

## LEXICOGRAFÍA COMPUTACIONAL Y LEXICOGRAFÍA DE CORPUS

CHANTAL PÉREZ HERNÁNDEZ  
*Universidad de Málaga*

ANTONIO MORENO ORTIZ  
*Universidad de Málaga*

PAMELA FABER  
*Universidad de Granada*

**RESUMEN.** *En este artículo ofrecemos una visión general de varias disciplinas de estudio relacionadas entre sí que han supuesto un cambio en las metodologías tradicionales de estudio lingüístico: la lexicografía computacional y la lexicografía de corpus. Enfatizamos la necesidad de emplear la evidencia de lengua en uso que puede derivarse del estudio de los corpórea textuales informatizados. Describimos a continuación algunas de las más destacadas herramientas computacionales de análisis de corpórea, en especial aquellas que pueden ser usadas en la compilación de diccionarios monolingües y bilingües. También tratamos los lexicones computacionales, así como los aspectos de la representación de conocimiento léxico que son relevantes para la clasificación que exponemos: diccionarios en formato electrónico, bases de datos léxicas y bases de conocimiento léxicas.*

**PALABRAS CLAVE.** *Lexicografía computacional, lexicografía de corpus, lexicones computacionales, diccionarios en formato electrónico, extracción, representación del conocimiento.*

**ABSTRACT.** *In this paper we provide an overview of a number of interrelated disciplines that have changed traditional methodologies in language studies, namely, computational lexicography and corpus lexicography. We stress the importance of deriving the description of a language from the naturally-occurring linguistic evidence that can be obtained through the analysis of corpora. We then describe some outstanding computational tools for the analysis of corpora and the ways they are used in the compilation of dictionaries, both monolingual and bilingual. Computational lexicons are also dealt with, as well as those aspects of lexical knowledge representation relevant for the classification we put forward: machine readable dictionaries, lexical data bases and lexical knowledge bases.*

**KEYWORDS.** *Computational lexicography, corpus lexicography, computational lexicons, electronic dictionaries, information retrieval, knowledge representation.*

## 1. LEXICOGRAFÍA COMPUTACIONAL Y LEXICOGRAFÍA DE CORPUS EN EL ÁMBITO DE LA LINGÜÍSTICA INFORMÁTICA

Para comprender los enormes avances realizados en los últimos veinte años en las dos disciplinas que nos ocupan, la *lexicografía computacional* y la *lexicografía de corpus*, es necesario tener en cuenta que ambas áreas de investigación han sufrido las limitaciones y se han beneficiado de los avances realizados en otra disciplina que las engloba: la *lingüística computacional*, la cual, a su vez, ha dependido siempre de los avances y tendencias en boga en la *lingüística teórica*.

En la relación entre *lingüística* (entendida como el estudio científico del lenguaje) y *lexicografía* no debemos olvidar la existencia de una rama de estudio que las enlaza, conocida como *lexicología*, que puede ser definida como el estudio científico del *lexicón* con el fin de revelar los principios que gobiernan su comportamiento y uso. Se puede considerar que mientras que el *análisis del lexicón* es la finalidad primordial de la lexicología, su *descripción* pertenece al dominio de la lexicografía, definida como el complejo proceso de compilación de diccionarios. Es decir, la *lexicografía* se ocupa de la descripción de una serie de fenómenos observables (el lexicón o vocabulario de una comunidad lingüística), los cuales define por medio de una serie de principios lingüísticos, tomados normalmente de la *lexicología* (Bennet et al. 1986: 4).

La *lingüística computacional*, tal y como se suele definir en los manuales introductorios a la materia, es el estudio de los sistemas de computación utilizados para la comprensión y la generación de lenguas naturales. Tres han sido tradicionalmente las aplicaciones principales de la lingüística computacional: la traducción automática (con una larga historia que parte de los años 50), la recuperación automática de información a partir de textos en lengua natural y la creación de interfaces en lengua natural hombre-máquina para la consulta de bases de datos (Grishman 1986: 15 y ss.).

Como ramas de la lingüística computacional, los términos *lexicología* y *lexicografía computacionales* se usan en muchas ocasiones como sinónimos<sup>1</sup> y, de hecho, se han desarrollado paralelamente en lo que se refiere a los avances tecnológicos, medios técnicos y desarrollo y aplicación de recursos computacionales para el estudio del lenguaje. Es de rigor sin embargo, destacar que los objetivos (si aplicamos estos dos términos en sentido estricto) son diferentes: la *lexicografía computacional* se refiere al uso de medios técnicos computacionales en los varios procesos que se siguen en la elaboración de un diccionario: desde que la primera idea parte del equipo editorial, pasando por decisiones que atañen a la macroestructura del diccionario (lista de lemas, orden, etc.), a su microestructura (el almacenamiento de la información durante el proceso de compilación de las entradas o los complejos medios de maquetado y edición en las fases posteriores). En este sentido, la praxis lexicográfica es más o menos computacional dependiendo del grado de tecnicidad de

la editorial, es decir, de los medios técnicos y las herramientas que pueda ofrecer a su equipo de lexicógrafos y editores. Hoy día los ordenadores se han convertido en herramientas de trabajo imprescindibles en todas las editoriales, al menos en las tareas que acabamos de señalar. Sin embargo, la proliferación de diccionarios en formato electrónico (*MRDs: Machine Readable Dictionaries*) y la introducción de los *cópora* textuales han ampliado enormemente el espectro de la *lexicografía computacional*.

Por otra parte, la *lexicología computacional* centra sus esfuerzos en la construcción de *lexicones computacionales* para el procesamiento del lenguaje natural. Los *lexicones* se consideran hoy día la base fundamental en la construcción de sistemas computacionales que posibilitan la interacción entre la máquina y el hombre. La importancia y centralidad del *lexicón* computacional en las aplicaciones de procesamiento de lenguaje natural 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 (1994: 267) 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 (1991: 205)

... [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.

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 (Boguraev & Briscoe 1989; Pustejovsky 1991). Esto ha provocado una demanda constante de información detallada sobre amplias áreas de vocabulario. La finalidad fundamental del procesamiento de lenguaje natural es la automatización de procesos lingüísticos, tales como la comprensión, producción o adquisición de una lengua, tareas que, por otra parte, 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 (1991: 3), durante años, los *lexicones* enfocados al procesamiento del lenguaje natural han sido los “hermanos pobres” de la lingüística computacional.

La mayoría de los sistemas diseñados hasta hace relativamente poco tiempo contenían sólo *lexicones* ilustrativos con no más de cien palabras<sup>2</sup> y, a pesar de los

numerosos avances en esta 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 deber ser representada. El conocimiento lingüístico que requiere un usuario “humano” y el que requiere un usuario “máquina” es totalmente diferente, de ahí que *lexicología* y *lexicografía computacionales*, a pesar de que se suelen usar como sinónimos, no sean exactamente lo mismo. La tarea de construir un lexicón completo para una lengua natural es enorme. El *Oxford English Dictionary* (OED), por ejemplo, contiene 250.000 entradas, 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”, y esto 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 semiautomática en bases de datos léxicas (información *fonológica, morfológica, sintáctica, semántica y pragmática* que se encuentra en los diccionarios en mayor o menor medida).

Tal y como veremos en la sección 3, se pueden distinguir dos grandes ámbitos de investigación en lo referente a la creación de lexicones computacionales: el de la *adquisición* y el de la *representación* de conocimiento léxico. El primer término suele ser empleado en empresas de reutilización de recursos existentes, normalmente diccionarios en formato magnético, 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, éstas son las dos fases principales contempladas en la construcción de un lexicón computacional y se pueden considerar como separadas pero interdependientes, por lo que repasaremos las metodologías más destacadas aplicables en cada una de estas dos fases.

Centrando nuestra atención en la otra disciplina que nos ocupa, la *lexicografía de corpus*, es indudable que no se puede entender su existencia sin tener en cuenta los postulados básicos de la *lingüística de corpus* y las conexiones que ambas poseen con la *lingüística* y la *lexicografía computacionales*.

En los últimos veinte años tanto lingüistas como lexicógrafos han sido testigos del resurgimiento de los métodos empíricos y estadísticos de análisis lingüístico, típicos de la década de los cincuenta (Church y Mercer 1993). En aquellos años era práctica común, por ejemplo, el estudio de las unidades léxicas basándose no sólo en su significado sino también en su concurrencia con otras palabras<sup>3</sup>. Debemos recordar que también en los años cincuenta, J. R. Firth, una figura eminente dentro de la tradición lingüística británica, publicaba *Papers in Linguistics*, donde el enfoque dado al estudio del lenguaje se resumía con la famosa frase “*you shall know a word for the company it keeps*” (Firth 1957: 11). Este interés empírico se desvaneció a finales de los años cincuenta, debido sobre todo a las críticas que Chomsky realizó a los métodos

empíricos e inductivos, dando paso a un largo periodo de estudios lingüísticos de carácter mentalista.

Sin lugar a dudas, la razón más poderosa para el actual resurgimiento de los estudios de corte empírico es la disponibilidad creciente de cantidades masivas de texto en formato magnético. Hasta hace sólo diez años, el corpus de un millón de palabras creado por Francis y Kúcera en la Universidad de Brown parecía enorme. Hoy por hoy, muchos centros de investigación poseen córpora que contienen cientos de millones de palabras.

