EPOS, XV (1999), págs. 425-434

HACIA UN MODELO DE FORMALIZACIÓN DEL CONOCIMIENTO LÉXICO CON FINES INFORMÁTICOS

ELENA BÁRCENA & TIM READ
UNED

RESUMEN

En este artículo se presenta un tipo de sistema de formalización del conocimiento léxico que ha sido ideado para compensar algunas de las carencias de los modelos lexicográficos más generalizados. El artículo comienza con una reflexión sobre los requisitos de forma y contenido que debe satisfacer un léxico general destinado a fines informáticos. Dichos requisitos han sido satisfechos por el modelo de diccionario lexemático-funcional originalmente ideado por Martín Mingorance a partir de las teorías de Coseriu y Dik, habiendo sido desarrollado y adaptado para su implementación computacional. Para ilustrar la arquitectura de este sistema, se presenta un prototipo que ha sido elaborado para los verbos del inglés y el español.

INTRODUCCIÓN

El artículo presenta una reflexión sobre los requisitos de forma y contenido que debe satisfacer un diccionario del léxico general destinado a fines

informáticos. Dichos requisitos han sido satisfechos por el modelo de diccionario lexemático-funcional (MLF) originalmente ideado por Martín Mingorance (1984) a partir de las teorías de Coseriu (1977) y Dik (1978)[1]. Dicho modelo ha sido desarrollado y adaptado por este equipo de investigación para su implementación computacional.

Entre las características del diseño del diccionario cabría destacar la completa información sintagmática contenida dentro de cada entrada léxica y su económica distribución, así como la organización jerárquica y onomasiológica del conjunto del diccionario, la cual está basada en la estructura léxico-conceptual de la lengua. Además, dicha estructura ha sido establecida empíricamente a partir de una selección de diccionarios y corpora monolingües ya existentes.

La versión automática del diccionario está dotada de una interfaz sofisticada, flexible, portátil e independiente de la base de datos. Esta interfaz está escrita en Java, lo cual permite el acceso al diccionario desde el mismo ordenador, en una red local o en la Internet por la World Wide Web (WWW). Para ilustrar la arquitectura de esta herramienta léxica se presenta un prototipo que ha sido elaborado por los autores para los verbos del inglés y el español.

En este artículo se sugieren futuras aplicaciones contrastivas y informáticas del diccionario gracias en parte a la profundidad y el rigor de su modelo descriptivo. El diseño composicional de las descripciones a partir de rasgos semánticos cuasi-elementales permite evitar la circularidad, revelar generalizaciones significativas a nivel intra- e interlingüístico y desarrollar un sistema automático de consulta léxica mono-, bi- y también multilingüe. Se prevé que de las definiciones de la base de datos léxica pueda ser extraída una interlengua léxica, lo cual permitiría la fácil incorporación de nuevos idiomas a la base de datos léxica, así como la traducción multidireccional.

Otros modelos de formalización del conocimiento léxico

El tipo estándar de organización de diccionario (monolingüe y bilingüe) es semasiológico (i.e., sigue un criterio formal), siendo el más co-

[1] Este trabajo se ha elaborado dentro del marco del proyecto «Desarrollo de una lógica léxica para la traducción asistida por ordenador a partir de una base de datos léxica inglés-español-francés-alemán multifuncional y reutilizable» (PB94-0437): subvencionado por el Ministerio de Educación y Ciencia. Cualquier error que se encuentre en este artículo es, sin embargo, responsabilidad exclusiva de los autores. Este artículo es una traducción resumida, adaptada y revisada de "MoLeX: Tomorrow's computational dictionary today", escrito por los mismos autores y publicado en 1997 en *ATLANTIS*, Vol. XIX, n.° 1, págs. 49-58.

mún, sin duda, el orden alfabético. Sin embargo, en estos diccionarios se pierde importante información sobre la organización global del vocabulario de las lenguas. Este hecho y el objetivo de producir un diccionario que fuera adecuado desde una perspectiva psicológica llevaron a algunos lexicógrafos a desarrollar fuentes de consulta léxicas organizadas onomasiológicamente (i.e., de acuerdo con el significado conceptual), siendo las más comunes los tesauros. Los tesauros, sin embargo, conllevan gran redundancia y proporcionan escasa o ninguna información sobre las propiedades combinatorias y selectivas de las palabras. Ciertamente, su organización está motivada psicológicamente hasta cierto punto (Roget, 1987), pero su estructura de arriba abajo (i.e., de general a particular) se considera bastante *ad hoc*.

