, Bendow, Byrd, Boguraev y Calzolari.

Es, por tanto, deseable un diálogo continuo entre todas las partes implicadas en el diseño, producción y uso de diccionarios con objeto de valorar la progresión, identificando áreas en donde los esfuerzos que se dirijan a la armonización y regulación logren incentivar la cooperación y la difusión de ideas y técnicas nuevas.

En este marco sería útil considerar la promoción de actividades en parcelas especializadas de este campo, con algunos de los siguientes cometidos:

- Definir los diferentes tipos de fuentes léxicas y terminológicas informatizadas.
- Establecer la situación real de los diccionarios automatizados.
- Investigar «las entradas de diccionario» desde diferentes puntos de vista de forma individual, las fuentes tanto en textos como en bancos de datos, los enlaces entre los elementos estructurales de las entradas dentro de un diccionario y las implicaciones de multilingüismo en el diseño de las entradas.
- Considerar la posibilidad de intercambios de datos léxicos.
- Evaluar los posibles usos de bases de datos léxicos en aplicaciones diversas, entre las que se pueden incluir la traducción asistida por ordenador.
- Considerar el impacto de la tecnología moderna en el diseño de diccionarios.
- Identificar áreas que sean susceptibles de armonización.
- Identificar corrientes nuevas en la investigación de diccionarios.
- Hacer recomendaciones sobre las prioridades en investigación y desarrollo.

El concepto de sistema de base multifuncional de datos léxicos (BMDLs):

De los diccionarios automatizados a bases multifuncionales de datos

La característica de multifuncionalismo le viene dada por la capacidad de la base de datos lingüísticos (BDL) de poder ser

utilizada con finalidades varias tanto en diferentes aplicaciones como por diversos tipos de usuarios. Al ampliar la noción de «auxiliar del diccionario» hay que inclinarse por un núcleo único de datos «neutros» de diccionario, al que pueden acceder muchos interfaces diferentes según las necesidades de todo un abanico de posibles aplicaciones. Diferentes procedimientos externos deben ser capaces de utilizar segmentos diferentes del contenido de un diccionario para determinadas aplicaciones, y cada usuario (sea humano o mecánico) percibirá sólo los datos que le son esenciales en cada caso, ignorando, casi por completo, todo lo relativo a su organización física interna.

La característica de «multifuncionalismo» le viene como resultado de su capacidad de acceso múltiple. Con la implementación de múltiples puntos de acceso se pueden buscar los aspectos diferentes de una palabra simplemente siguiendo caminos diferentes en la BDL. Cuando los datos originales se puedan considerar bajo perspectivas diversas, se podrá conseguir el efecto evidente de «multiplicar» la información ofrecida por los propios datos fuente.

Sin embargo, es posible crear, como derivados, muchos subléxicos potencialmente secundarios, que consistirían en partes del diccionario específicamente seleccionadas, como apéndices, diccionarios, sinónimos, tesoros, etc. De hecho, en una BDL amplia y estructurada, estos subléxicos se diferencian sólo en la forma como se seleccionan, clasifican y se relacionan entre sí los datos originales.

Bases multifuncionales de datos lingüísticos en la investigación teórica y aplicada

Las BMDL han de tender a conseguir información estructurada y acabada en muchos niveles descriptivos. Hacen un uso real de la cada vez más perfeccionada tecnología computacional, y parecen constituir uno de los campos de investigación más prometedores, ya que en gran parte tienden a ampliar y desarrollar los logros de proyectos anteriores de diccionarios automatizados, debido a su amplio abanico de áreas de aplicación.

Las BMDL tienen una gran gama de usuarios en potencia:

— Personas (especialistas y lexicólogos, lexicógrafos, lingüistas o usuarios corrientes de diccionarios de consulta).
— «Usuarios procesadores» (es decir, otros programas o sistemas complejos de los que las BMDL son parte integrante).

Una BMDL debe ser diseñada de la manera más flexible, tanto desde una perspectiva computacional como lingüística. Su implementación debe facilitar el uso en muchas ramas del campo de las industrias de la lengua que requieran hacer uso de un lexicón: es decir, prácticamente en todos los casos, porque, sea cual fuere la tarea que realicemos en la que participe el lenguaje natural, en algún momento tendremos que manejar palabras y, por tanto, enfrentarnos a los problemas del acceso léxico.