La investigación basada en corpus ha supuesto el nacimiento de nuevos métodos de estudio en áreas de estudio tan diversas como la adquisición de conocimiento léxico, la construcción de gramáticas, los estudios socioculturales, la estilística, la traducción automática, el reconocimiento del habla, la recuperación de información, la construcción de diccionarios electrónicos o la compilación de lexicones computacionales y bases de datos terminológicas. Este tipo de investigaciones se ha desarrollado en las dos últimas décadas de tal forma que, desde hace más de quince años, está empezando a considerarse una disciplina de estudio en sí misma, conocida como *lingüística de corpus* (o *del corpus*, ya que en español no parece existir consenso sobre su denominación), con la *lexicografía de corpus* como disciplina en desarrollo paralelo, aunque aún existan algunos académicos que se muestren reticentes a considerarla como una disciplina de estudio autónoma:

(...) but is corpus linguistics really comparable with these other hyphenated branches of linguistics? (socio-linguistics, psycholinguistics, text linguistics) No, because "corpus linguistics" refers not to a domain of study, but rather to a methodological basis for pursuing linguistic research (...) (Leech 1992: 105).

Lo cierto es que no es una disciplina unitaria, cuyos fines y métodos se presten a un fácil acotamiento. El hecho de que disciplinas tan variadas como las citadas anteriormente se sirvan de un corpus lingüístico informatizado para sus fines particulares ha llevado a algunos investigadores a considerar el corpus como una herramienta de apoyo o como una simple metodología de análisis. Esta argumentación puede ser apropiada en algunos casos, como por ejemplo la traducción automática, donde un corpus (normalmente uno paralelo, es decir, un texto y su traducción) se usa para obtener equivalentes de traducción de forma (semiautomática (Brown *et al.* 1990; Klavans y Tzoukermann 1990; Gale y Church 1993). Sin embargo, existe un ámbito de estudio en el que sí nos parece justificado hablar de *lingüística y lexicografía de corpus*. Nos referimos a aquellos casos en los que el corpus se usa para derivar de su estudio descripciones lingüísticas detalladas, ya sean con fines computacionales, teóricos o lexicográficos.

Considerándola como disciplina unitaria o no, es indudable que existen muchas publicaciones destacadas que nos animan a pensar que se encuentra en proceso de

establecerse como disciplina independiente, como por ejemplo el *International Journal of Corpus Linguistics*. Se cuenta ya, por otra parte, con publicaciones orientadas a asentar los presupuestos teóricos y metodológicos de la lingüística de corpus (Tognini-Bonelli 1996; Lager 1995), y se han publicado en los últimos diez años numerosos libros en los que se recogen artículos y actas de congresos que muestran las líneas de investigación basadas en corpus más destacadas, llevadas a cabo tanto en diversas universidades a ambos lados de océano, como en importantes centros de investigación, como los de IBM o AT&T. Entre estas publicaciones merecen especial mención las actas de los congresos organizados anualmente desde 1985 por el *Centre for the NEW OED and Text Research* en la Universidad de Waterloo (Ontario, Canadá), las actas del congreso sobre lexicografía computacional *Complex* (Kiefer, Kissi y Pajzs 1992), o los volúmenes especiales dedicados al corpus de las revistas *Literary and Linguistic Computing* (Ostler 1993), *Computational Linguistics* (Church y Mercer 1993), y el *International Journal of Lexicography* (Sinclair, Payne y Pérez 1996). Merecen ser destacadas también las numerosas recopilaciones en forma de libro que recogen contribuciones de diversos autores publicadas en los últimos años, como por ejemplo Baker, Francis y Tognini-Bonelli (1993); Hoey (1993); Svartvik (1992); Aarts, de Haan y Oostdijk (1993); Oostdijk y de Haan (1994); Boguraev y Pustejovsky (1996), Biber *et al.* (1998) o las publicaciones de carácter pedagógico, como por ejemplo McEnery y Wilson (1996) y Stubbs (1996), Kennedy (1998).

Las investigaciones basadas en corpus, tanto lingüísticas como lexicográficas, se han centrado mayoritariamente en la lengua inglesa, aunque en los últimos años se han promovido varias iniciativas para la construcción y uso de córpora en otras lenguas, sobre todo las pertenecientes a la Unión Europea y a algunos países del Este. De entre las publicaciones dedicadas al uso de corpus en lengua española cabe destacar Alvar y Villena (1994), Sánchez *et al.* (1995) y el informe llevado a cabo por el Observatorio Español de Industrias de la Lengua del Instituto Cervantes sobre recursos lingüísticos del español (Instituto Cervantes 1996).

Uno de los postulados básicos de los estudios basados en corpus es que la lengua debe estudiarse a través de ejemplos reales de uso, es decir, a partir del estudio de un corpus de texto informatizado, considerando el corpus como una muestra representativa del uso que los hablantes nativos hacen de una lengua. Se debe tener en cuenta también que un corpus se puede usar de formas muy diferentes, ya sea para *validar*, para *ejemplificar* o para construir una teoría de la lengua y los diferentes aspectos que ésta implica. Este hecho se hace patente en las diferentes denominaciones (con sus correspondientes diferencias teóricas y metodológicas) que se usan para referirse al uso de los córpora en la investigación lingüística: *corpus-based*, *corpus-driven*, *data-driven* y *text-analysis*, por nombrar sólo las más comunes. Estas diferencias en cuanto al uso que se hace de los córpora se corresponden también con posturas diferentes en lo que se refiere a aspectos fundamentales que se han de tener

en cuenta para considerar el corpus como una muestra representativa de la lengua de estudio: creación y diseño de corpus, tipo y forma de análisis, explotación y desarrollo de herramientas que lo manejan, tipo y cantidad de información metatextual que el corpus debe contener y, sobre todo, el grado de compromiso con la información que se deriva del corpus.

## 2. LEXICOGRAFÍA DE CORPUS

Definíamos la lexicografía en el apartado anterior como la descripción del vocabulario de una lengua, materializada en el complejo proceso de compilación de diccionarios. Esta descripción se hace por medio de una serie de principios lingüísticos, tomados normalmente de la lexicología (Lipka 1990; Tomaszczyk y Lewandowska 1990), y metodológicos, recogidos normalmente en manuales y publicaciones de lexicografía teórica (Alvar 1983; Haensch *et al.* 1988; Hausmann *et al.* 1989, 90, 91; Hartmann 1983; Chrisholm 1993, etc.). La teoría y práctica de la lexicografía implica múltiples aspectos de entre los que, para resaltar la importancia y utilidad de los corpóra, nos centraremos fundamentalmente en la obtención de información para la compilación de las entradas del diccionario.

Los lexicógrafos siempre han buscado fuentes de información para obtener la información necesaria para la descripción lingüística. Estas fuentes han sido, tradicionalmente (i) la intuición, (ii) otros diccionarios, (iii) fuentes tradicionales de recopilación manual de información sobre el uso de las palabras (citas de autores reconocidos, periódicos, libros, etc.) (Sinclair 1993). Por muy usual que haya sido durante siglos, confiar la descripción lingüística solamente en estas tres fuentes acarrea una serie de problemas.

La primera de ellas, el conocimiento intuitivo del lexicógrafo, plantea los mismos problemas que se han señalado en numerosas ocasiones en referencia al estudio lingüístico general. Sinclair, por ejemplo, ha recalcado en numerosos trabajos (1987b, 1991, 1992a, 1996, *inter alia*) las posibles inconsistencias e inexactitudes de las intuiciones lingüísticas, considerando incluso algunos casos en los que el hablante nativo simplemente no puede poseer el conocimiento intuitivo suficiente para postular una parte de la teoría o para describir el comportamiento de una palabra o unidad lingüística. Las introspecciones del lexicógrafo pueden no ajustarse a la realidad, o al menos a lo que es más frecuente en el uso de la lengua. Confiar sólo en la introspección puede llevar al lexicógrafo a no darse cuenta de ciertas regularidades en el uso o significado de las palabras, o a pasar por alto estructuras sintácticas o colocaciones que son relevantes y deben incluirse en el diccionario.

La segunda de las fuentes de información (*otros diccionarios*), plantea problemas de otra índole, aunque esta práctica es mucho más frecuente y tácitamente aceptada de lo que pueda parecer a primera vista. No debemos olvidar que las

descripciones lingüísticas hechas durante décadas son sin duda muy valiosas y acumulan gran cantidad de información que, sin duda, no se puede desdeñar a priori. Por otra parte, debemos tener en cuenta que, de este modo, es muy difícil asegurarse de que no se siguen incluyendo en diccionarios usos o acepciones obsoletas (al menos sin indicarlo expresamente), o que no se incluyen distinciones de significado que se han incluido durante décadas en los diccionarios pero que no se ajustan a la realidad del uso de los hablantes. El tercer problema que plantean descripciones lexicográficas anteriores es que, por supuesto, no constituyen una fuente de información apta para realizar una descripción actualizada de la lengua de estudio.

La tercera de las fuentes, la *recopilación manual de citas*, es un trabajo valiosísimo a la vez que tedioso y muy limitado, ya que sólo suelen recogerse citas que dan cuenta de curiosidades lingüísticas o usos que han llamado la atención del lexicógrafo.