Según Martín Mingorance, un diccionario debería reflejar la organización del lexicón mental humano y explicar el modo en que los idiomas individuales lexicalizan el conocimiento conceptual de sus hablantes. Muchos psicolingüistas como Apresjan (1993) mantienen que hay evidencia empírica sobre la naturaleza de las relaciones semánticas, que no son arbitrarias sino funciones mentales reales, y sobre la estructura jerárquica de la mente humana, donde las palabras que están relacionadas semánticamente entre sí se encuentran almacenadas cerca las unas de las otras (ver también Caramazza, 1996). Partícipe de esta línea de investigación, Martín Mingorance esbozó «de abajo arriba» el diseño de un lexicón relacional basado en la estructura del significado, donde los lexemas están organizados jerárquicamente entre sí dependiendo de la presencia de componentes semánticos comunes y diferenciales. Dicho léxico relacional capta los matices sutiles del significado de cada palabra y la distingue incluso de aquéllas con las que guarda una mayor proximidad semántica.

Martín Mingorance creía que la competencia lingüística depende en gran medida de la información obtenida del lexicón mental, lo cual implica la existencia de un conocimiento lingüístico complejo (e.g., morfosintáctico, semántico, pragmático) unido a cada elemento léxico. Motivado por este hecho, llevó a cabo un estudio del tipo de información paradigmática y sintagmática que sería necesario incluir en las entradas de un diccionario para explicar la producción y comprensión lingüísticas con éxito, y cómo hacerlo del modo más económico posible. Martín Mingorance incorporó aspectos cognitivos como estos en la elaboración de su teoría léxica que denominó MLF. Al margen de su motivación cognitiva, la riqueza descriptiva que se alcanza con este modelo sobre el uso y la combinabilidad de las palabras en la oración hace que sea adecuado para diversas aplicaciones informáticas.

El Modelo Lexemático-Funcional

El MLF está basado en la integración que realizó Martín Mingorance de dos teorías lingüísticas: la Gramática Funcional de Dik (GF) y la Teoría Lexemática de Coseriu, con el fin de obtener un eje sintagmático y otro paradigmático que expliquen la combinación y la selección léxicas respectivamente. El eje sintagmático basado en la GF utiliza marcos predicativos como fórmulas integradas que especifican los patrones de combinación que gobiernan los predicados de una lengua dada. En el eje paradigmático, los lexemas están organizados onomasiológicamente en campos semánticos siguiendo los dictados de la Teoría Lexemática. Microestructuralmente, las entradas léxicas están caracterizadas como unidades complejas de información sintáctica, semántica y pragmática. A nivel macroestructural, las entradas léxicas están interconectadas semánticamente de diversos modos formando jerarquías complejas. El modelo resultante integra una red jerárquica de lexemas relacionados semánticamente dentro de áreas generales de significado, con definiciones altamente informativas y económicas.

La GF y La Teoría Lexemática no son sólo complementarias sino compatibles desde perspectivas teóricas y prácticas. Como dice Dik (1978, p.46): «Aunque la idea del análisis léxico encaja bien con el modelo de la GF, las presuposiciones que conlleva no se siguen necesariamente de este modelo. Es decir, la GF también sería compatible con otras concepciones de definición del significado léxico» (*nuestra traducción*). Por ejemplo, la GF es funcional desde un punto de vista teleológico porque concibe la lengua como un instrumento de interacción verbal, no como un sistema abstracto e independiente de su uso real. Por su parte, la Teoría Lexemática es funcional desde una perspectiva estructural en el sentido de que en ella la lengua aparece determinada por un sistema de oposiciones funcionales. Además, los principios de la GF subyacentes a la estructura de las definiciones de significado y el proceso de descomposición léxica gradual para las definiciones tienen muchos puntos de convergencia con los principios de la teoría del campo léxico y la factorización de la Teoría Lexemática, tal y como aparece subsumido en el MLF.

