ALFINGE
Revista de Filología

# EL LEXICÓN EN LA LEXICOGRAFÍA COMPUTACIONAL: ADQUISICIÓN Y REPRESENTACIÓN DE INFORMACIÓN LÉXICA

*Antonio Moreno Ortiz*

NN: 0213-1854 nº 10, año 1998, págs. 249-272

# EL LEXICÓN EN LA LEXICOGRAFÍA COMPUTACIONAL: ADQUISICIÓN Y REPRESENTACIÓN DE INFORMACIÓN LÉXICA

## 1. El lexicón: concepto interdisciplinar

La importancia y centralidad del lexicón computacional en las aplicaciones de procesamiento de lenguaje natural (*NLP: Natural Language Processing*) en general es un hecho admitido por los más relevantes exponentes en el campo de la lingüística y lexicografía computacionales. La lista de referencias en este sentido sería inacabable; baste citar a modo de ejemplo representativo las palabras de la investigadora italiana Nicoletta Calzolari[1] cuando afirma:

> It is almost a tautology to affirm that a good computational lexicon is an essential component of any linguistic application within the so-called 'language industry', ranging from NLP systems to lexicographic enterprises.

En el mismo sentido se manifiesta Levin[2]

> ... [the lexicon] has often proved to be a bottleneck in the design of large-scale natural language systems, given the tremendous number of words in the English lexicon, coupled with the constant coinage of new words and shifts in the meaning of existing words.

---

1. Calzolari, N. 'Issues for Lexicon Building', A. Zampolli, N. Calzolari & M. Palmer (eds.) *Current Issues in Computational Linguistics: In Honour of Don Walker.* Liguistica Computazionale Vol. IX-X, Pisa, Giardini Editori e Stampatori, 1994, p. 267.
2. Levin, B. 'Building a Lexicon: The Contribution of Linguistics', en B. Boguraev (ed.). *Building a Lexicon. Special Issue. International Journal of Lexicography.* Vol. 4 Number 3, 1991, p. 205.

El problema del 'cuello de botella' es bien conocido en el entorno de la lexicografía y lingüística computacionales y ha sido reconocido por otros muchos investigadores relevantes[3]. En muchas ocasiones la construcción de lexicones para su uso en NLP ha sido realizada un tanto a la ligera, llegando en algunos casos a ser poco más que 'apéndices' de las gramáticas, y que de este modo no presentan grandes innovaciones con respecto a concepciones clásicas de más de seis décadas. En otros casos, una enorme cantidad de recursos económicos y humanos han sido puestos al servicio de la creación de lexicones computacionales creados a partir de diccionarios existentes, de los cuales se extrae la información léxica de un modo semi-automático. Aunque tales esfuerzos no dejan de tener un gran valor, pensamos que se hubiesen conseguido resultados mucho más aceptables si dichos esfuerzos se hubiesen orientado de otro modo, es decir, siguiendo una teoría lexicológica más apropiada, ya que en la mayoría de los casos ni siquiera se ha planteado la necesidad de recurrir a *una* teoría lexicológica. Pensamos que es necesaria, por tanto, la consideración del lexicón como elemento central de todo sistema de NLP, pero, al mismo tiempo, se debe contemplar desde una perspectiva más amplia, teniendo en cuenta los estudios elaborados por investigadores de otras disciplinas. Con el objeto de encuadrar nuestra visión global sobre la construcción de un lexicón, realizaremos un breve repaso del concepto de *lexicón*, tal y como se entiende hoy en día en los diversos ámbitos de estudio en los que se usa, desde la psicología hasta la lexicografía computacional. Dado su carácter multidisciplinar, este repaso no puede ser exhaustivo, sino que nos centraremos en las visiones del lexicón más destacadas

Tal y como reconoce Martha Evens[4] en su libro dedicado a modelos relacionales del lexicón, éste se ha convertido en el centro de atención de aquéllos que se dedican al estudio de los problemas relacionados con el lenguaje, sean éstos del tipo que sean. Dentro del ámbito de la lingüística existe hoy día un claro reconocimiento de que un análisis completo de la sintaxis y la semántica requiere un modelo de lexicón. También en otras disciplinas, como por ejemplo en la antropología, el estudio del léxico es fundamental, ya que los antropólogos no pueden describir una cultura sin hacer referencia al vocabulario que usan (o usaban) las personas que participan en las actividades propias de esa cultura.

---

3. Pustejovsky, J. 'The Generative Lexicon', *Computational Linguistics*, 17 (4), 1991. Boguraev B. & T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*. London, Longman, 1989.
4. Evens, M. W. *Relational Models of the Lexicon. Representing Knowledge in Semantic Networks*. Cambridge, Cambridge University Press, 1988.

En el ámbito de la psicología, los profesionales que se encargan de examinar el desarrollo y uso del lenguaje consideran que la organización del lexicón mental del hablante es una pieza clave. Las investigaciones que se centran en el estudio de la organización de la memoria han de estudiar también la organización de la información léxica y conceptual, así como la forma en la que accedemos a la información y la usamos para construir un discurso coherente. El estudio de las redes neuronales ha facilitado, en los últimos años, modelos más detallados sobre el acceso a la información léxica y otros procesos del lenguaje natural.

En el ámbito de la psicolingüística, el vocablo 'lexicón' se ha usado para hacer referencia al 'lexicón mental' del hablante de una lengua. Una de las cuestiones centrales de la psicolingüística contemporánea es el estudio de la adquisición del conocimiento léxico y de cómo éste se organiza en la memoria de un hablante para su acceso y uso inmediato. Para muchos psicolingüistas, entre ellos Miller[5] y Aitchinson[6], el hecho de que un hablante pueda acceder en milésimas de segundo a una cantidad ingente de vocabulario almacenado en su memoria[7], tanto en procesos de producción como de comprensión, es una prueba fehaciente de que el lexicón mental está organizado y estructurado de modo que posibilita el acceso inmediato. Emmorey y Fromkin[8] definen el lexicón mental como el componente de la gramática en el que

> (...) information about individual words and/or morphemes is entered, i.e. what a speaker/hearer of a language knows about the form of the entry (its phonology), its meaning (its semantic representation), and its combinatorial properties (its syntactic categorical properties).