Con la introducción del uso de los corpórea textuales informatizados, la calidad (y cantidad) de análisis lingüístico que los lexicógrafos pueden llevar a cabo en el proceso de compilación del diccionario se ha multiplicado de forma magnífica. La *lingüística de corpus* ha hecho patente la importancia de derivar la descripción lingüística de un análisis detallado de lengua usada de forma natural, ya que este estudio puede ayudar a revelar muchas regularidades (e irregularidades) en nuestro uso de la lengua que antes no se habían observado, o pueden ayudarnos a verlas de forma más uniforme, con una perspectiva más amplia y con índices de frecuencia relativa más fiables. De hecho, la introducción del uso del corpus en la praxis lexicográfica tiene ya una historia de casi veinte años, compartiendo en muchos casos recursos informáticos, técnicas y proyectos de investigación con la *lingüística de corpus*<sup>4</sup>, ya que las necesidades de los lexicógrafos como estudiosos de la lengua y su uso no difieren, al menos en los aspectos más básicos, de las de los lingüistas, sobre todo en lo que respecta a las fuentes de información para la extracción de conocimiento lingüístico.

La iniciativa pionera en la introducción del uso del corpus en la compilación de diccionarios fue la formada por la Universidad de Birmingham y la editorial Collins (en la actualidad Harper-Collins), conocida como COBUILD (*Collins Birmingham University International Language Database*). El diccionario *Collins Cobuild Dictionary of English Language* supuso, sin duda alguna, una revolución no sólo en el mundo editorial, sino que tuvo además una gran repercusión en otros ámbitos de estudio lingüístico y lexicológico. Las contribuciones recogidas en Sinclair (1987b) detallan varios aspectos del proceso de construcción del corpus, la creación de la base de datos y la posterior compilación del diccionario.

Lo más destacable e innovador de ese proyecto fue que, por primera vez, un diccionario se compilaba por medio del examen detallado de un corpus representativo de textos ingleses, orales y escritos (de 20 millones de palabras)<sup>5</sup>. Esto significaba, en palabras de su editor jefe, el Profesor John Sinclair, que además de las herramientas



con las que los lexicógrafos han contado durante años (es decir, un profundo conocimiento de la lengua y muchas lecturas, otros diccionarios y por supuesto ojos y oídos), este diccionario está basado en evidencia mensurable (Sinclair 1987a: XV).

Los lexicógrafos de Cobuild trabajaron durante siete años analizando el corpus para extraer de él información sobre el significado de las palabras, su uso, los patrones sintácticos que caracterizaban cada una de las diferentes acepciones y estudiar las colocaciones más frecuentes y que, por tanto, debían ser incluidas en un diccionario dirigido a los estudiantes de inglés. Este diccionario fue innovador en otros muchos aspectos, ya que la estructura de las definiciones y la organización de las entradas se aparta bastante de la praxis lexicográfica tradicional y la estructura de las entradas también es diferente.

Cobuild fueron los pioneros en el uso de los corpórea textuales informatizados aunque hoy en día, casi todas las editoriales importantes también ha adoptado su uso, en mayor o menor medida, en el proceso de compilación de los diccionarios. Tanto Oxford University Press como Addison-Wesley Longman y Larousse Kingsfisher Chambers han colaborado activamente en la creación del BNC (*British National Corpus*)<sup>6</sup>, Cambridge University Press ha basado su nuevo diccionario CIDE (*Cambridge International Dictionary of English*) en un corpus de 100 millones de palabras (Baugh, Harley y Jellis 1996) y en España, varias editoriales también cuentan con corpórea de diferentes tamaños y características: Vox Bibliograf posee un corpus de 10 millones de palabras, la editorial SM uno de 60.000 y la editorial SGEL posee el corpus CUMBRE, de 8 millones de palabras, cuya creación y uso se detalla en Sánchez et al. (1995). Esta inversión, tanto de recursos económicos como humanos, nos parece muy significativa del esfuerzo realizado por diversas editoriales, encaminado a extraer la información de sus diccionarios de corpórea textuales informatizados y su utilidad se hace patente en el hecho de que sus editores incluyan en las introducciones a sus diccionarios frases como “This magnificent new resource [BNC] has enabled us as never before ... to present a wholly accurate picture of the syntactic patterns of today’s English” (Jonathan Crowther, prefacio de la edición de 1995 del *Oxford Advanced Learner’s Dictionary*) o “the larger corpus [*The Bank of English*] enables us to make statements about the meanings, patterns, and uses of words with much greater confidence and accuracy of detail” (John M. Sinclair, introducción de la edición de 1995 de *Collins Cobuild English Dictionary*).

Para la mayoría de los lexicógrafos, los corpórea se han convertido en una herramienta lexicográfica fundamental para el estudio de las diferentes acepciones de una palabra que han de incluirse en las entradas léxicas y para el estudio de las colocaciones y la fraseología (véase, por ejemplo, los estudios contenidos en Sinclair 1987b, 1992; Sinclair y Kirby 1990; Clear 1993, 1994; Sánchez et al. 1995; Baugh, Harley y Jellis 1996). También ofrecen información decisiva sobre las diferencias de uso entre la lengua oral y la escrita y la frecuencia relativa de uso tanto de determinadas palabras, como de determinadas acepciones de una palabra, información

clave para la inclusión (o exclusión) de una entrada o una acepción en un diccionario. Las referencias a estudios y artículos sobre estos aspectos son innumerables, destacamos algunas fundamentales, como Hanks (1987, 1993); Atkins, Kelg y Levin (1986, 1988); Moon (1987); Atkins (1987, 1992, 1993) y Rayson, Leech y Hodges (1997).

A través del análisis exhaustivo de grandes cantidades de texto computerizado los lexicógrafos pueden obtener información indispensable sobre la gramática, las relaciones semánticas, la aceptabilidad de determinados usos, usos innovadores u obsoletos de palabras, palabras o expresiones de nueva creación, e incluso aspectos pragmáticos (véase, por ejemplo, Aarts 1991; Moon 1994; Hanks 1996). En este sentido, la macroestructura de los diccionarios ha cambiado notablemente en los últimos diez años. Cada vez se incluye más información *sobre* la lengua y su uso mientras que otro tipo de información que, quizás por tradición lexicográfica, seguía incluyéndose, como los libros de la Biblia, etimologías o tablas de conversión de monedas y mapas están empezando a desaparecer.

De igual importancia que en la lexicografía monolingüe, el uso del corpus es determinante para la creación de mejores, más completos y útiles diccionarios bilingües, que guíen al usuario de forma acertada en el proceso de la traducción de una a otra lengua o en la comprensión de un texto en lengua extranjera.

En cualquier caso, un corpus no es de gran utilidad si el lexicógrafo no cuenta con las herramientas de análisis adecuadas, que le permitan procesar el texto de formas diferentes y le ofrezcan un alto nivel de flexibilidad en el tipo de búsquedas que pueda realizar. Pasamos a continuación a ver algunas de las herramientas más usadas.

### *2.1. Análisis cualitativo y cuantitativo: herramientas computacionales para el tratamiento y explotación de los córpora informatizados*

Se suele hacer una distinción entre dos tipos generales de análisis del corpus: *cualitativo*, en el que se hace una descripción detallada y completa de un fenómeno lingüístico o del comportamiento de una palabra o grupo de palabras y *cuantitativo*, en el que se asignan índices de frecuencia a los fenómenos lingüísticos observados en el corpus y éstos pueden servir para construir modelos estadísticos más complejos que expliquen la evidencia hallada en el texto. Estos dos tipos de análisis no deben considerarse como excluyentes, sino más bien como complementarios, ya que el análisis cualitativo, por un lado, ofrece una gran riqueza y precisión en las observaciones realizadas y los fenómenos poco frecuentes pueden recibir igual atención que los muy frecuentes; por otro lado, el análisis cuantitativo puede ofrecer al lingüista o lexicógrafo información que sea estadísticamente significativa y resultados que pueden considerarse generalizables (McEnery y Wilson 1996: 63), por lo que es

hoy muy frecuente que se combinen ambos tipos de análisis. Mario Bunge argumenta al respecto (1995: 3):

There can be no opposition between quantitative and qualitative methods, since quantity and quality are mutually complementary rather than exclusive. Indeed, every quantity is either the numerosity of a collection of items sharing a certain quality, or the intensity of a quality. Hence, in the process of concept formation, quality precedes quantity.

La mayoría de los paquetes informáticos que se han desarrollado en los últimos años ofrecen la posibilidad de llevar a cabo ambos tipos de análisis, y en este sentido se han hecho enormes progresos y han aparecido diversas publicaciones que sirven de guía para el análisis estadístico con fines lingüísticos o lexicográficos (Butler 1985; Fielding y Lee 1991; Charniak 1993; Wilks, Slator y Guthrie 1996). Existe también en el mercado un importante número de programas (tanto comerciales como de libre distribución en ámbitos académicos) con interfaces de usuario muy fáciles de manejar y a la vez muy versátiles y sofisticados, aunque la mayoría de las grandes editoriales han desarrollado herramientas de análisis específicas para su propio corpus y que por tanto se adaptan perfectamente a cualquier tipo de información metatextual que se haya añadido (información sobre las clase morfológica de las palabras, información sintáctica, identificación del texto y especificaciones sobre su procedencia, tipo o variedad lingüística a la que pertenece, etc.) y además son herramientas que suelen ir adaptándose y desarrollándose para satisfacer las necesidades específicas de sus lexicógrafos.