Las posibles áreas de aplicación de las BMDL van desde la simple verificación ortográfica a la lematización, o a la investigación lexicológica y la práctica lexicográfica (por ejemplo, mejorar la coherencia y consistencia en la realización de diccionarios), o a numerosas aplicaciones de lingüística computacional, como los analizadores, los sistemas de pregunta y respuesta, la comunicación entre hombre y máquina, la traducción asistida por ordenador, la enseñanza de idiomas, etc.

Las BMDL de hecho pueden ser consideradas como punto de conjunción de los diferentes tipos de información al que debe tener acceso todo sistema de procesamiento de lenguaje natural: morfológico, sintáctico, semántico, pragmático o conceptual.

Criterios y métodos para acceder a material léxico ya existente

La implementación de múltiples puntos de acceso en la reestructuración de diccionarios sencillos computarizados como las BMDL nos permite aprovechar al máximo la información suministrada por las definiciones en lenguaje natural contenidas en los diccionarios normales.

Los productos de la tradición lexicográfica, o sea, los diccionarios impresos, son reconocidos ahora no solamente como una de las principales fuentes de datos y de información en torno al

lenguaje, sino también como sólidos bancos de conocimiento general, con un papel cognitivo de importancia. Al utilizar métodos y técnicas de BD, podemos conseguir de los diccionarios ya existentes en forma de lectura automatizada (aunque sólo sea con el simple propósito de la fotocomposición) información utilizable y explícita del lexicón, y podemos descubrir e implementar numerosas relaciones léxicas, morfológicas, semánticas y sintácticas entre las entradas de un diccionario.

Es posible, por ejemplo, a través de análisis y procedimientos apropiados, llevar a cabo una versión sistemática de definiciones en estructuras formalizadas que resuman su contenido informativo. Es posible utilizar un método inductivo a través de la generalización progresiva dada por elementos comunes.

Los problemas que habrá que resolver son los siguientes:

— ¿Es posible integrar recursos léxicos diferentes?
— ¿Es posible concebir BMDL que sean por lo menos parcialmente independientes de las diferentes teorías lingüísticas?
— ¿Cuál es la mejor manera de aprovechar al máximo los diccionarios existentes?

Usos de las BMDL: ejemplos, problemas, posibles acciones

El interés por las BMDL se muestra hoy en diferentes sectores de la investigación y de las aplicaciones. Indicaremos unos cuantos ejemplos y mencionaremos algunos de los problemas conectados con las posibles actividades de investigación y desarrollo.

Análisis computacional y sistemas de generación

Un gran problema de por sí es hasta qué punto y bajo qué condiciones es posible construir una BMDL de modo que la información lingüística pueda ser usada directa o indirectamente mediante sistemas computacionales (para análisis, generación y,

en general, procesamiento de lenguaje natural), que utilicen diferentes estructuras teóricas y computacionales. Este problema no tiene que ser subestimado, ya que la realización del componente léxico —en un sistema computacional de palabras reales— es lo que más tiempo y dinero cuesta.

Se han diseñado analizadores en numerosas concepciones teóricas y computacionales. Puesto que esas concepciones proponen con frecuencia elementos y operaciones lingüísticas diferentes, las consecuencias que se derivan de las diferencias en la construcción de gramáticas para el análisis son inmediatamente evidentes. Menos evidentes, aunque no por ello menos reales, son las consecuencias de naturaleza léxica. Aun así, la información que es, en cierto sentido, idéntica, está representada en lexicones que difieren considerablemente. Parece existir una sucesión de posibilidades para compartir información léxica entre dos sistemas:

1. La información está representada de forma idéntica en los sistemas; por ejemplo, los dos utilizan características con nombres y significados idénticos.
2. Se representa la información de la misma forma en los sistemas; por ejemplo, los dos utilizan características con significados iguales, pero con nombres diferentes.
3. Los fenómenos lingüísticos se analizan de la misma forma en los sistemas, pero las representaciones léxicas no son comparables; por ejemplo, una determinada construcción sintáctica está representada por una única característica en uno de los sistemas, y por la concatenación de muchas en el otro, como se ve en el ejemplo presentado de formulación del control objeto/objeto.
4. Uno de los sistemas hace distinciones más matizadas que el otro en sus análisis; por ejemplo, uno distingue la formulación del objeto desde el control objeto, mientras que el otro no lo hace.
5. Los fenómenos lingüísticos se analizan de forma no comparable en los sistemas; por ejemplo, uno utiliza una base sintáctica en su análisis, mientras que el otro lo hace desde una semántica.