La complejidad de la memoria léxica ha fascinado a muchos psicolingüistas, quienes han propuesto diferentes métodos para explorar y analizar los procesos cognitivos que se producen en su uso. Sin embargo, el estado actual de las investigaciones sobre memoria léxica y las dificultades para poder acceder al

5. Miller, G. A. (1986) 'Dictionaries in the Mind', *Language and Cognitive Processes*, Vol. 1, No. 3, pp. 171-185. Miller, G. A. (1991) 'Semantic Networks of English', *Cognition*, 41, pp. 197-229.
6. Jean Aitchinson (1987) *Words in the Mind: An Introduction to the Mental Lexicon*. Oxford, Basil Blackwell.
7. Aitchinson estima que el número de palabras que un hablante medio conoce es superior a 50.000 y puede llegar hasta 250.000. En cuanto al reconocimiento de palabras, un hablante nativo puede reconocer una palabra en menos de 200 ms., y en algunos casos, incluso antes de que se oiga.
8. Emmorey, K. & V. Fromkin 'The Mental Lexicon', en F. Newmeyer (ed.) *Linguistics: The Cambridge Survey Vol. III*, Cambridge: CUP, 1988, p. 12.

interior de la mente humana para observar su funcionamiento han provocado que los métodos propuestos sean principalmente analógicos o, tal y como reconoce Miller, se centren en el análisis de una pequeña porción del léxico.

En el ámbito computacional, los lexicones se consideran hoy día la base fundamental en la construcción de sistemas informáticos que posibiliten la interacción entre la máquina y el hombre. No se pueden construir sistemas de procesamiento de lenguaje natural que sean lo suficientemente robustos como para ocuparse de problemas del 'mundo real', sin antes diseñar lexicones de gran magnitud que contengan información léxica detallada.

El interés creciente por la generación de textos, la traducción automática, el desarrollo de sistemas de reconocimiento de habla, los interfaces en lenguaje natural y otras muchas aplicaciones que tienen como objeto el tratamiento automatizado del lenguaje humano, ha provocado una demanda constante de información léxica detallada sobre áreas de vocabulario cada vez más amplias. Sin embargo, la construcción de un lexicón para su utilización en este tipo de tareas requiere una elaboración concienzuda. En el resto de este trabajo estudiamos la metodología que debería seguirse para conseguir que un repositorio de información léxica de estas características ofrezca todas las garantías en cuanto a reutilización y multifuncionalidad, dos conceptos que han marcado la lexicografía computacional en esta década.

## 2. El lexicón en el procesamiento de lenguaje natural: la lexicografía computacional

Como ya hemos mencionado, todas las aplicaciones que tienen como objeto el tratamiento computacional del lenguaje natural consideran el lexicón como el componente central, lo que ha provocado una demanda constante de información léxica detallada sobre amplias áreas de vocabulario. La finalidad fundamental del procesamiento de lenguaje natural es la automatización de los procesos lingüísticos, tales como la comprensión, producción o adquisición de una lengua, tareas que los usuarios de una lengua realizan fluida y naturalmente. Tanto para los humanos como para las máquinas, todas estas tareas implican un conocimiento profundo del vocabulario de una lengua aunque, tal y como señala Boguraev[9], durante años los lexicones enfocados al procesamiento del lenguaje natural han sido los 'hermanos pobres' de la lingüística computacional.

9. B. Boguraev 'Building a Lexicon: The Contribution of Computers', en B. Boguraev (ed.). *Building a Lexicon. Special Issue. International Journal of Lexicography*. Vol. 4 Number 3, 1991, p. 3.

La mayoría de los sistemas diseñados hasta hace relativamente poco tiempo contenían sólo lexicones ilustrativos con no más de cien palabras[10] y, a pesar de los numerosos avances en este área, aún hoy no existe consenso sobre la naturaleza de la información que el lexicón debe contener ni, por supuesto, sobre la manera en la que la información debe ser representada. La tarea de construir un lexicón completo para una lengua natural es enorme. El *Oxford English Dictionary* (OED), por ejemplo, contiene 250.000 entradas de palabras independientes, y a pesar de tan elevado número, no incluye muchas palabras pertenecientes al vocabulario técnico.

Resulta por tanto muy costoso, tanto en recursos humanos como en tiempo y dinero, construir un lexicón 'a mano', lo que ha llevado a muchos investigadores a considerar las versiones electrónicas de los diccionarios impresos como fuentes potenciales de información léxica, que puede ser vertida de forma automática o semi-automática en sistemas de procesamiento de lenguaje natural.

El ámbito de investigación del procesamiento del lenguaje natural hace converger los intereses de lingüistas computacionales, psicolingüistas, informáticos e ingenieros de sistemas. Todos ellos, desde diferentes perspectivas teóricas y prácticas, intentan desarrollar una teoría que sea totalmente explícita (y por tanto automatizable) de los procesos lingüísticos. Aunque hasta la fecha no existe una teoría completa, no debemos olvidar que hay bastantes ejemplos de sistemas de NLP que han sido construidos sobre la base de un entendimiento parcial de algunos de esos procesos. Se han construido, por ejemplo, sistemas de síntesis hablada de textos (*text-to-speech-synthesisers*); estos sistemas no intentan 'entender' el *input*, sino que se apoyan en un análisis lingüístico superficial de las palabras que conforman el texto y en su organización sintáctica en oraciones. También se han conseguido sistemas de reconocimiento del habla que convierten una cadena de habla en su correspondiente escrito con bastante exactitud.

Bastante menos, sin embargo, se ha avanzado en las áreas de generación o comprensión del lenguaje, aunque hay un gran número de proyectos de investigación dedicados a ello. La mayoría de los sistemas de procesamiento de lenguaje natural adoptan un enfoque que se ha dado en llamar 'basado en el

10. Hecho que en la mayoría de los casos no se hacía explícito en informes, libros o tesis publicadas. Una anécdota relatada en Wilks et al. (*Electric Words: Dictionaries, Computers and Meanings*. Cambridge, Mass, The MIT Press, 1996) refleja esta situación con bastante precisión: hace cinco años, se le preguntó a un grupo de investigadores del campo de NLP cuál era *realmente* el número de palabras contenidas en los lexicones de sus sistemas. La media de estas respuestas fue de 36, una cifra, en palabras de los autores, '*often taken to be a misprint when it appears, though it was all too true...*'.

conocimiento'[11], ya que para llevar a cabo la tarea para la que están diseñados, necesitan incorporar conocimiento lingüístico explícito, junto con otros tipos de conocimiento de carácter más general. Por ejemplo, un sistema de síntesis que convierta un texto en su correspondiente cadena hablada, necesita 'conocimiento' sobre la pronunciación de las letras y las secuencias de letras, así como de las palabras individuales que no siguen las reglas generales. También precisa conocimiento sobre los patrones rítmicos de acentuación y de cómo la organización sintáctica afecta la prosodia y la entonación. Sin esta información el sistema no podría sintetizar el texto correctamente. Por tanto, la importancia del lexicón es fundamental y su construcción debería estar guiada por parámetros bien determinados si pretendemos mejorar la utilidad de los sistemas automáticos de tratamiento del lenguaje natural.