Algunos programas de manejo de corpus disponibles se distribuyen de forma gratuita para ser usados con fines académicos (por ejemplo *Conc*, del *Summer Institute of Linguistics*; *FreeText Browser*, de la Universidad de Michigan o *TACT*, del departamento de *Computing in the Humanities and Social Sciences* de la Universidad de Toronto). Dentro de los programas comerciales, los más usados han sido tradicionalmente *Oxford Concordancing Program*, *MicroConcord* (ambos de *Oxford University Press*) y *WordCruncher* (*Wordcruncher Publishing Technologies*), junto con un nuevo conjunto de herramientas para el manejo de corpus desarrollado por Michael Scott para *Oxford University Press*, conocido como *Wordsmith Tools*<sup>7</sup>.

Casi todos estos programas nos ofrecen las herramientas básicas de manejo de corpus, como por ejemplo la capacidad de realizar *listados* de las formas (*types*) que aparecen en un corpus, ordenados de diferentes maneras, ya sea por orden alfabético, frecuencia, o en algunos casos por orden alfabético inverso e índices estadísticos sobre el número de palabras, oraciones o párrafos y la longitud de éstos. Estos listados pueden ser de gran utilidad lexicográfica, ya que pueden ayudar a decidir la lista de voces que han de incluirse en un diccionario, teniendo en cuenta su frecuencia de uso, o para decidir qué vocabulario básico debe incluir un diccionario escolar. También pueden ofrecernos índices de frecuencia en los que muestre la ratio *formas/palabras* (*types/tokens*), es decir el número total de palabras de un texto frente al número de

palabras *diferentes* que a parecen en el mismo o comparar los índices en varios ficheros de texto, tal y como aparece en la figura 1, en el que se muestra una captura de pantalla tomada del programa *Wordsmith Tools* en el que se compara la lista de palabras y la ratio forma/palabra de dos ficheros de texto diferentes. Este tipo de cálculos pueden ser fundamentales para establecer el grado de representatividad del corpus que estamos usando. Sánchez y Cantos (1997), por ejemplo, desarrollan un procedimiento estadístico para predecir la relación entre formas y palabras en un corpus, de forma que éste puede subdividirse en secciones más pequeñas o subcórpora, que son más fáciles de manipular y analizar pero que guardan la estructura y la consistencia interna del corpus completo y que son similares en lo que respecta a variación lingüística y a variabilidad.

N	1	2	3
Text File	OVERALL	SPANISH2.TXT	SPANISH1.TXT
Bytes	2.260.341	1.266.411	993.930
Tokens	354.974	206.597	158.377
Types	31.476	22.071	19.683
Type/Token Ratio	8.62	10.68	12.43
Standardised Type/Token	48.02	48.22	47.75
Ave Word Length	4.73	4.72	4.85
Sentences	10.217	6.090	4.127
Sent. length	33.81	32.23	36.14
sd. Sent. Length	33.61	33.11	34.21
Paragraphs	521	261	260
Para. length	646.53	724.34	568.43
sd. Para. length	414.07	444.38	365.75
Headings	0	0	0
Heading length			
sd. Heading length			
1-letter words	20.287	11.978	8.309
2-letter words	89.542	50.695	38.847
3-letter words	51.255	29.618	21.637
4-letter words	32.638	18.976	13.662

Fig. 1. Comparación de índices de frecuencia de dos ficheros realizado con Wordsmith Tools.

Tanto *Wordsmith Tools* como *TACT* cuentan con una serie de herramientas para preprocesar el texto antes del análisis. Estas herramientas nos permiten, por ejemplo, añadir etiquetas morfosintácticas (*tags*) al texto, a partir de un diccionario creado con las formas extraídas del texto, lematizar el texto, asignando diferentes formas a una misma forma canónica, o crear una lista de palabras que, por ejemplo, dada su alta frecuencia no queremos incluir en nuestra búsqueda (*Stop Words Lists*).

Otra de las herramientas de manejo de corpus más importante y versátil para la lexicografía son los programas que nos proporcionan de forma automática líneas de *concordancia* de una palabra. Una concordancia, normalmente llamada KWIC (*Key Word in Context*) es una colección que recoge todas las apariciones de una palabra en un texto o conjunto de textos, junto con un número determinado (normalmente por el lexicógrafo) de caracteres de co-texto anterior y posterior (la palabra que se está

estudiando o *nodo*, suele aparecer en medio, resaltada en pantalla con un formato o color diferente). De esta forma el lingüista puede ver a la vez una gran cantidad de ejemplos de uso de una palabra o un grupo de palabras. Las posibilidades de trabajo con las líneas de concordancia dependerán en gran medida del paquete informático que estemos manejando. La mayoría de ellos nos permitirán obtener un número determinado de líneas (100, 200, o todas las que aparezcan en el texto) y ordenarlas posteriormente de diferentes maneras: alfabéticamente, de acuerdo con la palabra inmediatamente anterior o posterior al nodo o en relación a la palabra que aparezca dos, tres, etc. posiciones a la derecha o izquierda de nuestro nodo (el nodo también puede ser, a su vez, una sola palabra o un grupo de palabras). Algunos programas están limitados en cuanto al número de líneas de concordancia que pueden ofrecernos, como por ejemplo *Micro Concord*, que al servirse únicamente de la memoria convencional de DOS, suele limitar el número de líneas que puede extraer a una cifra entre 1500 y 1700. La figura 2, por ejemplo, es una captura de pantalla que muestra algunas líneas de concordancia de la palabra inglesa "term" (ordenadas según la primera palabra que aparece antes del nodo), extraídas con la herramienta *Concord* de *Wordsmith Tools*:

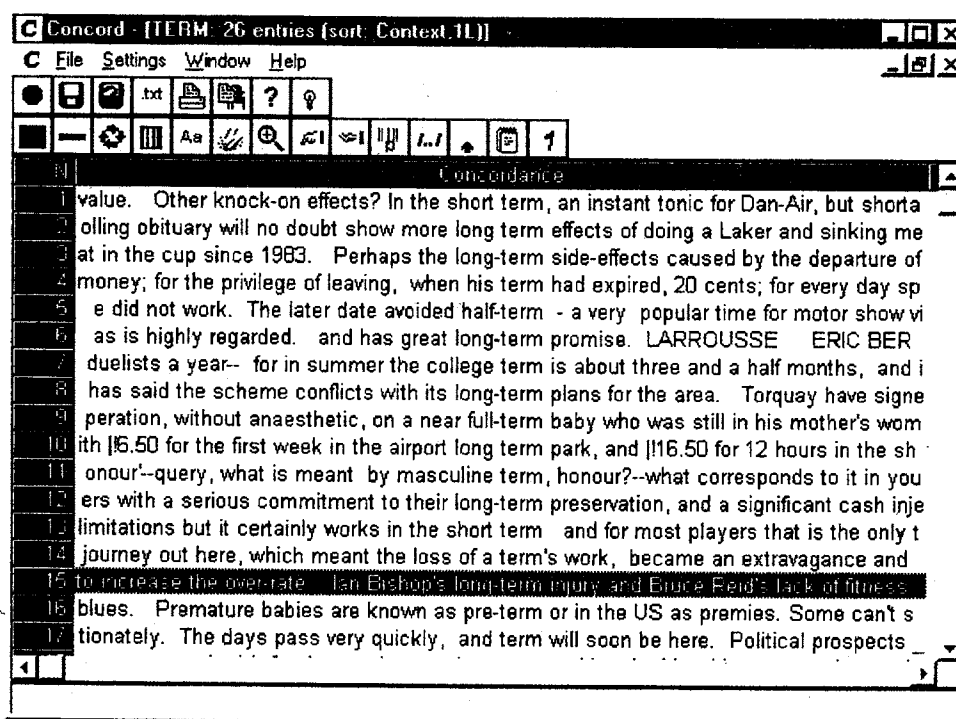


Fig. 2. Líneas de concordancia extraídas con la utilidad *Concord* de *Wordsmith Tools*.