A menudo dos lexicones no difieren en modo alguno de los caminos indicados, más bien en una mezcla de los mismos, o sea,

algunos de los fenómenos se analizan y representan de forma idéntica, otros muestran el mismo análisis y representaciones diferentes, mientras que otros poseen análisis y representaciones diferentes. Las diferencias de 1 a 3 pueden ser sometidas a traducción no asistida por ordenador, mientras que 4 requeriría, como mínimo, la intervención humana en la traducción desde el sistema menos perfeccionado al más perfecto.

Desde un punto de vista preteórico, es posible imaginar varios aspectos de especificación léxica que sean cruciales en el sistema ideal de generación de textos, pero no tanto en el análisis (al menos en el caso de un analizador que admite entrada de datos elaborados, cooperativos e idiomáticos). La selección léxica necesita ser manejada tanto en el eje paradigmático como en el sintagmático.

En el nivel paradigmático, para que un término sea seleccionado, al menos ha de ser apropiado tanto morfológica como semánticamente. Este último requisito implica que el término debe ser no sólo denotacionalmente apropiado, sino también contener la cantidad apropiada de información —esto es, debe poseer un nivel preciso de generalidad y no contener presuposiciones o implicaciones que no sean consistentes con las finalidades del sistema—. De manera ideal, consideraciones tales como el registro de habla tendrían que ser respetadas también en este nivel.

En el nivel sintagmático, se deben respetar las restricciones de coocurrencia de los diferentes niveles de precisión. En el nivel menos elaborado, el sistema debe producir estructuras complementarias sintácticamente gramaticales, mientras que en los niveles de mayor elaboración el sistema tendría que imponer limitaciones semánticas a estos complementos, utilizar frases idiomáticas (en el sentido de frases no composicionales), mantener un registro de habla consistente, e incluso (en un sistema ideal) respetar los tipos de preferencias colocacionales que algunos elementos léxicos poseen para determinados componentes y patrones sintácticos.

Mientras que podemos llamar de manera apropiada a todo lo anterior «consideraciones léxicas», la información utilizada para hacer estas opciones frecuentemente está contenida en los componentes de un generador de textos más que en los del lexicón. Los sistemas existentes difieren considerablemente, tanto en lo referente a los fenómenos que se manejan, como en el lugar en que tratan

la gramática. Esto no debe sorprender, ya que se han creado los sistemas para realizar diversas tareas no similares y que se basan en varias teorías diferentes de comunicación, conocimiento y gramática. En consecuencia, la información que almacena un sistema en el lexicón puede que no esté representada en modo alguno en el otro sistema, o puede que esté distribuida (explícita o implícitamente) a través de cualquiera de los componentes que se ocupan de la semántica, conocimiento cultural, planos textuales y sintaxis.

Después de haber valorado lo que ha sido la información léxica y lo que podría redistribuirse en sistemas computacionales de generación de textos, habría que hacer un esfuerzo ahora para examinar las posibilidades de armonización y reutilización de las fuentes.

Sistema de traducción, diccionarios bilingües

El relanzamiento de proyectos de investigación en los campos de la traducción automática y de la traducción asistida por ordenador ha despertado la atención de investigadores y de instituciones financieras afines por el problema de la construcción de grandes diccionarios multilingües. La necesidad de la investigación se ha hecho evidente en lingüística contrastiva, necesidad claramente relacionada con la de poseer *corpora* técnicos y de referencia de naturaleza multilingüe. El coste y trabajo de la realización de diccionarios bilingües adecuados y de su conexión con BD terminológicas (existentes) son tales que la Comunidad Económica Europea ha financiado un estudio sobre la reutilización de fuentes léxicas.

Para garantizar la corrección en la traducción y la coherencia terminológica, se emplean los glosarios de ordenadores y diccionarios de lectura automatizada. La utilización diferencial de tales materiales, en formas diferentes de traducción, tanto por medio del hombre como de máquinas, es un tema muy importante. La cuestión central del debate serán los diferentes tipos de información requerida por el hombre frente a la máquina, y cómo se relaciona eso con la forma de traducción. Por ejemplo, el hombre tiene acceso directo a los bancos de datos terminológicos, estén

on-line o no, y contienen información útil sólo durante la traducción, y ésta, en general, apenas puede ser empleada por los ordenadores.

En el otro extremo, los diccionarios a los que se tiene acceso por medio de sistemas de traducción totalmente automatizados poseen información útil durante la traducción, pero normalmente no la puede emplear el hombre para muchas otras cosas. En medio están los diccionarios utilizados en actividades humanas asistidas por ordenador; de manera óptima, deberían poseer las dos clases de información. Entonces surge la pregunta: ¿existe una síntesis de las dos clases de información, válida para las dos formas de traducción?