Se pueden distinguir dos grandes ámbitos de investigación en lo referente a la creación de lexicones computacionales: el de la *adquisición* y el de *representación* de conocimiento léxico. El primer término suele ser empleado en empresas de reutilización de recursos existentes, normalmente diccionarios en formato magnético (*MRD: Machine Readable Dictionary*), pero también a la adquisición de información léxica mediante córpora textuales. El término 'representación', por otra parte, se enmarca en el más amplio campo de la representación del conocimiento y los sistemas de información. En general, estas son las dos fases principales contempladas en la construcción de un lexicón computacional y se pueden considerar como separadas pero interdependientes. Veamos cuales son las metodologías aplicables en cada una de estas dos fases.

2.1. Adquisición de conocimiento léxico

La adquisición de la información léxica necesaria para popular lexicones computacionales plantea serios problemas, tanto en lo que se refiere a la efectividad de los diferentes métodos que se han empleado como a la inversión de tiempo, dinero y recursos humanos y computacionales que estos métodos requieren.

Se puede considerar que existen tres métodos o fuentes principales para la adquisición de conocimiento léxico: (i) adquisición manual de información léxica, (ii) diccionarios en formato magnético (MRDs) y (iii) los córpora textuales informatizados.

Los tres métodos plantean ventajas y desventajas, tanto en lo que se refiere a los recursos que requieren como a la efectividad que han demostrado hasta ahora.

11. En este ámbito, el término 'conocimiento' se utiliza de una manera muy general, como oposición a los sistemas estadísticos y matemáticos.

2.1.1. La adquisición mediante MRDs

Aunque en principio las fuentes electrónicas pueden aportar una gran cantidad de información lingüística muy valiosa, que puede servir como punto de partida para la creación de una base de datos léxica (*LDB: Lexical Data Base*), en la práctica es difícil aprovechar toda la información que esas fuentes electrónicas contienen. Los diccionarios en formato magnético, por ejemplo, parecen particularmente apropiados como base para la construcción de lexicones automáticos[12], ya que la información que en ellos se encuentra está estructurada en cada una de las entradas, y parece posible extraer cierta información con bastante facilidad. Sin embargo, después de muchos años de investigación y de multitud de proyectos dedicados a ello, los resultados obtenidos en la adquisición de información léxica a partir de MRDs están lejos de ser satisfactorios.

El problema fundamental es que los diccionarios están diseñados para ser usados por humanos. Los usuarios son hablantes nativos de una lengua, que saben, al menos implícitamente, como está estructurado el lexicón de su lengua. Los lexicógrafos, a la hora de compilar un diccionario, explotan el conocimiento lingüístico de sus usuarios potenciales, de modo que las entradas contienen sólo la información necesaria para que un hablante de una lengua sea capaz de conectarla con su conocimiento lingüístico general. Incluso los diccionarios diseñados especialmente para los estudiantes de una lengua (*learners' dictionaries*) tienen en cuenta las propiedades generales del lenguaje, aunque contengan información mucho más detallada (sobre todo a nivel sintáctico y de uso) que cualquier otro tipo de diccionario.

Tal como reconoce Levin[13], el valor que posee el uso de los diccionarios electrónicos en la construcción de una base de conocimiento léxico se ve limitado, en muchas ocasiones, por la esencia misma del arte de la lexicografía: los diccionarios están elaborados por lexicógrafos, que son 'seres humanos' (y no 'máquinas'), que trabajan bajo grandes presiones de tiempo y espacio.

Esto provoca que la mayoría de ellos sean inconsistentes e incompletos[14], y que, por ejemplo, palabras que tienen un comportamiento similar (morfológico, sintáctico, semántico, etc.) no reciban un tratamiento homogéneo en los

12. De hecho, los primeros intentos de usar diccionarios electrónicos en el proceso de construcción de bases de conocimiento léxico se remontan a finales de los años 60.
13. B. Levin, 'Building a Lexicon: The Contribution of Linguistics', en B. Boguraev (ed.). *Building a Lexicon. Special Issue. International Journal of Lexicography*. Vol. 4 Number 3, 1991.
14. S. Atkins, J. Kelg & B. Levin 'Anatomy of a Verb Entry', *International Journal of Lexicography*, Vol. 1, No. 2, 1988. B. Boguraev & T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*. London: Longman, 1989.

diccionarios, ya sea por falta de tiempo, por haber sido compiladas por diferentes lexicógrafos, o simplemente porque el lexicógrafo no fue capaz de reconocer las similitudes[15].

Han sido numerosos los proyectos orientados a la extracción de información de versiones electrónicas de diccionarios impresos en papel. Si atendemos a la cantidad de bibliografía que se puede encontrar relativa a este tema, puede parecer a primera vista que un gran número de diccionarios ha sido usados con este propósito, aunque en realidad no es así, puesto que casi todos los proyectos en este área se han centrado en un número reducido de diccionarios, bien por problemas con los derechos de publicación o bien por la falta de las cintas magnéticas correspondientes a las versiones publicadas en papel.

De hecho, los diccionarios usados se reducen a los siguientes: *Oxford Advanced Learner's Dictionary of Current English* (OALD), *The Collins Cobuild English Language Dictionary* (COBUILD), *Longman Dictionary of Contemporary English* (LDOCE), *Webster's Seventh Collegiate Dictionary* (W7), *Merriam-Webster Pocket Dictionary* (MWPD). Las diferencias que se puede apreciar en las entradas léxicas de estos diccionarios han sido ya analizadas en diversas publicaciones[16] por lo que no nos detendremos a hacerlo aquí.

Existe, sin embargo, una distinción común a todos ellos, la que se hace entre los 'datos' (el contenido léxico propiamente dicho) y la 'estructura' (el formato, los códigos y las distinciones tipográficas dentro de cada entrada). Esta distinción es muy relevante, ya que los 'datos' constituyen una fuente de información 'explícita' que se pensaba que podía ser extraída con facilidad, y de hecho la mayoría de los proyectos iniciales estaban orientados a obtener información de la parte de las entradas que contenía los datos léxicos. En estos proyectos no se hacía uso del potencial de información que la 'estructura' de una entrada léxica también ofrece.

---

15. También podríamos detenernos a considerar las importantes diferencias que se observan si consultamos la misma entrada léxica en varios diccionarios, no sólo en cuanto a la división de los significados de una palabra, sino también en cuanto a su comportamiento sintáctico, colocacional, etc. Esta diferencia se hace mayor si la información contenida en las entradas se compara con la que se podría extraer de las ocurrencias de esa palabra en un corpus textual informatizado. No ahondaremos en esos aspectos aquí, aunque un punto de referencia muy interesante en este sentido, con especial énfasis en la construcción de MRDs válidos para NLP, se encuentra en Atkins (1991).
16. S. Atkins. 'Building a Lexicon: The Contribution of Lexicography' , en B. Boguraev (ed.). *Building a Lexicon. Special Issue. International Journal of Lexicography*. Vol. 4 Number 3, 1991.