Estos diferentes tipos de ordenación permitirán al lingüista o lexicógrafo centrar su atención en el co-texto inmediatamente anterior o posterior de la palabra (por ejemplo, para estudiar tipos comunes de sujetos y complementos en el caso de un verbo), o en el tipo de modificación adjetival que acompaña a un sustantivo

determinado o, al revés, el tipo de sustantivos a los que acompaña un adjetivo determinado. Muchos de estos programas permiten el uso de caracteres comodines (*wildcards*), con los que se pueden buscar diferentes formas de una misma palabra o realizar búsquedas difusas, múltiples y de frases idiomáticas con un cierto grado de variación. Con la mayoría de los programas que existen en el mercado, también podremos identificar la fuente original de una línea de concordancia determinada, ampliar el co-texto o acceder al texto original al que un ejemplo determinado pertenece. Los ficheros de líneas de concordancia pueden almacenarse en el ordenador para después editarlos y manipularlos con un procesador de texto. Como decimos, todas estas posibilidades dependerán del paquete informático que se use, ya que algunos son más limitados que otros, tanto en la cantidad de texto que pueden manejar a la vez como en la variedad de análisis que ofrecen. La figura 3 muestra algunas líneas de concordancia de la palabra inglesa “term”, extraídas con *Micro Concord (OUP)*:

MicroConcord search SW: term

90 characters per entry

Sort : SW/1L

r: low grade facilities destroyed in Iraq. Long-term contamination in small areas likely. ng to a 120mph wind, so you don't hear it. Short-term memory also comes low on the list - ney supply, and thus inflation. The Medium Term Financial Strategy as it was dubbed and means that the ERM has replaced the Medium Term Financial Strategy as the bedrock of the G llow the international players to enjoy an end-of-term pillow fight. In South-west London of innocence Michael Henderson on the end-of-term attractions of the Rosslyn Park Sev operation, without anaesthetic, on a near full-term baby who was still in his mother's wo heme did not work. The later date avoided half-term - a very popular time for motor sho f the club." Us, name-drop? Surely not. A long-term injury to Brian Gayle has prompted Ba t to increase the over-rate. Ian Bishop's long-term injury and Bruce Reid's lack of fitne (Even better, stretch it into Sunday.) In long-term, relationships, hanging on means that e a cool !!5.5 billion in a full year. The long-term success of this Budget may well depen ny other new entrant into the market. The long-term aim would be to have numbers that ide accident." So when you finally get into a long-term relationship . . . "Who knows. There' ndship and communication. You could meet a long-term lover in a bar. Many of my friends ar hs had been dropping amorous hints about a long-term relationship upped and disappeared to work here if only someone will commit to a long-term investment,' says Mark Edwards, a twe nomy with wider ownership of wealth; and a long-term commitment to future generations whic tral striking role, when Agana picked up a long-term injury, and has responded with eight r have a realistic shot at developing into long-term love. Part of negotiating your way th u find it difficult to get aroused by your long-term partner and it may be necessary to dr ten minutes to lodge then securely in your long-term memory. 4 Remember pictures. Instead fficult for chemical companies to tackle a longer-term difficulty facing up to the envir market tightens, anyone willing to take a longer-term view will be able to take advantage education budget. But the most effective longer-term measure for a green Budget would be

for the next thing. Whatever that is, his longer-term ambition seems to be increasingly clear and should not distract policy-makers from the longer-term, underlying issue. Britain's key economic term planning. Let me see <ZGY> this. No medium-term planning let me get it right. This and disrupted agriculture could be the main medium-term effects, say the World Conservation is yet to decide whether there should be a mid-term election and, if so, who would head the thing fashionable to see him as a potential one-term President. He'd had a tremendous start baby blues. Premature babies are known as pre-term or in the US as premies. Some can't suffer distortions from savings allocation, not a short-term fix hiding behind false claims that at by Richard Branson, he was engaged on a short-term contract by BA chairman Lord King, whose evidence of the office is worth more than a short-term political gain.' It's just the Tories and a chains will do much good. They are after short-term profit, and Hollywood makes that for much larger and a much longer trial, and any short-term benefits from early AZT treatment do have stopped launching into niches to make short-term profits. And it is thought that both of them then she says We need to standardize our short-term planning and this is what we're doing diagnosed as having Korsakoff's Psychosis - short-term memory loss resulting from long-term alcoholism. There are advantages for those who want some short-term guarantee of interest-rate stability and the ease of payments encourages this view: in the short-term, at least, Britain's foreign exchange market is interested in returning to quotas in the short-term nor in changing the market monitoring to clarify the discussion, let us use the British term of 'personal allowances' to refer to this.

"What is 'GNILLIC'?" "That is the Eskimo term for 'snow.'" "So you knew the English for snow, be they complex, real or rational. The term modern algebra can then be used to describe the page there is reference to a "bistro", a term which apparently did not come into use in the 19th century. I expected that scientists should have coined a term for so ostensibly unscientific a pursuit, such as a rock band. Parallax, otherwise a term which in physics relates to the change in position of an object for the contemporary use of "bitch" as a term of endearment is culled from the letters of the word bitch, which is now dead. Instead, "luggable" is used as a term of abuse for portables that ought to come in a suitcase. At trial judges should be free to impose a term of imprisonment that they believe fitted the crime. However, now, the "real books" phrase has become a term of disapproval, a convenient shibboleth for those who wonder if their party deserves another term of office. And, frankly, more of the same is in the air. The team leader probably in the autumn term will be language and music that will be interesting. I checked, a stonker was a colloquial term for what a man gets if he's on a long train journey. I rents. I also go to boarding school during term time. All this makes it impossible for me to have a holiday. My first term as Environment Secretary, is well established (1981). Michael Heseltine, in his first term as Environment Secretary, also warned the public that he would not seek a second five-year term of the institution that provides more than 100 years of continuous service. The Conservative Party had been elected for a fourth term and appears to have convinced itself that it was a general reference point, but as a general term of abuse. Intermittently they stamp on the advantage of the defendant, and in the long term to the police as well. While there is no doubt that the practice is widespread and which may have unknown long term effects. Amniocentesis involves extracting fluid from the fetus. The Government should reduce the number of offenders, said: "In the long term the Government should reduce the number of offenders to a level that would allow it to rely to offset inflation and in the longer term it offers enormous opportunities. It's the only way to reduce the extra indexing of the threshold will be necessary. As the government says "higher petrol prices in the longer term are both necessary and unavoidable". No petrol price rises each month. At the end of the mortgage term the borrower owes nothing. Repayment mortgage each month, and at the end of the mortgage term (usually 25 years) the debt will be cleared.

<p>ve jobs we enjoy, a daughter in her second term at university and a 14-year-old son at home still looked like a shoo-in for a second term in the White House, Margaret Thatcher was</p> <p>idn't invest in new players just for short term success, although that is what we wanted.</p> <p>ad of England. It looks like, in the short term at least, he made the right choice. "Beati</p> <p>originally an early 20th century US slang term of abuse either for any lesbian or for any</p> <p>the noise of closing doors, in the spring term of the first year of the "sixth-form" co</p> <p>t school only during the present teacher's term of office. The regular scholars, if the tr</p> <p>is exobiology? In fact, it's the technical term for the study of alien life in space. And</p> <p>. Since the turn of the year, though, the term "expansive" has disappeared from Cooke's</p> <p>ests, on the manner and sense in which the term is used: positive and life-affirming? or n</p> <p>r). He seems unaware that criticism of the term "adolescence" in relation to sexual behav</p> <p>c violence yes I agree with the use of the term domestic violence but this is not the term</p> <p>ow where they were. I don't know where the term "junglist" (hardcore Techno's dominant st</p> <p>inal festival band (provided you allow the term "original" a certain latitude). The field</p> <p>ton, Cheshire Commonplace criticism of the term 'adolescence' MARK SHIMPSON accused me of</p> <p>modern classic from the man who coined the term 'homophobia'. LESBIAN Somewhere Like This</p> <p>azz Dancing, Class of '89') redefines the term 'engaging': as in, I'd like to engage his</p> <p>h and many men did not even understand the term 'sexual intercourse' used in the titles".</p> <p>y camps and high camp. Black comedy is the term which springs to mind, but it would be mis</p> <p>e Government has been careful to avoid the term "victory" in relation to these events. I</p> <p>as a proxy for an academic record; but one term's work cannot yield "evidence" of potent</p> <p>ministrative legal area), during one legal term, there were 24 travelling days listed. Mul</p> <p>embarrassment causes blushing. In the long term, emotional stress affects the pituitary gl</p> <p>gger' and 'faggot'. It's such a pejorative term, one I associate with a mode of thinking,</p>	<p>term at university and a 14-year-old son at home still looked like a shoo-in for a second term in the White House, Margaret Thatcher was</p> <p>idn't invest in new players just for short term success, although that is what we wanted.</p> <p>ad of England. It looks like, in the short term at least, he made the right choice. "Beati</p> <p>originally an early 20th century US slang term of abuse either for any lesbian or for any</p> <p>the noise of closing doors, in the spring term of the first year of the "sixth-form" co</p> <p>t school only during the present teacher's term of office. The regular scholars, if the tr</p> <p>is exobiology? In fact, it's the technical term for the study of alien life in space. And</p> <p>. Since the turn of the year, though, the term "expansive" has disappeared from Cooke's</p> <p>ests, on the manner and sense in which the term is used: positive and life-affirming? or n</p> <p>r). He seems unaware that criticism of the term "adolescence" in relation to sexual behav</p> <p>c violence yes I agree with the use of the term domestic violence but this is not the term</p> <p>ow where they were. I don't know where the term "junglist" (hardcore Techno's dominant st</p> <p>inal festival band (provided you allow the term "original" a certain latitude). The field</p> <p>ton, Cheshire Commonplace criticism of the term 'adolescence' MARK SHIMPSON accused me of</p> <p>modern classic from the man who coined the term 'homophobia'. LESBIAN Somewhere Like This</p> <p>azz Dancing, Class of '89') redefines the term 'engaging': as in, I'd like to engage his</p> <p>h and many men did not even understand the term 'sexual intercourse' used in the titles".</p> <p>y camps and high camp. Black comedy is the term which springs to mind, but it would be mis</p> <p>e Government has been careful to avoid the term "victory" in relation to these events. I</p> <p>as a proxy for an academic record; but one term's work cannot yield "evidence" of potent</p> <p>ministrative legal area), during one legal term, there were 24 travelling days listed. Mul</p> <p>embarrassment causes blushing. In the long term, emotional stress affects the pituitary gl</p> <p>gger' and 'faggot'. It's such a pejorative term, one I associate with a mode of thinking,</p>
--	---

Fig. 3. Líneas de concordancia extraídas con Micro Concord de la palabra term.