Las cuestiones que deberían ser examinadas son las siguientes:

— ¿Son los diccionarios bilingües impresos una buena o válida fuente de información para la implementación de BDL bilingües, como lo son de forma comprobada los diccionarios impresos monolingües?
— ¿Hay necesidad de otras fuentes de información, como, por ejemplo, las de datos textuales (bi o plurilingües)?
— ¿Es posible ampliar los métodos computacionales usados en la implementación de BD monolingües para las bilingües?
— ¿Son los diccionarios bilingües útiles para conectar dos bases de datos léxicos monolingües?
— ¿Cuál es la mejor estructura para generar vínculos correctos y válidos no simplemente entre palabras léxicas, sino también entre conceptos léxicos?

Bases de conocimiento

Una base léxica de conocimiento (BLC) es una representación computarizada de la información de que disponemos sobre significados de conceptos y de sus relaciones. Proporciona sistemas informatizados con su contrapartida para la comprensión humana del léxico. Para la actividad operativa de una BLC es fundamental que autocomprenda sus propios datos. Las BLC son más que bases de datos en virtud de que sus datos no son almacenados de forma redundante, de que sus valores de datos son comprobados

para ser corregidos, tanto en la forma como en el contenido, y por lo general porque no graban solamente la información nueva que reciben, sino que también la asimilan, y porque en lugar de dar sin más respuestas a consultas de información, lo que hacen es generarlas.

Una BLC no contiene información de cara al público. Contiene información para uso de otros programas; sólo se almacenan los programas que espera recibir y aquellos otros que asumen valores correctos. Una base de conocimiento sabe la clase conceptual o las relaciones ISA que corresponden a todos los valores correctos de todos sus datos.

Los textos existentes de lectura automatizada, ya sean las entradas muy estructuradas de diccionarios u otros libros de referencia, ya sean textos descriptivos de redacción telegráfica, pueden servir como fuentes de información para las bases léxicas de conocimiento.

Para extraer la información de estas fuentes es necesario procesarlas en varias etapas: a partir del texto de fotocomposición se puede transliterar a caracteres generales de ordenador, y entonces se anulan las instrucciones para formatear el texto con objeto de obtener una lista de valores y atributos de la información sintáctica y semántica que contienen, y a partir de aquí se pueden ejecutar etapas adicionales de procesamiento para que se creen bases de datos desde las cuales sea posible agrupar las bases léxicas de conocimiento.

Usando estas técnicas es posible derivar jerarquías ISA y categorizaciones temáticas a partir de las entradas del diccionario, reglas gramaticales y semánticas a partir de códigos de diccionario, al igual que cualquier otra información útil sobre las propiedades de las palabras, como su grado de ambigüedad y capacidad de redefinición gramatical. Los anuarios pueden proporcionar información sobre las diferentes clases de nombres propios al igual que relaciones críticas entre los miembros de estas clases. Las enciclopedias aumentan estos datos aún más y proporcionan conocimientos básicos de vocabulario sobre conceptos elevados. De las entradas de las enciclopedias se pueden sacar definiciones de los conceptos. Por último, a partir de textos descriptivos de redacción telegráfica se pueden obtener actualizaciones y añadidos para todas las fuentes de referencia anteriores, modificando así los

valores cambiantes en anuarios y enciclopedias, al mismo tiempo que se aumenta el lexicón de los diccionarios.

Las técnicas de extracción de información de las fuentes textuales incluyen los tratamientos convencionales de textos y técnicas informáticas más avanzadas de comprensión lingüística de textos. Dado que es imposible diseñar analizadores generales para la entrada ilimitada de texto, es relativamente fácil hacerlo con analizadores expertos de textos en determinados modelos sintácticos. Tales herramientas reconocen fácilmente determinadas clases de información en el texto, y con el tiempo, y a través del procesado de millones de palabras-texto, se puede derivar una cantidad considerable de información cultural. Estos expertos en textos se han diseñado para obtener datos geográficos y biográficos, para el reconocimiento de neologismos, y por lo general como apoyo para la recogida de nombres propios y compuestos que hay que añadir a una base léxica de conocimiento.

Investigación psicolingüística

La palabra es una unidad muy importante en el sistema del lenguaje natural. No solamente las palabras son unidades que portan significado, sino que también funcionan como interfaz entre nuestro sistema cognitivo y el mundo. Por esta y otras razones, las palabras desempeñan un papel esencial en la investigacion lingüística.