Posteriormente, algunos investigadores observaron que hay muchos aspectos en la estructura de las entradas (tanto a nivel individual como en su interrelación al formar parte de la macroestructura del diccionario) que contienen, de forma 'implícita', información que puede ser muy relevante, ya que los códigos que controlan el formato de la entrada, así como los diferentes tipos de letra y otros caracteres especiales son siempre significativos a la hora de leer una entrada en un diccionario. Un lector humano se acostumbra a ellos con rapidez y es capaz de darles la interpretación adecuada. En este sentido, algunos proyectos orientados a la extracción de información de MRDs han intentado dar cuenta tanto de la información explícita en las entradas como de la implícita, aunque esto último es bastante más complejo de lo que a priori puede parecer.

Los primeros trabajos realizados con los diccionarios electrónicos se dedicaron a estudiar frecuencias de palabras en las definiciones, una tarea muy costosa en términos computacionales, sobre todo si tenemos en cuenta los recursos informáticos de la época. Al mismo tiempo, y quizás influenciados por las investigaciones llevadas a cabo en el ámbito de la IA relativas a las redes semánticas, se estaban empezando a estudiar los 'enlaces' (*links*), 'cadenas' (*chains*) y 'círculos' (*circles*) que se forman en un diccionario a través de las palabras que se usan en sus definiciones, con vistas a la construcción automática de taxonomías.

En esta línea de investigación, el trabajo de una investigadora que aún hoy (veinte años después) sigue liderando la comunidad lexicográfica computacional, Karen Sparck-Jones, demostró que la 'circularidad' debe, en principio, existir en un diccionario, ya que cada palabra usada en las definiciones ha de ser, a su vez, definida en el diccionario. Algunas de estas circularidades mantienen una distancia semántica reducida, como por ejemplo las definiciones mutuas de 'good' y 'excellent', y son por tanto fáciles de observar y asimilar por un lector humano, pero son muy difíciles de localizar a nivel formal y esto puede dificultar enormemente la labor de extracción de información de las definiciones, sobre todo si se aplican nociones empíricas de derivación circular.

A partir de la segunda mitad de los años ochenta se puede apreciar un cambio en las investigaciones relacionadas con los diccionarios en formato magnético, cambio que vino precedido por la sucesiva publicación de diccionarios especializados para estudiantes de inglés. La estructura de estos diccionarios parecía, a priori, muy adecuada para su uso en NLP, ya que cuentan con una formalización interna mucho mayor que otros diccionarios y son mucho más explícitos en lo que se refiere a las características sintácticas, morfológicas y semánticas de cada una de las entradas. De entre estos diccionarios, los que han recibido una mayor atención han sido, sin lugar a dudas, el LDOCE y COBUILD, y en menor medida el OALD.

La versión magnética del LDOCE contiene 41.000 entradas, con información adicional a la que se encuentra en la edición en papel. Sus autores defienden que las entradas han sido definidas usando un vocabulario 'controlado' de 2.000 palabras y que las entradas tienen una sintaxis simple, lo que parece reducir la circularidad en las definiciones a la que hacíamos referencia anteriormente. Esto ha causado que un gran número de investigadores hayan dedicado sus esfuerzos al estudio de las definiciones, empleando métodos muy diversos que van desde la aplicación de análisis estadísticos para la asignación de significado a técnicas para la localización del *genus* y la *diferencia específica* de las definiciones.

Sin embargo, tal y como se demuestra en estos estudios, las palabras que integran el vocabulario controlado son seis veces más ambiguas que las demás palabras que aparecen en el diccionario, ya que cada una de ellas tiene una media de 12 significados diferentes, frente a una media de 2 en el resto de las palabras. Además, este tipo de definiciones con vocabulario controlado hace que éstas sean más largas y que las referencias cruzadas entre definiciones (tanto explícitas como implícitas) sean mucho más frecuentes. Debido a estas dos características, sólo un sistema de NLP con una capacidad de comprensión lingüística muy sofisticada podría hacer uso de la información contenida en las definiciones; paradójicamente, conseguir un sistema con esta capacidad nos hace volver a los problemas iniciales que hacían de los MRDs herramientas de posible utilidad en los sistemas de NLP.

El LDOCE también cuenta con un sistema de 110 códigos gramaticales, junto con un grupo de identificadores semánticos, tales como 'abstracto', 'concreto', 'animado', etc., que se usan para asignar restricciones de selección a los argumentos de los verbos. El sistema de codificación gramatical usado deriva de un modelo lingüístico específico (basado en Quirk et al. 1972), lo que ha provocado que no sea apropiado (o incluso incompatible) con los *parsers* automáticos de algunos sistemas de NLP. Uno de los problemas más serios que ha planteado el uso del LDOCE es que los códigos, en algunos casos, mezclan información sintáctica y semántica, mientras que en otros sólo ofrecen información sintáctica superficial y en otros casos han sido modificados por el lexicógrafo para hacer que el aspecto visual de la entrada sea más claro o más compacto. Es necesario analizar los códigos con algoritmos muy complejos para poder separar la información semántica de la sintáctica y aun así los procesos que se han desarrollado hasta la fecha no han alcanzado resultados demasiado satisfactorios, incluso en aquellos casos en los que las rutinas automatizadas se han combinado con procesos manuales.

Hemos nombrado ya algunos de los problemas y desventajas que los MRDs plantean, en cuanto a la falta de consistencia e inexactitud de la información que

contienen (sea ésta sintáctica, semántica, morfológica o de uso), pero aún nos parece más importante la falta de aquella información detallada que no aparece en ningún diccionario y que un lexicón diseñado para un sistema de NLP necesita, por no mencionar aquellas unidades léxicas que, por falta de espacio o por motivos editoriales, no aparecen en el diccionario. Otro problema destacable son los errores tipográficos contenidos en las cintas originales de los MRDs: corregir estos errores es muy costoso tanto en tiempo como en recursos humanos[17].

También hemos de destacar, sin embargo, que no todas las investigaciones realizadas con MRDs han sido infructuosas. No nos parece apropiado detenernos aquí a hacer un repaso exhaustivo de los numerosos proyectos llevados a cabo para la extracción de información de MRDs, puesto que nuestra intención inicial era sólo destacar los inconvenientes y las ventajas que éstos ofrecen en cuanto a la adquisición de conocimiento léxico, por lo que, para ofrecer una visión equilibrada de las investigaciones llevadas a cabo con MRDs, debemos también nombrar algunas iniciativas en las que el uso de diccionarios electrónicos ha dado resultados positivos. Boguraev & Briscoe[18], por ejemplo, implementaron con éxito un algoritmo para convertir a formato PATR los códigos gramaticales que el LDOCE asigna a los verbos según los complementos que seleccionan. Usando las definiciones del LDOCE, por ejemplo, Pustejovsky[19] ha diseñado un sistema capaz de construir entradas verbales de forma semi-automática. Por otro lado, el proyecto ACQUILEX también ha ofrecido resultados bastante satisfactorios en cuanto a la construcción de redes semánticas extraídas de diccionarios.