La GF fue adoptada por Martín Mingorance por varias razones, incluyendo su adecuación psicológica y su integración en una teoría de la comunicación verbal. La GF está dirigida por el lexicón y da cuenta de la combinabilidad de los elementos léxicos en la secuencia linear, expresada en términos funcionales y abstractos. Las unidades léxicas se conciben como representaciones estructuradas y codificadas en forma de marcos predicativos. Las definiciones de significado siguen la estrategia conocida como descomposición léxica gradual

y, de modo similar a las predicaciones subyacentes oracionales, están expresadas en términos de predicados existentes en la lengua correspondiente, aunque la diversidad y el orden de dichos predicados están parcialmente restringidos para poder profundizar en la dimensión semántica de los vocabularios de las lenguas. Debido a la complejidad de idear un conjunto finito de unidades universales de descripción semántica (e.g., rasgos binarios, primitivos abstractos, predicados atómicos), tal y como han manifestado repetidas veces numerosos investigadores, los componentes de significado empleados en la GF son sintagmas de la lengua natural, aunque hay diversas restricciones en su expresión. Apresjan (1993, p.86) observa a este respecto: «En la mayoría de las lenguas hay primitivos semánticos, o más bien, *cuasi primitivos*, para describir el concepto básico de cada lexema. Los predicados que aparecen en las definiciones de significado son elementos léxicos de la lengua objeto» (*nuestra traducción y nuestro énfasis*).

Debería mencionarse que en el MLF hay un refinamiento del contenido de los marcos predicativos tal y como son tratados por la GF. Por ejemplo, para cada predicado dado hay una lista de uno o más marcos predicativos preferidos, lo cual justifica la flexibilidad sintáctica y evita la rígida y problemática distinción entre argumentos (términos obligatorios) y satélites (términos opcionales) en la GF. Además, la descriptividad de los marcos predicativos está enriquecida con información adicional; por ejemplo, la descripción sintagmática de los verbos incluye información colocacional y funciones sintácticas, las cuales no están presentes en la GF estándar.

Martín Mingorance incorporó la Teoría Lexemática en su modelo para ampliar el conocimiento del lexicón más allá del nivel de las entradas individuales, reflejando la estructura relacional global de éste y así construir el diccionario de una lengua como una red jerárquica de lexemas interconectados semánticamente.

Con este modelo relacional es posible captar las estructuras jerárquicas semánticas de los campos léxicos completos que forman los vocabularios de las lenguas, las relaciones (hiponimia, sinonimia y antonimia) entre sus elementos y los sutiles matices semánticos de los mismos. Además, siguiendo el método composicional de definición y el principio de la herencia de rasgos de ciertas jerarquías, los hipónimos de un archilexema dado se convierten en el *definiens* de otras palabras en niveles más específicos o inferiores de la jerarquía, y así sucesivamente. Los componentes cuasi-primitivos que forman las definiciones de significado son heredados por todos los lexemas que se encuentran más abajo en la jerarquía, de tal modo que cada definición de un lexema dado está claramente expresada en términos tan simples como la combinación de la definición del elemento inmediatamente superior más el

significado nuclear propio del lexema a través de uno o más componentes semánticos. Este económico método de definición evita la circularidad y revela generalizaciones significativas sobre la organización semántica estructural del lexicón de una lengua.