Las palabras se utilizan en la investigación psicolingüística como unidades de investigación en sí mismas, o como subunidades dentro de unidades mayores de interés. Si se estudian ciertas propiedades de estas últimas (las subunidades), entonces se definirán clases diferentes de unidades, reflejando cada una de forma diferente la propiedad que se estudia.

La lógica de la investigación experimental determina, por tanto, que estos tipos diferentes de unidades no tengan que diferir en ningún otro aspecto.

Por tanto, es necesario comparar estas clases en muchas de las propiedades que tiene la palabra y que pueden desempeñar un papel en ese experimento determinado.

La razón para la necesidad de conocer las propiedades de las palabras es doble: una es el interés por las propiedades en sí, y la otra tiene que ver con los principios de diseño experimental.

Una base de datos léxicos es extremadamente útil en psicolingüística experimental como modo de acceso a las muchas propiedades de las palabras de una lengua. La estructura e información de un BDL (propiedades de la palabra tales como categoría sintáctica, composición ortográfica, estructura morfológica, composición fonológica, estructura silábica, rasgos de acento, contorno de la palabra, frecuencia) tienen que investigarse en función de las consultas que debe formular un psicolingüista experimental a una base de datos léxicos.

Las aplicaciones prácticas (por ejemplo, en el aprendizaje de idiomas) deben ser sometidas a debate.

En la consecución de objetivos psicolingüísticos, algunos investigadores han usado BDL para comprobar la plausibilidad y aplicabilidad de algunos conceptos en las teorías de reconocimiento auditivo de la palabra y, en general, para responder a ciertas preguntas cruciales sobre el proceso del lenguaje humano.

De una vez por todas, no se puede facilitar una «lista de la compra». Hay, sin duda, propiedades esenciales que una BDL debería poseer. Son bastante comunes a las exigidas por lingüistas y lingüistas informáticos computacionales. Particularmente, los investigadores en psicolingüística piden ya recuentos adecuados y periódicamente actualizados de frecuencia.

Sistemas de diccionario para el mercado de masas

El mercado de diccionarios de lectura automatizada (DLA) abarcaría tres tipos principales de aplicación:

— Donde se utiliza el DLA como herramienta en el tratamiento de textos.
— Como herramienta de trabajo intelectual en la investigación literaria, lingüística o cualquier otra académica.
— Referencias generales —una alternativa más flexible al uso del diccionario impreso.

Los DLA en la oficina: Dos tendencias parecen influir en el uso de los DLA en la oficina: la primera parece ser el trasvase del trabajo de escribir a máquina, desde auxiliares y oficinistas a jefes y ejecutivos. Es el resultado natural del número de estaciones de trabajo que hay en mesas de ejecutivos, y el desarrollo de redes de trabajo. El ejecutivo que tiene que mecanografiar el correo electrónico va a necesitar más que un procesador de textos, quizá algo parecido al Proyecto Epístola de IBM. La segunda tendencia —y relacionada con la anterior— es el avance espectacular de la tecnología del ordenador, lo que significa que la capacidad de procesamiento y de almacenamiento no constituye un factor de limitación.

Éste será quizá el primer mercado para los DLA. Sin embargo, las exigencias de utilización puede que no sean homogéneas, al menos inicialmente. Sin duda alguna, habrá un desarrollo de estos mercados de exigencias muy específicas y de requisitos especializados. Pero esto será ya en la segunda etapa. Lo que se necesitará inicialmente será un DLA que ofrezca una gama completa de información léxica dotado de software que permita al usuario conseguir el tipo adecuado y el nivel de información con pocas dificultades.

El hecho de que un DLA contenga cantidades de datos innecesarios para el usuario no será una barrera de la manera que lo sería un diccionario convencional.

Sin duda alguna será preciso adaptar tanto el contenido como el software a las necesidades del ejecutivo. Los tipos de información que sin duda se necesitarán serán, por ejemplo, vocabulario especializado relacionado con los campos técnicos/profesionales, nuevos términos y acceso fácil a un tesoro.

En la casa: El ritmo acelerado del cambio tecnológico es posible que fomente también un mercado de DLA de uso doméstico.

Como consecuencia de la gran capacidad cada vez mayor de almacenaje y procesamiento, es probable que la informática doméstica ofrezca una completa gama de funciones que incluya no sólo las aplicaciones individuales (como juegos y tratamiento de textos), sino también medios de comunicación (como el correo electrónico y la banca desde casa) y de información (bolsa de valores e información enciclopédica y léxica).