Éstos son sólo algunos ejemplos de investigaciones cuyos resultados pueden considerarse bastante satisfactorios, aunque analizándolos cuantitativamente debamos plantearnos si el tiempo empleado en construir sus lexicones con esos métodos no es muy similar al tiempo que se emplearía en construir un lexicón manualmente. Otro problema se refiere a la excesiva dependencia teórica que estos lexicones plantean, ya que el formato PATR sólo permite su conversión posterior a gramáticas basadas en formalismos de unificación o en el caso de Pustejovsky las entradas verbales en el lexicón son construidas (o mejor dicho, traducidas) a fórmulas complejas de lógica de primer orden.

17. Por ejemplo, se tardó casi un año en comprobar y corregir la cinta magnética que contenía el OALD ya que un elevado número de errores fueron introducidos en el proceso de teclear en el ordenador la información contenida en el diccionario en papel.
18. B. Boguraev & T. Briscoe (eds.) *Computational Lexicography for Natural Language Processing*. London, Longman, 1989.
19. J. Pustejovsky 'On the Acquisition of Lexical Entries: the Perceptual Origin of Thematic Relations', *Proceedings of the 25th Annual Conference of the Association for Computational Linguistics*, 1987, pp. 172-178.

En términos generales, la mayoría de los problemas que el uso de MRDs ha planteado en la construcción de lexicones computacionales parecen derivarse no sólo de su condición de producto realizado por y para los humanos, sino también de la gran diversidad de teorías, tanto sintácticas como de otro tipo, que pueden subyacer a la construcción de cada sistema para el que se han intentado usar. Cada una de estas teorías puede representar información similar de manera muy diferente o puede incluso trazar una línea divisoria diferente entre la información que ha de aparecer en el lexicón y la información que debe aparecer en otros componentes del sistema. Otra de las razones que se han esgrimido en contra del uso de diccionarios electrónicos para la adquisición de conocimiento léxico es el hecho bien conocido y estudiado de que, mientras que el lenguaje es un objeto dinámico que evoluciona constantemente, los diccionarios son, por definición, objetos estáticos. El lapso de tiempo que transcurre entre el proceso de compilación y la edición, publicación y distribución de un diccionario, hace imposible que pueda ser un reflejo totalmente actualizado de una lengua, situación que se va agravando cuanto más tiempo ha pasado desde su publicación.

2.1.2. La adquisición léxica mediante córpora textuales

Este problema, junto con alguno de los ya mencionados anteriormente, ha provocado que en los últimos diez años se haya considerado en algunos proyectos de enorme magnitud (como por ejemplo WordNet, o Cyc) la entrada manual de datos como el método más económico y seguro de adquisición de conocimiento léxico, aunque consideraciones de este tipo también han llevado a contemplar los córpora textuales informatizados como fuentes potenciales para la adquisición de información léxica actualizada. Esta tendencia es consecuencia del reciente resurgimiento de la aplicación de métodos empíricos y estadísticos al análisis lingüístico, y ha desarrollado una corriente propia en el ámbito de la lexicografía comercial que se conoce como *Lexicografía de Corpus*.

Debido al carácter dinámico y evolutivo del lenguaje al que hacíamos referencia antes, la lexicografía de corpus considera que el proceso de compilación de un nuevo diccionario debe derivarse del estudio y análisis exhaustivo de la lengua, tal y como ésta es usada por sus hablantes en situaciones reales, es decir, a través del estudio de un corpus representativo de textos, tanto orales como escritos, de una lengua. Las crecientes posibilidades de obtener y almacenar enormes cantidades de texto en formato magnético han hecho posible que algunas editoriales hayan usado intensivamente los córpora textuales en el proceso de compilación de sus diccionarios, tanto en la creación de las entradas léxicas del diccionario

como en la división de significados de las entradas, la selección de los ejemplos de uso o la información gramatical y colocacional que se incluye en las entradas.

No sólo se ha avanzado, en términos computacionales, en lo que se refiere al poder de almacenamiento de textos; también se ha avanzado enormemente en el desarrollo de herramientas computacionales que facilitan el manejo y estudio de los córpora textuales, aunque ésta siga siendo, en muchos aspectos, un área bastante controvertida[20], sobre todo en lo que concierne a sus aspectos teóricos y metodológicos. La mayoría de los experimentos llevados a cabo para la adquisición de información léxica a través de córpora se hallan aún en fase experimental, por lo que quizás sea aún pronto para extraer conclusiones definitivas sobre su utilidad. En el momento presente, los córpora textuales han demostrado ser de gran utilidad en el ámbito de la lexicografía comercial y están siendo aplicados con éxito a otras áreas del procesamiento de lenguaje natural, como por ejemplo en la categorización de nombres propios o en la desambiguación léxica por medio de la aplicación de métodos estadísticos.

Aunque ésta es un área en la que se está avanzando con gran rapidez, parece claro que queda aún un largo camino por recorrer, ya que la información que se puede obtener hoy día de los córpora a través de análisis cuantitativos representa sólo una parte de la que un lexicón computacional requiere, y la extracción automática de información es aún muy costosa en lo que respecta a recursos computacionales y humanos.

Tal y como ya señalamos en referencia al uso de diccionarios en formato electrónico, la extracción automática de información léxica de un corpus textual informatizado requiere de antemano la capacidad de analizar automáticamente el texto de diversas maneras, para lo que se necesita un sistema de procesamiento de lenguaje natural con unas capacidades de comprensión lingüística muy sofisticadas.

---

20. Tanto la Lingüística de Corpus como su disciplina 'hermana', la Lexicografía de Corpus, son ámbitos de estudio de reciente creación, cuyos principios teóricos y metodológicos están aún en proceso de definición, tarea que, debido a su carácter eminentemente aplicado y experimental, no resulta una tarea nada fácil. No podemos detenernos aquí a analizar estos aspectos pero baste señalar que los investigadores pioneros en este área mantienen posturas diferentes en aspectos tan importantes como el diseño de un corpus representativo de una lengua, la explotación probabilística o no-probabilística del corpus, relación (o elección) entre la calidad o la cantidad de texto a utilizar, etc.

21. Por ejemplo, D. Hindle & M. Rooth 'Structural Ambiguity and Lexical Relations', *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics.* ACL, 1991; B. Boguraev & J. Pustejovsky *Corpus Processing for Lexical Acquisition.* Cambridge, Mass: The MIT Press, 1996.

Por estas razones, un gran número de investigadores[21] apuntan al uso conjunto de varias fuentes para la adquisición de conocimiento léxico, puesto que en ninguna de ellas aisladamente se puede encontrar toda la que un lexicón requiere. Nuestra línea de investigación también apunta en esta dirección, ya que estamos convencidos de que los córpora pueden ofrecer información léxica muy relevante, sobre todo en aspectos relativos a los hábitos colocacionales de las unidades léxicas o sus propiedades combinatorias, y son una herramienta de gran utilidad para la extracción de ejemplos reales de uso, así como en el enriquecimiento y refinamiento de la información ya contenida en un lexicón computacional[22].