Únicamente con echar un vistazo a estas líneas de concordancia (que son una pequeña fracción de las que aparecen en nuestro corpus) pueden verse no sólo los contextos de uso más frecuentes de la palabra "term" (*term of abuse, term of disapproval, term of office, jail term, prison term, in the short/long term*), sino también alguno de los compuestos en los que aparece: *long-term, short-term* (con y sin guión), *medium-term, mid-term, medium-to-long term, end-of-term, full-term (baby), one-term (President)*, etc.

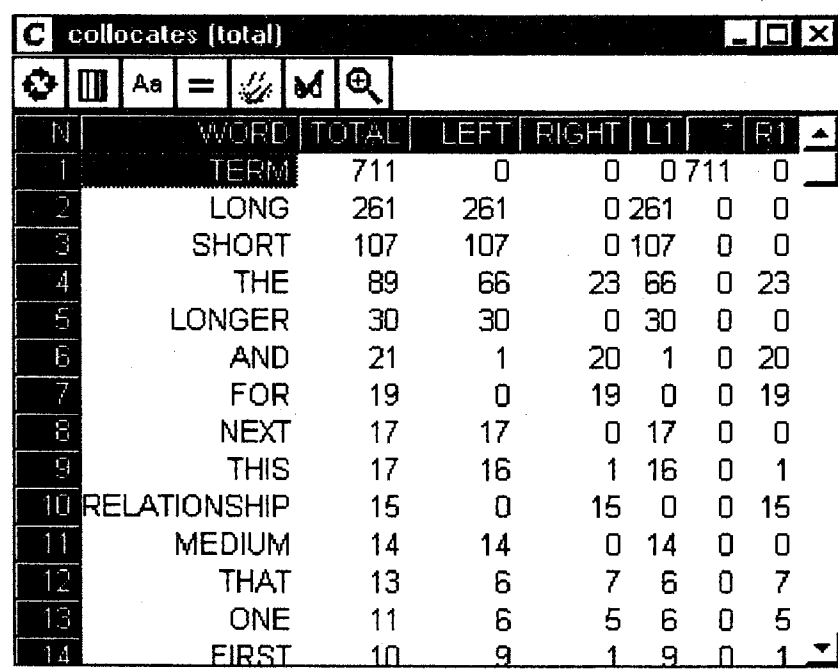
La mayoría de las herramientas incluyen también una serie de cálculos estadísticos, que pueden ir desde simples *índices de frecuencia* de aparición de una determinada forma (o formas) en el corpus e *índices de asociación de palabras* (colocaciones), hasta cálculos estadísticos muy complejos, desarrollados en centros de investigación especializados, en muchos casos orientados a la traducción automática, la adquisición automática de información léxica o la recuperación de información.

El estudio de los hábitos colocacionales de las palabras es uno de los caballos de batalla de la lexicografía, tanto monolingüe como bilingüe<sup>8</sup>, y sin embargo, es una de las áreas en la que los usuarios potenciales de un diccionario necesitan más ayuda, ya



que no resulta nada fácil llegar a dominar las combinaciones de palabras que se perciben como idiomáticas en una lengua extranjera. Los lexicógrafos, a la hora de estudiar una palabra o grupo de palabras y sus hábitos colocacionales, necesitan herramientas que les asistan en el análisis de las diversas combinaciones que pueden observarse en un corpus, sobre todo en aquellos casos en los que el corpus cuenta con un número muy elevado de palabras y/o cuando la palabra en cuestión presenta un índice de aparición muy alto, por lo que sería prácticamente imposible estudiar todas y cada una de las líneas de concordancia manualmente (Clear 1994).

Por esta razón, es muy útil contar con herramientas computacionales que ofrezcan listados de colocaciones, así como la posibilidad de ordenarlas según diferentes cálculos estadísticos. La figura 4, por ejemplo, muestra las colocaciones más frecuentes de la palabra “term”, en relación a las líneas de concordancia que habíamos extraído anteriormente:



N	WORD	TOTAL	LEFT	RIGHT	L1	R1
1	TERM	711	0	0	0	711
2	LONG	261	261	0	261	0
3	SHORT	107	107	0	107	0
4	THE	89	66	23	66	23
5	LONGER	30	30	0	30	0
6	AND	21	1	20	1	20
7	FOR	19	0	19	0	19
8	NEXT	17	17	0	17	0
9	THIS	17	16	1	16	1
10	RELATIONSHIP	15	0	15	0	15
11	MEDIUM	14	14	0	14	0
12	THAT	13	6	7	6	7
13	ONE	11	6	5	6	5
14	FIRST	10	9	1	9	1

Fig. 4. Colocaciones más frecuentes de term extraídas con Concord (Wordsmith Tools).

Algunos de esos cálculos estadísticos son muy útiles para la lexicografía, como por ejemplo uno de los índices que muestran la frecuencia de asociación denominado “índice de información mutua” (*MI Score*), en el que se mide la “fuerza” de asociación entre dos palabras, es decir, la cantidad de información que la aparición de una palabra nos da sobre la aparición de otra (Church y Hanks 1990). Esta medida estadística calcula la probabilidad de que las dos palabras ( $x$  y  $z$ ) aparezcan juntas, calculando la probabilidad de que  $x$  y  $z$  aparezcan de forma independiente y después comparando los dos valores. Si existe una asociación fuerte entre  $x$  y  $z$ , la probabilidad de que aparezcan juntas deberá ser mucho mayor que la de que aparezcan por separado. En

caso de que los dos valores de frecuencia sean muy similares, la concurrencia de las dos palabras no suele considerarse muy significativa.

Clear (1993) discute en profundidad la utilidad lexicográfica del índice de *información mutua*, comparándolo con otro índice que se usa con bastante frecuencia en lexicografía, el *T-score*, que mide, no como el anterior, la fuerza de la asociación de dos palabras, sino el grado de confianza con que se puede decir que existe una asociación de palabras. Las palabras que poseen un índice de frecuencia más alto en el corpus (preposiciones, pronombres o artículos) ofrecerán también un índice de colocación *t-score* mayor, de forma que índices significativos de esta medida suelen señalar colocaciones muy fuertes o asociaciones entre palabras léxicas y gramaticales (por ejemplo, preposiciones con verbos o con adjetivos), mientras que el índice de información mutua suele indicar asociaciones que son estadísticamente significativas (aunque la frecuencia de aparición de los elementos de la colocación en el corpus sea muy baja), por lo que suele señalar asociaciones semánticas entre palabras o elementos de una unidad fraseológica.

A modo de ejemplo, compárense una sección el índice de frecuencia de asociación *t-score* y el *índice de información mutua (MI)* de la palabra “term”, obtenidos a través del servicio *CobuildDirect*, ofrecido por la editorial Collins Cobuild en su servidor de internet, el cual posibilita la consulta de una parte de su corpus, el citado *Bank of English* (50 millones de palabras, lematizadas y etiquetadas), al que se accede a través de una potente herramienta de manejo de corpus denominada *lookup*. El programa permite seleccionar secciones del corpus, crear concordancias, hacer búsquedas complejas y, como vemos en las dos figuras que siguen, ofrecen diversos cálculos estadísticos:

long	35566	3108	54.974874	during	22181	137	9.403872
short	11038	1307	35.781761	interest	12573	115	9.300319
the	2872094	4793	18.863081	investment	5173	95	9.102411
longer	7811	347	18.118833	fixed	2359	87	9.020313
this	224039	697	16.097594	is	499929	873	9.003541
for	482791	1121	15.974017	future	11901	103	8.725160
a	1228514	2147	14.145253	jail	2450	76	8.376587
term	8714	198	13.319368	second	25016	121	8.238859
in	958631	1670	12.384545	relationship	7062	81	8.047317
end	28018	185	11.100465	next	32264	130	7.966097
year	76008	276	11.058452	last	67959	190	7.798093
medium	1631	123	10.911984	psychotherapy	278	59	7.637204
used	28519	170	10.382740	contract	4831	68	7.534923
use	25110	155	10.001151	contracts	1667	57	7.281756
rates	5737	110	9.823960	savings	2607	59	7.269069
effects	3888	103	9.683765	mid	3678	60	7.169467

Figura 5. T-score de las colocaciones de la palabra term (*CobuildDirect*).

endearment	24	13	8.802222	legislator	43	4	6.260227
michaelmas	19	5	7.760641	pathways	154	14	6.227056
psychotherapy	278	59	7.450312	long	35566	3108	6.170042
coined	116	23	7.352181	imprecise	37	3	6.061981
duisenberg	16	3	7.271555	involvements	37	3	6.061981
dyads	22	4	7.227156	maturities	50	4	6.042613
crocks	37	6	7.062080	colloquial	39	3	5.986024
incapacity	41	6	6.913967	medium	1631	123	5.957435
absentees	22	3	6.812077	crock	81	6	5.931571
gits	24	3	6.686534	generic	149	11	5.926721
derogatory	65	8	6.664164	vp	75	5	5.779553
short	11038	1307	6.608378	outweigh	91	6	5.763610
penal	93	11	6.606799	longitudinal	97	6	5.671482
prioress	32	3	6.271455	viability	115	7	5.648295

Figura 6. MI-score de las colocaciones de la palabra term (CobuildDirect).