Estos medios serán no sólo un instrumento recreativo, como lo ha sido hasta ahora el ordenador doméstico. También será una herramienta para la escritura creativa, para el diseño, para la composición musical y literaria, etc. El valor extra que se añade por la disponibilidad de información léxica puede justificar en muchos casos el desembolso exigido para la compra de un sistema de este tipo.

En la enseñanza: Este sector, sobre todo la utilización del ordenador para un mejor y más conveniente uso del material de consulta, se está empezando a estudiar.

Ahora se comienza a pensar en la utilización de DLA en la adquisición de un segundo idioma para perfeccionamiento de la expresión escrita. Cito de un artículo de Benbow. El desarrollo de un mercado para el nativo de lengua inglesa estará, como en otros campos, dictado por la disponibilidad de una tecnología apropiada.

Se pueden prever importantes y amplios usos de los DLA (o una serie determinada de ellos) en centros de enseñanza y formación, en todos los niveles educativos. No es probable, sin embargo, que la disponibilidad sólo de los DLA sea suficiente para generar un mercado importante.

Si los centros educativos y de formación están utilizando equipos sobre los que se podría instalar los DLA, por un modesto desembolso para instalar DLA (o una serie de ellos) se podría dotar al sistema de valor significativo.

En cuanto a la enseñanza del inglés como lengua extranjera, se podría ver más fácilmente el desarrollo de un mercado autónomo de DLA. La enseñanza de la lengua inglesa siempre ha sido rápida en adoptar la tecnología apropiada, como el uso del magnetófono y el casete en los laboratorios de idiomas y el desarrollo del vídeo como instrumento de enseñanza.

Obviamente, el tipo de información requerida por un hablante de lengua inglesa no nativo tendrá que diferir bastante del que necesita el nativo: la información sobre la tipología verbal y el uso contextual serán prioritarios, por citar un caso. Por tanto, como este tipo de información aparece ya en diccionarios de enseñanza de lengua inglesa, este requisito de especialización no supondrá un problema.

Consecuencias para la lexicografía

1. Los usos de un DLA serán diferentes de los de un diccionario impreso. Una encuesta reciente del OED (Old English Dictionary) indica que esto será lo que ocurrirá por lo menos en lo que se refiere a este diccionario. La diferencia en cuanto al uso, sin embargo, no estriba simplemente en diferentes medios de información sobre los que el texto se apoya. El usuario de DLA en muchos casos estará buscando información no fácilmente conseguible en un diccionario impreso, y las vías de acceso a la información serán infinitamente más variadas que las que puede ofrecer un diccionario impreso. Un ejemplo fundamental sería una consulta a la inversa (hallar un término a partir de su significado) y otras exploraciones semánticas.

2. Surgirán nuevos tipos de usuarios al disponer de DLA. El uso del diccionario impreso se limita en la práctica a conseguir información sobre lemas conocidos. Con los DLA será posible conseguir información sobre grupos de palabras relacionados por uno o varios factores; por clasificación temática, clase gramatical, pronunciación, lengua de origen, etc. La lista no tiene límites. Por ello, las implicaciones completas son imposibles de predecir.

3. Los diccionarios impresos y sus homólogos de lectura automatizada tomarán rumbos opuestos. Los usos varios, los grupos diferentes de usuarios, la distinta naturaleza harán que la correspondencia inicial entre diccionarios impresos y sus homólogos de lectura automatizada desaparezca pronto. Si, como parece probable, los DLA proliferan, surgirá el problema de mantener los normales. La cantidad de material puede que sustituya a la calidad como principal preocupación de los fabricantes de DLA; el planteamiento de un equilibrio aceptable debería ser la preocupación principal de todos aquellos que trabajen en el desarrollo de DLA.

4. El software será una clave determinante en la utilidad de un DLA. Se necesitará mucha investigación y desarrollo para proporcionar un software lo suficientemente aceptable que haga que los DLA sean realmente valiosos.

5. La influencia del lexicógrafo puede afectar al desarrollo de un DLA. En la preparación de un diccionario impreso el papel

que desempeña un lexicógrafo es capital. Puede que no sea ése el caso en el desarrollo de los DLA. La influencia del usuario, del especialista en ciencias de la información y del especialista en marketing puede que sea significativamente más importante en la fabricación de un DLA. Esto nos lleva al problema de la calidad: el lexicógrafo es el factor más importante para el control de calidad. Si su papel queda degradado, la calidad del producto estará en peligro.