Un caso interesante de uso conjunto de varias fuentes es la investigación llevada a cabo por Hearst y Schüetze[23], ya que usan una base de datos construida manualmente, WordNet[24], y aplican métodos estadísticos a un corpus para mejorar la clasificación semántica y las relaciones que aparecen en la misma. Su intención es adaptar el contenido de WordNet para que sea capaz de asignar una etiqueta que caracterice documentos de acuerdo al tema del que tratan. En su trabajo, ellos explican el proceso mediante el que se obtienen representaciones semánticas de un gran número de palabras extrayéndolas de cálculos estadísticos de co-ocurrencia léxica, aumentando y reubicando los elementos del lexicón, y haciéndolo más apropiado para otras tareas específicas a un dominio determinado (*domain-specific task*), como por ejemplo la recuperación de información (*information retrieval*).

Otro ejemplo de entrada manual de conocimiento léxico lo encontramos en la creación de la herramienta Ontos, dentro del proyecto de traducción automática basada en el conocimiento KBMT-89[25], que tenía la finalidad de desarrollar motores de traducción automática basada en el conocimiento, ya que sus investigadores están convencidos de que la información léxica que un sistema de traducción automática robusto requiere no puede encontrarse en fuentes disponibles en formato electrónico.

---

22. Véase A. Moreno Ortiz, A. 'Current Techniques in Lexical Information Retrieval and Manipulation', en J. Pérez Guerra (ed.) *Proceedings of the XIXth International Conference of AEDEAN*. Universidad de Vigo, 1996, 425-430.
23. Hearst, M. & H. Schüetze (1996) 'Customising a Lexicon to Better Suit a Computational Task', B. Boguraev & J. Pustejovsky (eds.) *Corpus Processing for Lexical Acquisition*. Cambridge, Mass: The MIT Press, 1996.
24. G. A. Miller, 'Dictionaries in the Mind', Language and Cognitive Processes, Vol. I, No. 3, 1986, 171-185. G. A. Miller, 'Semantic Networks of English', Cognition, 41, 1991, pp. 197-229. Miller, G. A. (1993) 'Introduction to WordNet: An On-line Lexical Database', en G. A. Miller et al. (eds.) *Five Papers on WordNet™*. CSL Report 43.
25. R. D. Brown & S. Nirenburg, 'Human-Computer Interaction for Semantic Disambiguation', en *Proceedings of the 13th International Conference on Computational Linguistics (COLING '90)*. Vol. 3, 1990, pp. 42-47.

En conclusión, las fuentes electrónicas (los MRDs y los córpora textuales) están lejos aún de ofrecer la información léxica detallada que un lexicón computacional requiere y que, en la mayoría de los casos, el esfuerzo y dinero que se debería invertir para extraer de ellos una cantidad mínima de tal información puede ser bastante mayor a la que supondría la población manual de un lexicón computacional. Los problemas que plantean el uso de tales fuentes han llevado a que proyectos de gran magnitud se hayan llevado a cabo por medio de métodos manuales. Por ejemplo el proyecto Cyc[26], que está aún en fase inicial, está orientado a la construcción de una base de conocimiento que contenga el conocimiento humano necesario para hacer inferencias, por lo que sus investigadores están vertiendo manualmente lo que ellos consideran que conforma la información morfológica, sintáctica, semántica y pragmática que los hablantes asociamos con una palabra.

2.2. Representación de conocimiento léxico: los lexicones computacionales

Una distinción previa es necesaria en referencia a los términos *información* y *conocimiento*: hasta ahora los hemos usado como sinónimos, aunque en el ámbito de la tecnología de la información hacen referencia a conceptos distintos, ya que las aplicaciones que han de gestionarlos son de naturaleza distinta (bases de datos vs. bases de conocimiento).

En general, podemos decir que el *conocimiento* es abstracto y generalizador. Los *datos* son concretos y describen de forma detallada entidades del mundo real. *Información* es lo que se obtiene cuando los datos son analizados, ya sea por un agente humano o por una aplicación externa. Por tanto, de una base de datos se extrae información, mientras que ésta se encuentra explícitamente e implícitamente almacenada en una base de conocimiento.

Una interpretación en términos cognitivos de este extremo es que una base de conocimiento contiene información representada de una forma más parecida a cómo los humanos la almacenamos[27]. El término 'conocimiento', hace referencia específica a las reglas en las que la información ha de ser usada, así como a diversos procesos cognitivos en relación con la actualización de la información contenida en un sistema. Si un sistema contiene información representada a modo de generalizaciones sobre datos y es capaz de usar esta información de forma inteligente, actualizarla según necesidades y proveer *información acerca de la información* que contiene (ser 'consciente' de sus

26. R. V. Guha & D. B. Lenat 'CyC: a Midterm Report', *Artificial Intelligence*, 11, 1990, pp. 32-59.

27. Véase Walker, A. et al. (ed.) *Knowledge Systems and Prolog*. Reading, Mass, Addison-Wesley, 1987.

limitaciones), entonces es un sistema de conocimiento. Si, por el contrario, un sistema contiene una gran cantidad de datos específicos sobre un determinado universo y provee los mecanismos necesarios para recuperar y modificar esa información, ya sea por un programa o por un usuario humano, entonces es un sistema de base de datos.

En el entorno de la representación léxica, sin embargo, los términos *knowledge (data)base, lexical data base, lexical base, lexical knowledge base*, tienden a ser utilizados de una forma no muy clara en las publicaciones internacionales que tratan de este tema. Casi todos los autores los usan indistintamente, y en caso de definirlos, ofrecen definiciones bastante vagas o genéricas. La definición que Boguraev & Pustejovsky[28], ofrecen en su reciente libro dedicado a los avances en la construcción de lexicones computacionales es la siguiente:

> (...) the term *'lexical knowledge base'* has become a widely accepted one. Researchers use it to refer to a large-scale repository of lexical information, which incorporates more than just static descriptions of words, e.g., by means of clusters of properties and associated values. A lexical knowledge base would state
>
> - constraints on word behaviour
> - dependence of word interpretation on context, and
> - distribution of linguistic generalisations
>
> It is essentially a dynamic object, as it incorporates, in addition to its information types, the ability to perform inference over them and thus induce word meaning in context. This is the sense of 'computational lexicon', the population of which is giving new meaning to lexical acquisition from MRDs...

Como observamos, una base de conocimiento léxico se diferencia de un diccionario en formato magnético en que contiene y gestiona no sólo datos estáticos descriptivos de entradas léxicas, sino también reglas generalizadoras sobre el funcionamiento de esas unidades léxicas en contexto.