Estos dos cálculos estadísticos están integrados en el programa diseñado para el estudio lexicográfico de las colocaciones desarrollado por Oxford University Press, denominado *collocate*, que además incluye la posibilidad de estudiar variaciones posicionales en los elementos de la colocación, permitiendo estudiar el co-texto derecho o izquierdo de la colocación independientemente, posibilidad que no ofrecían el índice de *información mutua* y el *t-score*, ya que no proporcionan información sobre la posición de los elementos de la colocación.

Otra de las áreas de aplicación lexicográfica del estudio de las colocaciones es la discriminación de significados (*sense discrimination*), es decir, el estudio de las diferentes acepciones de una palabra que deben incluirse en la entrada. Diferentes significados de una palabra suelen asociarse con colocaciones diferentes y con diversos patrones sintácticos. Baugh, Harley y Jellis (1996: 40), por ejemplo, destacan cómo el estudio de las colocaciones ayudó en el proceso de compilación del *CIDE*, tanto en el estudio del significado como en el de los patrones sintácticos asociados a los diferentes significados. Para estos autores, el corpus fue una herramienta fundamental a la hora de hacer distinciones de significados, y comparándolo con métodos tradicionales argumentan que “*through using the corpus, CIDE lexicographers often found that previous dictionaries defined quite rare senses of words but missed important, common ones*” (ibid.: 41).

Además del citado programa *collocate*, Clear (1994) muestra una herramienta computacional diseñada para discriminar los diferentes sentidos de una palabra usando listas de colocaciones extraídas de un corpus. Trabajando con una lista determinada de colocaciones de una palabra, esta herramienta procesa un número de líneas de concordancia, usando las colocaciones asociadas con un significado determinado como *indicios* (*clues*), y todas las demás colocaciones como *contrarios*

(*antis*). Después, añade información estadística sobre palabras que aparecen frecuentemente asociadas a las colocaciones (tanto las tomadas como *indicios* de un significado como las *contrarias*), de forma que agrupa las líneas de concordancia de acuerdo con la aparición (en un co-texto de 512 caracteres) de alguna de las colocaciones y sus palabras asociadas. Según se desprende de la discusión final de los resultados, esta metodología, aunque necesite refinarse para conseguir resultados más acertados, posee una utilidad lexicográfica enorme, sobre todo en las fases de análisis de significado más avanzadas, ya que puede ofrecer al lexicógrafo las concordancias agrupadas de acuerdo con los diferentes significados de una palabra y facilitar, por ejemplo, la selección de un ejemplo o el estudio de las restricciones de selección de una palabra.

Otro aspecto en el que los corpóra poseen una gran utilidad lexicográfica es en la selección de los ejemplos que se han de incluir junto con las definiciones en las entradas. Los ejemplos son de vital importancia en el proceso de compilación de un diccionario, sobre todo en aquéllos que están orientados al aprendizaje de una lengua extranjera, ya que pueden usarse para mostrar contextos típicos de uso, ilustrar restricciones de selección o características pragmáticas de una palabra para guiar a los usuarios, ofreciéndoles ejemplos similares a la frase que ellos intentan construir o entender. En muchos casos, los ejemplos no son tomados directamente del corpus, sino que el lexicógrafo los adapta, “inspirándose” o derivándolos de los que ha encontrado en el corpus, aunque no existe consenso sobre el grado en el que los ejemplos deben ser modificados antes de incluirlos en el diccionario (Fox (1987: 138) y Baugh, Harley y Jellis (1996: 43), por ejemplo, argumentan de forma diferente en lo que respecta a la autenticidad de los ejemplos).

## 2.2. *Lexicografía bilingüe basada en corpus*

A la hora de crear un diccionario bilingüe los lexicógrafos (independientemente de la editorial a la que pertenezcan) tienen una meta común, la de ofrecer al usuario una representación lo más acertada posible de las correspondencias que existen entre dos sistemas lingüísticos diferentes (el de la lengua de origen (LO) y el de la lengua meta (LM)). Además, han de conseguir esto con un diseño y presentación que combinen la claridad y la exhaustividad y que sean atractivos para los usuarios potenciales. Para ello, los lexicógrafos bilingües siguen normalmente un proceso que conlleva tres fases (Atkins 1990, Clari 1994): (i) la de-generalización de la lengua origen, (ii) la asociación de significados entre lengua origen y lengua meta y (iii) la nueva generalización de los datos que resultan de las dos operaciones anteriores (también llamadas *análisis*, *transferencia* y *síntesis*).

Los dos sistemas lingüísticos con los que los lexicógrafos están trabajando (LO y LM) no pueden ser comparados si no se tienen en cuenta una serie de parámetros gramaticales, sintagmáticos, semánticos y estilísticos. Es por tanto de vital

importancia para el lexicógrafo detallar la categoría gramatical de la palabra, su comportamiento sintáctico y morfológico, el registro en el que se usa y por supuesto, su significado o significados.

En este proceso de tres fases los compiladores (nativos de esa lengua), independientemente de la lengua meta, deciden la lista de entradas que han de incluirse en el diccionario y después analizan cada una de las palabras de la LO, de acuerdo con los parámetros que antes mencionábamos, asegurándose de que su análisis refleja los usos más comunes y centrales de esa palabra en esa lengua. De este análisis se obtiene un *framework* o marco de trabajo, que se da a los traductores (que normalmente son nativos de la LM). Su misión es encontrar la mejor correspondencia, el mejor *equivalente de traducción* en la lengua meta, de acuerdo con cada uno de los rasgos sintácticos, semánticos y estilísticos especificados por el equipo de compiladores. En algunos casos, también deben dar ejemplos y contextos de uso para las traducciones propuestas.

Después de varias comprobaciones y revisiones por parte de los dos grupos de lexicógrafos (LO y LM), esa masa de información debe ser reorganizada en un formato que sea claro y fácil de manejar para el usuario. El mismo proceso se sigue en las dos partes del diccionario para asegurar que se cubren satisfactoriamente las necesidades de los usuarios de ambas lenguas.

Estas tres fases son cruciales, pero quizá la más importante es la del establecimiento del *marco de trabajo* en la lengua origen. Las entradas del marco de trabajo de un diccionario bilingüe son similares a las de uno monolingüe pero, en este caso, las observaciones en cuanto al contexto y co-texto de la palabra en la lengua de origen deben llevarse hasta límites mucho mayores de degeneralización que en un diccionario monolingüe. La tarea del compilador es estudiar cada palabra en su contexto para dar cuenta de los patrones que son observables. Este estudio de la lengua de origen se hace independiente de la lengua meta y la descripción de la lengua de origen debe permanecer, en este estadio inicial, aislado de las influencias de las diferenciaciones de significado de la lengua meta.

En aquellos proyectos lexicográficos que ya han introducido el uso de los corpóra, los lexicógrafos tienen durante la fase de compilación mucha más facilidad para analizar y estudiar (con las herramientas a las que hacíamos referencia antes) el comportamiento de las palabras y los significados que están asociados a estos comportamientos. Pueden también ver cuáles son las colocaciones más frecuentes, las preferencias de selección con respecto a los sujetos o los objetos de los verbos, los adjetivos que típicamente acompañan a determinados sustantivos, las preposiciones que se usan más frecuentemente en determinadas estructuras, usos diferentes de palabras con significados similares, etc. Todos estos tipos de información son fundamentales para el hablante no nativo de la lengua y constituyen uno de los aspectos que presenta mayor dificultad a la hora de dominar una lengua extranjera: el

ser capaz de usar estructuras que no sólo sean gramaticalmente correctas, sino que además sean idiomáticas.

En la segunda fase, *la traducción a la lengua meta*, comienza el establecimiento de las equivalencias. Los traductores, que son los primeros “clientes” del marco de trabajo monolingüe, tienen acceso hoy día a mucha más información sobre la lengua de origen que antes, sobre todo en aquellos casos en los que se han compilado las entradas a través del estudio de corpora. Si el corpus se pone a disposición del equipo de traductores, éstos pueden comprobar y verificar los equivalentes de traducción propuestos traduciendo la palabra de la lengua de origen en muchos contextos diferentes, ya que en muchas ocasiones, patrones sintácticos o contextos que no se habían considerado significativos en el marco de trabajo monolingüe pueden ser determinantes para la asignación de un equivalente de traducción (Sinclair, Payne y Pérez 1996). De igual importancia es la utilización del corpus en la lengua meta, ya que el comportamiento del equivalente de traducción en su propio co-texto y contexto merece una atención que hasta ahora no se le ha prestado en prácticamente ningún diccionario bilingüe.