El desarrollo de un marco que organice y describa predicados siguiendo principios lexemáticos conduce a un nivel de comprensión profundo de las lenguas individuales y de la compleja relación entre la sintaxis y la semántica. En cuanto a la estructura semántica de los vocabularios, por ejemplo, se han establecido una serie de principios sobre la recurrencia de las dimensiones a través de los campos léxicos. Además, varios aspectos sintácticos de las palabras están fuertemente ligados a su significado. La estrecha relación entre la sintaxis y la semántica ha sido observada por un gran número de autores como Dixon (1991, p.75): «Las palabras léxicas de una lengua pueden ser agrupadas en una serie de tipos semánticos, cada uno de los cuales tiene un componente de significado común y un conjunto típico de propiedades gramaticales» (*nuestra traducción*). Muchas de estas propiedades, como los clasemas, son recurrentes en otros dominios y campos léxicos. Martín Mingorance desarrolló la noción de esquema predicativo como un tipo de marco predicativo ampliado que capta la información sintáctica, semántica y pragmática compartida por todos los lexemas de cada dimensión. Esto es, una vez más, crucial para organizar las entradas de diccionario de un modo más económico, ya que sólo la nueva información sintagmática necesita ser explicitada para cada predicado que se halle en una posición inferior en la jerarquía, así como para captar generalizaciones sintácticas y semánticas significativas entre elementos léxicos relacionados semánticamente.

La versión automática

Consultar sistemas de referencia léxica basados en papel es a menudo una tarea lenta y tediosa. Durante mucho tiempo ha habido gran interés por parte de los lexicógrafos y de los usuarios en hacer el proceso de consulta léxica más rápido y grato. No requirió mucha imaginación darse cuenta del papel que los ordenadores podrían jugar en este proceso (no sólo en la provisión de sistemas de referencia léxica, sino también, por ejemplo, de procesadores textuales y correctores ortográficos y gramaticales).

Los sistemas automáticos lexicográficos tienen muchas ventajas sobre las versiones en papel, incluyendo la búsqueda rápida y flexible, ya que se puede acceder a los datos de los diccionarios automáticos a través de palabras clave (o partes de ellas), además de las entradas mismas, o incluso a través de com-

binaciones de palabras clave. Cabe destacar también la gran capacidad de almacenamiento, ya que a medida que la memoria y los tamaños de los discos duros aumentan y los medios de almacenamiento óptico (como los discos compactos) se hacen disponibles, los léxicos *on-line* (en línea) ofrecen una cobertura lingüística cada vez mayor que sus correspondientes versiones en papel.

Por otra parte, los diccionarios de papel son normalmente monolingües o bilingües o, en el mejor de los casos, cubren un reducido número de lenguas. Las fuentes *on-line*, sin embargo, ofrecen múltiples combinaciones de idiomas. Los diccionarios automáticos pueden también usarse solos o como parte de entornos integrados, tales como la llamada «mesa de trabajo del traductor» (*translator's workbench*), en la que todas las herramientas se encuentran disponibles y listas para su utilización. Además, a los diccionarios automáticos se les puede añadir fácilmente nuevas entradas, como parte del diccionario principal o como sub-diccionarios especializados separados. Por último, en un contexto de investigación, las herramientas informáticas proporcionan un modo práctico de investigar las lenguas y la funcionalidad e implicaciones de las teorías lingüísticas.

Además, la lexicografía computacional ha encontrado muy recientemente un entorno de uso y un mercado amplios y dinámicos en la Internet. La evolución de la Internet ha sido muy rápida: comenzó en Estados Unidos en 1969 para facilitar la investigación en materia de defensa. En 1983 había quinientos sistemas conectados y a finales de 1994, más de cuatro millones. El diseño de un nuevo protocolo: HTTP (HyperText Transfer Protocol) por parte del laboratorio de investigación CERN en 1990 mejoró notablemente la cantidad y el tipo de información que se transfería, a través del lenguaje HTML (HyperText Markup Language). Esto marcó el comienzo de la WWW, que ha hecho aumentar en gran medida el uso de la Internet. La WWW facilita, además, el diseño de diccionarios on-line que pueden ser consultados desde ordenadores localizados en distintos países.