Hablando en términos generales, se pueden identificar al menos *cinco tipos de conocimiento* que son relevantes en cualquier sistema de procesamiento de lenguaje natural, y el hecho de que la mayoría de este conocimiento sea almacenado en el lexicón hace de él el eje central de cualquier sistema de NLP: fonológico, morfológico, sintáctico, semántico y pragmático. Cada uno de estos cinco tipos (generales) de conocimiento puede ser caracterizado

28. Boguraev, B. & J. Pustejovsky, *Corpus Processing for Lexical Acquisition*. Cambridge, Mass: The MIT Press, 1996, p.6.

por medio de un conjunto de *reglas*. Por ejemplo, es una regla de tipo sintáctico en inglés que los adjetivos aparezcan siguiendo al verbo *to be* y precediendo al nombre en las frases nominales:

Ej. *That house is beautiful* *That beautiful house*

Sin embargo, esta regla no se puede aplicar en el caso de un adjetivo como por ejemplo '*lone*':

Ej.: *That lone house* **That house is lone*

Este tipo de particularidades deben hacerse explícitas en el lexicón. En este caso, es necesario representar el hecho de que un grupo de adjetivos no pueden ser usados como atributos (en inglés, a esta posición se la conoce tradicionalmente como '*predicative position*'). Si no podemos encontrar una regla general que identifique a los miembros de este grupo, habrá que marcar a cada miembro con un rasgo, como por ejemplo '*non-predicative*', y restringir de este modo la aplicación de la regla sintáctica a los adjetivos que no estén marcados con ese rasgo.

Este tipo de reglas sintácticas se pueden representar en la forma de reglas de estructura sintagmáticas (*phrase structure rules*), que tienen la siguiente forma genérica:

Mother —> $\text{Daughter}_i$, $\text{Daugher}_j$ ..... $\text{Daughter}_n$

en la que la categoría sintáctica 'madre' puede contener una o más categorías 'hija' y las categorías consisten en un nombre con algunos rasgos sintácticos opcionales. Una gramática escrita con esta notación que sólo genere ejemplos gramaticalmente correctos sería:

1. S —-> NP VP
2. VP —-> V AP[prd +]
3. NP —-> Det N
4. NP —-> Det AP[prd x] N
5. AP[prd x] —-> A [prd x]

En estas reglas se especifica que una oración (S) está formada por una frase nominal (NP) y una frase verbal (VP), y que una frase verbal, a su vez, puede estar formada por un verbo y una frase adjetiva (AP) marcada como [prd +]. El rasgo [prd] tiene dos valores (- / +) y se usa para indicar si una categoría adjetiva puede o no aparecer como atributo. Las reglas 3 y 4 especifican que una frase nominal está formada por un determinante (Det) y un nombre (N), con una frase adjetiva que puede aparecer opcionalmente. Este frase adjetiva está marcada como [prd x], donde x es una variable que representa los posibles valores de este rasgo. La última regla especifica que una frase adjetiva está formada por un adjetivo con valor [prd] que ha de ser idéntico al de la frase adjetiva.

Al proceso de aplicar una gramática a una oración para determinar si es gramaticalmente correcta o no se le conoce como análisis sintáctico (*parsing*). Sin embargo, para aplicar esta gramática de forma mecánica y automatizada a una oración, es necesario contar con un lexicón que ofrezca información al analizador sintáctico (*parser*) sobre las categorías gramaticales que están asociadas a las palabras que aparecen en la oración que se desea analizar. Siguiendo con nuestro ejemplo anterior, un lexicón (incompleto, por supuesto) para nuestra oración sería:

beautiful : A [prd+], A [prd-]
house : N
is : V
lone : A [prd -]
that : Det

El proceso de análisis sintáctico sería el siguiente:

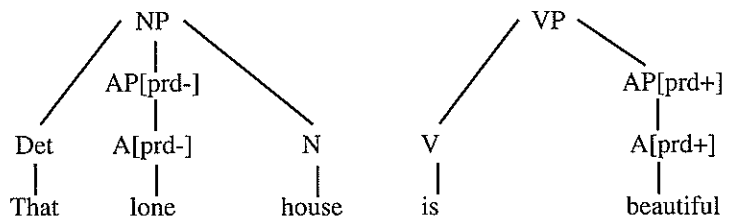
1. Para cada una de las palabras del lexicón, se consulta el lexicón y se le asigna una categoría sintáctica.
2. Se intenta hacer corresponder las reglas que contienen categorías léxicas 'hijas' explícitas con las categorías léxicas de las palabras de la oración, yendo de izquierda a derecha. Después, las categorías madre se relacionan con las hijas tal y como se especifica en cada regla.
3. Se construye un árbol de representación con un nodo raíz S (oración) llenando los huecos con más reglas que conectan las demás categorías.

Por ejemplo, el primer paso en el análisis de la siguiente oración:

*That lone house is beautiful*

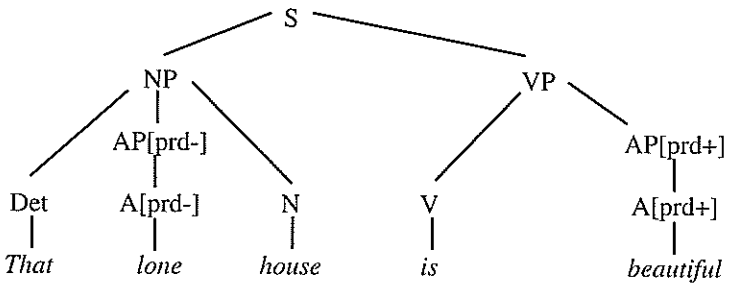
sería:


El segundo paso nos daría el siguiente árbol parcial:



En este caso, el valor variable [prdX] contenido en la regla 5 se ha asignado a un adjetivo (*beautiful*) que puede actuar con dos valores, aunque en este caso se ha seleccionado el valor [prd+]. Si se hubiera seleccionado el valor [prd-] de nuestro pequeño lexicón, la regla número 2 no hubiera correspondido con el valor asignado, y el proceso de *parsing* debería volver a realizarse.

El tercer paso añadiría el nodo S (oración), asignando así la regla nº 1:




Con este ejemplo pretendemos ilustrar la interacción necesaria entre las reglas (gramática) y el lexicón. En nuestro ejemplo, el lexicón se adaptaba a la gramática que habíamos diseñado, pero ambos tendrían que ser extendidos cada vez que se introdujeran reglas nuevas en la gramática o se añadieran palabras al lexicón. Si añadimos, por ejemplo, un verbo no copulativo, como *hate*, necesitaríamos hacer una distinción entre diferentes tipos de verbos, tanto en la gramática como en el lexicón, para evitar que se generen oraciones incorrectas como por ejemplo:

* *That lone house hates beautiful*

Esto demuestra la necesidad de que en cualquier sistema de procesamiento de lenguaje natural exista una gran interconexión entre las reglas generales que se incorporan a la gramática y la información incluida en las entradas del lexicón, ya que el lexicón deberá aportar toda la información que no sea predecible de las reglas, y deberá 'rellenar' estas reglas de modo que funcionen correctamente.