6. Los DLA no podrán reemplazar por completo a los diccionarios. No debe pasarse por alto el hecho de que para las cuestiones más simples (por ejemplo, el significado, grafía, pronunciación de una palabra determinada) el diccionario impreso constituye la mejor y más conveniente manera de conseguir la información solicitada (a menos que, evidentemente, en el momento que se necesite la información se esté trabajando en una estación de trabajo con acceso a un DLA).

7. Es posible que para muchos usuarios de diccionario la existencia de grabados e ilustraciones sea un buen complemento. Lo mismo puede decirse del vídeo y del material audiovisual que puede aparecer como material de ayuda de las definiciones verbales. Ciertas características técnicas adicionales —como la reproducción sintética de la pronunciación de una forma determinada— aumentarán el valor del DLA.

Reutilización de diccionarios de lectura automatizada

El número de diccionarios de lectura automatizada aumenta debido a la difusión de la fotocomposición. La posibilidad que ofrecen las modernas tecnologías en la distribución de productos lexicográficos tradicionales o nuevos constituirá otro factor para disponer cada vez más de recursos léxicos en forma de lectura automatizada.

Sin duda, los diccionarios tradicionales no contienen toda la información necesaria para los sistemas de procesamiento del lenguaje natural.

Por otra parte, poseen, sin embargo, bastante información más

o menos organizada y pertinente con respecto a los sistemas computacionales. No obstante, parece conveniente evaluar, en el nivel metodológico y de organización, las posibilidades existentes para emplear la información disponible en DLA en diferentes aplicaciones informáticas.

La reorganización de DLA en forma de bases de datos con una metodología propia de integración puede que abra, al usuario especialista común, el acceso a información alfabetizada de recuperación difícil.

Los DLA y la investigación en lingüística computacional

Considerando el gran número de sistemas que hay dentro del paradigma de la lingüística computacional, sorprende que exista un número tan reducido de entradas de diccionario disponibles en estos sistemas.

Hay que admitir que la mayor parte de estos sistemas son experimentales; sin embargo, para las aplicaciones con palabras reales, teniendo en cuenta los avances recientes en la tecnología computacional que hacen que estas aplicaciones sean factibles, se necesitan vocabularios de mucha mayor extensión. Dado el número de cortapisas reales que encierran muchos de los sistemas, siendo los más críticos las formalizaciones idiosincráticas para redactar entradas de diccionario, parecen ser un reto, y de hecho lo son, en la labor de crear un vocabulario amplio y sólido para cualquier aplicación, los diferentes formatos de lexicón, y las diferentes consideraciones sobre lo que constituye la información lingüísticamente relevante (sintáctica, semántica y pragmática), cuyo lugar apropiado es el lexicón. Y apenas sorprende, por tanto, que un número de investigadores esté tras la consecución de diccionarios de lectura automatizada (DLA). Esperan que la información ya ordenada, categorizada, indexada y, sobre todo, disponible en forma de lectura automatizada, sea convenientemente utilizada, si no para tener un lexicón considerable «de la nada», sí por lo menos para crear automáticamente una parte sustancial del mismo, consiguiendo de la nada lo que se tiene la esperanza que sea

un objeto internamente consistente (y coherente). Un lexicón logrado de esta forma ahorrará esfuerzos y creará el volumen de vocabulario terminal que podría ser por consiguiente ampliado, si fuera necesario, y adaptado a las tareas y aplicaciones pertinentes.

Existe un trabajo paralelo afín, en línea con la corriente actual de «programación basada en el conocimiento», que plantea el problema de la adquisición de segmentos fundamentales de conocimiento estructurado sobre la propia palabra. La esperanza gira aquí de manera similar en torno a la creencia de que se pueden encontrar caminos para localizar y extraer parte del conocimiento requerido para realizar un tipo de funciones (semi)inteligentes de fuentes de lectura automatizada, por ejemplo, de diccionarios y enciclopedias.

Será útil analizar los caminos que recorren muchos investigadores en busca de esos fines, usando diccionarios diferentes y utilizando técnicas diversas en apoyo de aplicaciones varias. Se pondrá énfasis especial en determinar la cantidad de información disponible en forma de lectura automatizada y cuánta podrá utilizarse en la investigación lingüística computacional.

Dado que es cierto que los sistemas difieren en organización, estructura y contenido de sus lexicones, aún es posible aislar ciertos tipos de información, determinados por la función y las aplicaciones particulares, que deberían estar disponibles a nivel léxico. Una categorización de los requisitos léxicos, a medida que varían en una gama de aplicaciones dentro del mismo marco de la lingüística computacional, aportará una nueva dimensión para evaluar la utilidad de la información léxica que pueda conseguirse en forma de lectura mecanizada.