Con la Internet multitud de personas pueden acceder simultáneamente a un mismo diccionario, aunque su uso se ve restringido por el número relativamente reducido de conexiones privadas con la Internet actualmente, lo cual está cambiando a medida que las nuevas tecnologías causan una caída gradual del costo de las comunicaciones. Otra característica de los diccionarios en la Internet es que es más fácil modificarlos al haber sólo una copia en el servidor y no una copia en cada ordenador. Siguiendo el mismo argumento, no es necesario emplear tanto espacio del disco duro de cada ordenador. Los sistemas de referencia léxica en la Internet puede que no sean demasiado atractivos en los lugares donde las conexiones son todavía lentas, pero esto también está cambiando.

Un problema más grave de estos diccionarios y otras herramientas de Procesamiento del Lenguaje Natural (PLN) hasta hace poco tiempo era la falta de

sofisticación de los tipos de interfaz disponibles. HTML puede ser lo suficientemente sofisticado para construir diccionarios on-line, pero no proporciona el tipo de funcionalidad a la que están acostumbrados los usuarios en los diccionarios de sus ordenadores personales, como la habilidad para buscar entradas (Heslop & Budnick, 1996). Mientras que la llegada del CGI (Common Gateway Interface) ha permitido a los documentos de HTML acceder a programas basados en un servidor que ofrecen funciones de búsqueda básicas para documentos HTML, no ha proporcionado la suficiente funcionalidad para construir herramientas de PLN on-line sofisticadas.

Java lo ha hecho. Desarrollado por Sun Microsystems, Java fue presentado oficialmente en 1995 como un lenguaje de programación sencillo, robusto, dinámico, *multi-threaded*, orientado a objetos y que funciona independientemente del hardware o sistema operativo debido a su máquina virtual (Naughton, 1996). Los programas escritos en Java pueden funcionar en la WWW y ofrecen la misma funcionalidad que los que están localizados en el disco duro del ordenador del usuario. Como consecuencia de características como éstas, Java es un lenguaje de programación muy útil para herramientas de PLN basadas en la WWW.

Todas estas consideraciones fueron tomadas para el diseño de la versión automática del diccionario basado en el MLF. Como es corriente hoy en materia de Lexicografía y en general en la Lingüística, se ha desarrollado un prototipo computacional para probar y evaluar el MLF. Además, su construcción ha de verse como un paso hacia la implementación del diccionario completo, que está siendo llevada a cabo actualmente. Por lo tanto, una vez que el MLF fue diseñado y el diccionario bilingüe construido, el próximo paso fue el desarrollo de un prototipo computacional para los verbos en inglés y español.

Como la mayoría de los programas con una interfaz visual, el prototipo desarrollado es un sistema conducido por eventos, por lo que, después de una fase inicial de arranque, la acción del usuario determina la información que se presenta en la pantalla. La interfaz desarrollada permite al usuario consultar las distintas acepciones de una palabra dada. Hay dos posibles tipos de ambigüedad: ambigüedad en la lengua fuente o en la lengua meta. En el primer tipo, una misma palabra aparece bajo más de una dimensión (o incluso en la misma dimensión). Entonces, se le ofrecen al usuario todas las posibles interpretaciones de la palabra ambigua (distinguidas entre sí por índices) por medio de las distintas definiciones y se le pide que seleccione la interpretación que prefiere. En el caso de la ambigüedad en la lengua meta, una palabra con una sola acepción aparece más de una vez debido a que equivale a más de una palabra en la lengua meta. Las distintas ocurrencias de esta palabra pueden distinguirse a través de las dimensiones correspondientes o simplemente con

el equivalente en otra lengua. Conviene observar que las características del formato de la base de datos proporcionan información relacionada con el modelo descriptivo. Por ejemplo, las relaciones de hiponimia existentes entre las distintas dimensiones y entre los lexemas de cada dimensión se representan visualmente por medio de una sangría izquierda.