El lexicón también ha de incluir otros tipos de información idiosincrásica o no derivable de reglas, como por ejemplo, en el caso del inglés, la pronunciación, que se considera normalmente como un aspecto lingüístico que no se puede derivar del significado de las palabras o de su forma morfológica.

El significado de las palabras tampoco parece ser derivable de reglas generales, por lo que debe ser definido para cada entrada léxica, aunque en ciertos casos se puedan hacer consideraciones morfológicas; por ejemplo, el signifi-

cado del verbo *think* no es predecible, pero el de *rethink* puede serlo, a través de una regla morfológica que combine el significado del prefijo *–re* con el significado de *think*[29] (lo que podríamos glosar como *'to think again'*). Consideraciones de este tipo indican que el conocimiento morfológico debe estar también integrado en el lexicón.

Un lexicón válido para tareas de procesamiento de lenguaje natural ha de contener, por tanto, información fonológica, morfológica, sintáctica, semántica y pragmática, pero además esta información debe ser estructurada de forma que permita su reutilización para diversas tareas.

Esta estructuración conforma un *modelo de datos* o *esquema conceptual*, para ser más precisos, que va a determinar lo siguiente:

- cuáles son y qué forma tienen los objetos o entidades léxicas consideradas
- qué tipo de datos pueden o deben ser introducidos para cada una de esas entidades
- qué interrelaciones mantienen las distintas entidades que conforman la base léxica
- cuáles son las propiedades definitorias de cada una de esas entradas
- qué restricciones se imponen sobre la información que se asigna a cada una de las entidades

Evidentemente, este modelo de datos va a determinar el lexicón resultante, por lo que su diseño conlleva un buen número de decisiones importantes que condicionarán características fundamentales como son la reutilización de la información léxica almacenada en él y su funcionalidad en aplicaciones distintas. Dada su cardinal importancia, es sorprendente, sin embargo, la escasa atención que el diseño de lexicones computacionales ha recibido en la pasada década. Fue únicamente a raíz de proyectos financiados de cooperación internacional cuando recibió la atención debida. Nos referimos especialmente al proyecto MULTILEX[30] donde se pretendió lograr un estándar para la construcción de lexicones multilingües que garantizase la distribución de estos importantes recursos léxicos. El proyecto de ESPRIT MULTILEX se llevó a cabo con los siguientes objetivos[31]:

- La definición de estándares para la creación de un lexicón europeo multilingüe y multifuncional. Este estándar debería basarse en las

29. En otros casos, este tipo de regla morfológica para derivar significados no sería productiva como por ejemplo con los verbos *recall* o *represent*, cuyos significados no se corresponden con *'to call again'* o *'to present again'*.
30. MULTILEX, *Multilex WP9 Final Report: MLEXd*, 1993.
31. Krasemann, H. (coord.) Prop. 1212. *Implementation of the Eurotra-2 Workbench*. CAP GEMINI SCS. 1991, p. 6.

líneas estipuladas por ET-7, en los requisitos de las distintas empresas europeas incluidas en el proyecto y en los resultados de otros proyectos como GENELEX.

- El desarrollo de bases de datos léxicas siguiendo estos estándares, así como el desarrollo de un conjunto de herramientas de almacenamiento y recuperación sobre las bases de datos léxicas.

Lo realmente importante de este proyecto fue la consideración del lexicón de las lenguas naturales en términos de modelado de datos, lo cual implica la inclusión del concepto de 'arquitectura lingüística'. Según la definición del *Expert Advisory Group on Language Engineering Standards* (EAGLES)[32],

> The linguistic architecture defines the basic objects of the model and their relations. It also specifies the general terminology which is common to the whole standard and used to talk about dictionaries, their components and the interaction of these.

El modelo de MULTILEX fue diseñado desde un principio para ser multilingüe, debiendo dar cabida a todas las lenguas de la UE, por lo que la arquitectura general debería ser independiente de la lengua. Además el modelo debía ser multifuncional, capaz de integrar información utilizable por diversas teorías y aplicable a distintas aplicaciones de NLP. Por lo tanto, creemos que debería ser un punto de referencia para proyectos de construcción de lexicones computacionales.

Pensamos que futuros esfuerzos enfocados a la creación de lexicones computacionales deberían tener en cuenta las recomendaciones EAGLES con el objeto de asegurar su reutilización y posible distribución, sobre todo en lo que se refiere a la técnica de modelado de datos y la independencia de la teoría gramatica que se postula en ellas. Estas recomendaciones son lo suficientemente flexibles como para permitir la adaptación a las necesidades concretas de la aplicación para la que el lexicón esté destinado en principio, pero además va a permitir su reutilización en posibles aplicaciones y cooperaciones futuras.

## 3. Conclusión

El carácter indiscutiblemente interdisciplinar del lexicón hace necesaria la integración de técnicas, metodologías y resultados provenientes de los cam-

32. EAGLES EAGLES (Expert Advisory Group on Language Engineering Standards) Lexicon Architecture. Document EAG-CLWG-LEXARCH/B. October 1993, p. 14.

pos que hemos mencionado. El inmenso éxito de proyectos WordNet se debe principalmente a contar con una sólida base psicolingüística, pero también al hecho de que es totalmente reutilizable y distribuible.

En cuanto al apartado de adquisición de conocimiento léxico, pensamos que es menos determinante que el de representación, aunque igualmente importante. Como ya vimos, los mejores resultados se obtienen mediante la utilización de las diversas técnicas que hemos mencionado: la reutilización de recursos previamente existentes, los córpora textuales y el contraste de los datos así obtenidos con el conocimiento intuitivo del hablante nativo. Sólo mediante esta integración se conseguirán lexicones computacionales cuantitativa y cualitativamente válidos para su utilización en tareas complejas de NLP.

Hemos querido resaltar la centralidad del esquema de representación subyacente a la base de datos léxica. No podemos dejar de subrayar este aspecto, que ha impedido en muchos casos la reutilización de recursos léxicos muy valiosos. Al mismo tiempo, este modelo de datos debería adherirse, en lo posible, a los estándares existentes, y proponer mecanismos para su adaptación total a éstos.

No hemos profundizado demasiado en una de las repercusiones que la adopción de esta postura tiene con respecto a la lingüística: la necesidad de independencia de la información léxica respecto de la teoría gramatical o lexicológica empleada para el análisis léxico. Esto no significa que no deba emplearse una teoría sino que deben utilizarse modelos de datos estándar conjuntamente con interfaces que proporcionen *vistas* de los datos relevantes a la teoría con la que se esté trabajando. Esta metodología garantiza al lingüista trabajar y profundizar en una teoría lingüística determinada, al mismo tiempo que se garantiza la utilidad y aplicación práctica del fruto de su trabajo: la información lingüística.