Una fuente de lectura automatizada puede ser percibida como una base de datos léxicos que ofrece información, por ejemplo, sobre los núcleos silábicos de una palabra o sobre un comportamiento sintáctico idiosincrático. Dicha información puede que sea esencial para un sistema de síntesis del lenguaje o para un programa de análisis, pero es de poca utilidad, por ejemplo, para un componente de interpretación de un sistema finito de lenguaje natural. En el último caso, la intención es la de considerar el diccionario como una especie de base de conocimiento, donde la información de naturaleza semántica más general estaría codificada, o en un formato de texto libre, o en alguna forma de etiqueta-

ción semántica. De interés general en ambos casos son cuestiones tales como los tipos de datos que estarán disponibles en forma de lectura mecanizada, y el tipo de sistema computacional a los que le puedan ser útiles los datos.

Una gran parte del trabajo se ha empleado en intentar dar soluciones computacionales a un gran número de problemas, que radican en la información extraíble de un DLA. El análisis sintáctico, el desarrollo gramatical, la selección de sentido de la palabra, la síntesis del lenguaje, la sólida interpretación textual, la adquisición de conocimientos, la organización de la información, el acceso léxico con ayuda fonética, el análisis de la frase, son sólo unas cuantas de las funciones que se han favorecido sustancialmente, ya que se disponía de grandes diccionarios *on-line*.

Una experiencia colectiva hace factible, y por supuesto muy necesario, examinar aspectos más generales del trabajo con diccionarios de lectura automatizada. Ahora ya es posible preguntarse por cuestiones tales como ¿cuál es el costo total (el esfuerzo de mano de obra dedicado al proyecto) de intentar enjaezar lo que con frecuencia es un objetivo tosco y voluminoso? ¿Qué tipo de información se puede esperar razonablemente de un DLA? ¿Hasta dónde es fiable esta información? ¿Cuál sería la mejor manera de utilizarlo en un contexto determinado? El consenso sobre estos temas, junto a la opinión sobre el puesto que debe ocupar en esa escala la gama de diccionarios disponibles, será de gran utilidad para la comunidad investigadora.

Estructura y acceso en los lexicones automatizados

El problema de la estructura y acceso en relación con la organización de un lexicón automatizado se puede ver desde varias perspectivas:

— Como problemas de software y de hardware.
— Como problemas de estructura conceptual o lógica de los datos, es decir, en una base de datos o en una red.
— Como problemas de estructura «léxica» y de relaciones

entre las palabras y, por tanto, de la conexión y concepto lingüísticos.
— Como problemas de naturaleza más psicolingüística, si se estimula el proceso mental de acceso léxico.
— Como problemas relacionados con los fines perseguidos.
— Como problema relacionado con los usuarios (personas o procesos) de los datos.

El tipo de organización y el consiguiente acceso que se requieren son muy diferentes si tenemos que crear un diccionario de simple lectura automatizada, o una base de datos «estructurada» o incluso una base de conocimiento más compleja. En el primer caso, sin duda es suficiente una «cadena» o «formato de textos», mientras que en los casos restantes es posible que una organización más estructurada/codificada proporcione los mejores resultados.

La reestructuración de los datos de un DLA organizado en forma de base de datos puede que haga posible lograr información adicional sobre los datos originales, multiplicados por un factor dado por el número de perspectivas diferentes que se establezcan en los datos, por ejemplo, de las diferentes relaciones de ordenación o de los varios caminos que conecten los datos.

El lexicón aparecerá de hecho dividido en tantos subgrupos como relaciones hayan sido determinadas y formalizadas. Al representar el lexicón por el conjunto de esas relaciones, podemos tener acceso al diccionario a través de las unidades léxicas, o a través de rasgos, o de relaciones, o podemos buscar la red para comprobar dónde se empareja con la pregunta formulada, y recoge las diferentes partes del contenido léxico en relación tanto con el punto de entrada como de las opciones activadas en ese punto.

El concepto del lexicón en sí mismo puede ser ampliamente tratado y cambiado en el marco informático, sobre todo si se considera como un objeto mucho más interrelacionado, donde las palabras están compuestas en muchos niveles, y cada tipo de complejidad tiene que ser analizado y tratado, de manera que cada una de estas complejidades en varios estratos pueda ser captada o por reglas generales o por funciones apropiadas de acceso.