Conclusión

Este artículo ha resumido los rasgos fundamentales de un sistema de referencia léxica *on-line* para la descripción del léxico básico de las lenguas naturales destinado principalmente a fines informáticos, que sigue un nuevo modelo de diccionario basado en el MLF. Como hemos visto, el MLF es una teoría con sólidos fundamentos cognitivos en la que el diccionario se considerá como la representación del lexicón mental del propio hablante. El MLF ha sido diseñado para construir una descripción onomasiológica del vocabulario de una lengua, con los predicados agrupados en clases o dominios semánticos, y una descripción completa de sus predicados. El MLF integra la GF y la Teoría Lexemática para dar cuenta de las relaciones sintagmáticas y paradigmáticas que existen en el lexicón, las cuales están basadas en los principios complementarios de combinación y selección presentes en las definiciones. El contenido del lexicón está ampliado más allá del nivel de las entradas individuales para reflejar su estructura relacional global, y de este modo diseñar el lexicón como una red jerárquica de lexemas conectados semánticamente. Esta red es dinámica y abierta a la incorporación progresiva de más entradas.

El prototipo informático ha sido diseñado en dos partes: la base de datos de conocimiento léxico y una interfaz escrita en Java. La organización de la información en la base de datos permite a los lingüistas que trabajan en el prototipo acceder y modificar dicha información fácilmente. Como las tablas y campos en la base de datos están organizados onomasiológicamente, es también fácil trabajar en sus contenidos sin tener que tratar con estructuras de datos complejas dictadas por el programa principal (cf. los archivos de datos que a menudo están escritos en Prolog [Dik, 1992]).

La interfaz del lexicón existe como un documento de HTML y contiene un *applet* de Java que trata la manipulación de las lenguas de la base de datos. Mientras que la interfaz depende en gran medida del explorador empleado (e.g., Netscape), su aspecto es muy similar al de las páginas de la WWW, independientemente de si el lexicón funciona en el mismo ordenador, en una red local o en la Internet. La sofisticación de la interfaz procede del *applet* de Java que accede a la base de datos y realiza el procesamiento lingüístico. Además,

la combinación de HTML y Java implica que el lexicón funcionará en cualquier plataforma que apoye un explorador de la WWW, ya que los dos exploradores principales (Microsoft Internet Explorer y Netscape) son compatibles con Java. Toda la ayuda y documentación on-line tienen también la forma de documentos de HTML, por lo que el usuario no tiene que aprender una nueva interfaz o modo de interaccionar con la información. Una ventaja añadida para los usuarios de la versión del diccionario basada en la Internet es la inclusión de una herramienta de correo electrónico en la que se pueden plantear preguntas y proporcionar retroalimentación al grupo que desarrolla el lexicón.

BIBLIOGRAFÍA

APRESIAN, J.D. (1993): «Systemic lexicography as a basis of dictionary-making», *Dictionaries: Journal of the Dictionary Society of North America* 14.

CARAMAZZA, A. (1996): «The brain's dictionary», *Nature* 380.

COSERIU, E. (1977): *Principios de semántica estructural*, Madrid, Gredos.

DIK, S.C. (1978): *Functional Grammar*, Dordrecht, Foris.

DIK, S.C. (1992): *Functional Grammar in Prolog. An Integrated Implementation for English, French, and Dutch*, Berlín, Mouton de Gruyter.

DIXON, R.M.W. (1991): *A new approach to English grammar on semantic principles*, Oxford, Clarendon.

HESLOP, B. & L. BUNDNICK. (1996): *Publicar con HTML en Internet*, Madrid, Paraninfo.

MARTÍN MINGORANCE, L. (1984): «Lexical Fields and Stepwise Lexical Decomposition in a Contrastive English-Spanish Verb Valency Dictionary». En R.R.K. HARTMANN (ed.) *LEXeter 83 Proceedings: Papers from the International Conference on Lexicography at Exeter*, Tübingen, Max Niemeyer.

NAUGHTON, P. (1996): *The Java Handbook*, Nueva York, Longman.

ROGET, P. (1987): «Introduction». En B. KIRKPATRICK (ed.) *Roget's Thesaurus of English words & phrases*, Harlow, Longman, 6.ª ed.