La mayoría de los problemas que se presentan a la hora de compilar (y a la hora de usar) un diccionario bilingüe provienen de la noción de *equivalencia de traducción*. Tal y como subraya Hartmann (1994), el concepto tradicional de equivalencia de traducción se limitaba a relacionar palabras de una lengua con sus equivalentes, considerándolas unidades formales en sistemas lingüísticos paralelos. Esta visión se hace todavía más patente con la aparente facilidad con la que los diccionarios bilingües nos ofrecen “ecuaciones” léxicas para ser insertadas en una porción de texto. Snell-Hornby (1984: 274), por su parte, también resalta el hecho de que los diccionarios operan con palabras aisladas aunque en la realidad esas palabras deben usarse en textos particulares y en una gran variedad de contextos diferentes.

Todos los que hemos usado alguna vez un diccionario bilingüe hemos tenido la experiencia de ir a buscar una palabra determinada y que la traducción o traducciones propuestas no nos satisfagan, no porque sean incorrectas, sino porque no estamos seguros de que puedan reproducir en la lengua meta no sólo el significado léxico, sino que además se ajusten a las restricciones y a las preferencias colocacionales del contexto y sean capaces de aportar un alto grado de idiomatidad al texto meta y suenen “naturales”. Tal y como señala Snell-Hornby (1984: 279), los diccionarios bilingües no serán capaces de cumplir la función para la que han sido creados si son sólo repositorios de lexemas aislados y equivalentes estáticos. Es necesario que estén preparados para “reveal the dynamic system of relationships within and between languages, the function of words in their contexts and the interdependence of language, culture and social interaction”.

Sin lugar a dudas, al igual que en el caso de la lexicografía monolingüe, el uso de un corpus en ambas lenguas puede ofrecer al lexicógrafo bilingüe una riqueza de información mucho mayor, de forma que le permita explicitar en las entradas del

diccionario, no sólo cuáles son los equivalentes de traducción de una palabra, sino cuáles son las restricciones o las limitaciones de la equivalencia y en qué contextos será un equivalente apropiado.

### 3. LEXICONES COMPUTACIONALES

Pasamos a continuación a tratar la otra área de investigación que se engloba dentro de la *lexicografía computacional*: la creación de lexicones computacionales. Como ya hemos mencionado, existen dos etapas fundamentales en cualquier empresa de construcción de un lexicón computacional: la fase de *adquisición* y la de *representación* de información léxica. Sin duda alguna, la segunda es la más relevante, en el sentido de que es la que determina si el producto resultante va a ser un lexicón computacional o simplemente un diccionario en formato magnético. La primera, por otra parte, es indispensable, y de ella se va a derivar la calidad y veracidad de la información contenida en el lexicón.

#### 3.1. *Adquisición de información léxica*

En lo que respecta a la *adquisición* de información léxica, podemos distinguir tres fuentes de información principales que se han usado tradicionalmente: (i) otros diccionarios, (ii) MRDs y (iii) córpora textuales. A estas tres fuentes habría que sumar una cuarta, el conocimiento lingüístico del lexicógrafo. Ya hemos mencionado en el apartado dedicado a la lexicografía las ventajas y desventajas que supone usar otros diccionarios y las intuiciones lingüísticas del lexicógrafo como fuentes de información en la compilación de un diccionario, por lo que no volveremos a repetirlo aquí ya que su uso plantea los mismos problemas y limitaciones (o quizá mayores, tal y como veremos después) en la creación de lexicones computacionales. Centraremos, por tanto, nuestra atención en el uso de fuentes de información en formato magnético, como son los diccionarios electrónicos (MRDs) y los córpora de texto informatizados.

En primer lugar es importante aclarar que un MRD no es lo mismo que un lexicón computacional. En principio, un MRD es la versión en formato magnético de un diccionario tradicional publicado en papel<sup>9</sup>. Por tanto, se trata de uno o varios ficheros que contienen texto, normalmente sin más estructuración que la que encontraríamos en el diccionario en papel. Esto significa que el tipo acceso a la información contenida en el MRD sigue siendo básicamente secuencial, aunque, claro está, podemos hacer uso de búsquedas de texto simples, búsquedas con expresiones regulares y todo aquello que es aplicable a los ficheros de texto en general.

Un lexicón computacional, también llamado LDB (*Lexical Data Base*), por otra parte, es una base de datos que organiza y estructura la información original en tablas,

registros y campos y que permite un acceso mucho más flexible y rápido mediante el empleo de índices, consultas, etc., así como la imposición de todas aquellas restricciones de integridad de los datos y de seguridad que el administrador de la base de datos considere oportuno.

Convertir un MRD en un lexicón computacional, por tanto, consiste en desarrollar programas que lean los ficheros de texto originales, delimiten detalladamente los campos que van a constituir la base de datos y transfieran estos datos a su lugar correspondiente (campo, registro y tabla, en el caso de un modelo relacional). En este sentido, el mayor problema que plantean los MRDs es que los diccionarios están hechos para ser utilizados por humanos, que saben manejar muy bien las inconsistencias y que pueden usar su conocimiento lingüístico para suplir o hacer todo tipo de inferencias lingüísticas en aquellas partes de la entrada de un diccionario que no están completan. La estructura definicional de los diccionarios es, desde el “punto de vista” de una base de datos, bastante inconsistente, por no decir caótica. Por ejemplo, la siguiente entrada está tomada al azar del LDOCE:

**cosmopolitan**<sup>1</sup> /kɒzmə'pɒlɪtən||kɒ:zmə'pɒ:-/ *adj* **1** consisting of people from many different parts of the world: *London is a very cosmopolitan city* **2** (of a person, belief, opinion, etc.) not narrow-minded; showing wide experience of different people and places: *She has a very cosmopolitan outlook on life.* **3** *tech* (of an animal or plant) existing in most parts of the world

Desde el punto de vista lexicográfico, la entrada es apropiada, toda la información está bien organizada, el uso de la negrita y cursiva ayuda a localizar la información, los números en negrita señalan la polisemia, existen restrictores semánticos, indicadores de registro idiomático, ejemplos, etc. Sin embargo, a la hora de “montar” esta entrada en una base de datos aceptable para su uso en tareas de NLP, el lingüista computacional se enfrenta a una serie de dificultades. Algunas de las más obvias son las siguientes:

- algunas entradas (como la de nuestro ejemplo) marcan la existencia de otra entrada con el mismo lema mediante un número en superíndice, pero otras no. El programa debe leer este marcador como independiente del lema porque de lo contrario no se podría establecer ninguna relación entre, por ejemplo, “cosmopolitan 1” y “cosmopolitan 2”.
- pueden existir 0, 1 ó  $n$  transcripciones fonéticas, pero además, como vemos en nuestro ejemplo, cuando la segunda acaba igual que la primera, no se incluye toda su transcripción, sino que se marca mediante un guión, convención que no plantea ningún problema al usuario humano, ya que el uso de guiones en los que el usuario debe insertar información para completar la entrada es práctica común en la lexicografía, pero plantea grandes problemas para que un programa pueda manejarlos.



- cada entrada puede tener 1 ó  $n$  subentradas que designan significados polisémicos u homonímicos; cuando hay más de una se marcan con un índice, pero no cuando sólo hay una. Además, a veces existe un punto antes del nuevo índice (“*outlook on life. 3 ...*”), pero otras veces no (“*cosmopolitan city 2 ...*”).
- la definición de cada una de las subentradas puede ir seguida de 0, 1 ó  $n$  ejemplos de uso. Cuando estos se dan, van seguidos del signo de dos puntos (“:”).
- a veces existen restrictores de uso (“of a person, ...”), pero en otras ocasiones no, y además no se hace explícita la forma en la que estos restrictores se relacionan sintáctica y co-textualmente con el lema.

Hemos querido utilizar una entrada del LDOCE como ejemplo, porque es sin duda el MRD más consistente que tenemos y el que más se ha usado en diversos proyectos de construcción de lexicones computacionales. Por ejemplo, se marca la existencia de múltiples transcripciones fonéticas con la doble barra vertical (“||”), cuando hay más de un ejemplo, se marca mediante la barra vertical (“|”), etc.; incluso así, ya hemos visto las dificultades que se plantean. Existen multitud de problemas que no hemos mencionado y que suponen enormes obstáculos para un buen aprovechamiento de estas fuentes de información.

Tal y como reconoce Levin (1991), el valor que posee el uso de los diccionarios electrónicos en la construcción de una base de conocimiento léxico<sup>10</sup> 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 inconscientes e incompletos (Atkins, Kelg y Levin 1988; Boguraev y Briscoe 1989), y que, por ejemplo, palabras que tienen un comportamiento similar (morfológico, sintáctico, semántico, etc.) no reciban un tratamiento homogéneo en los diccionarios, ya sea por falta de tiempo, por haber sido compiladas por diferentes lexicógrafos, o simplemente por que el lexicógrafo no fue capaz de reconocer las similitudes.

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 han 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 bastante reducido, bien por problemas con los derechos de publicación, bien por no disponer de la versión magnética correspondiente a la versión publicada en papel. De hecho, los diccionarios más usados hasta la fecha pueden reducirse 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), y *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 (Boguraev y Briscoe 1989; Bogurev 1991; Atkins 1991).

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. Sin embargo, como hemos observado en la entrada que hemos usado como ejemplo, la estructura de las entradas, los tipos de letra, estilos, etc., delimitan los "trozos de información" relevantes y se deben tener en cuenta. El problema, de nuevo, es la inconsistencia que exhiben estas marcas y que dificulta su correcta interpretación por el programa.

Hemos nombrado ya algunos de los problemas y desventajas que el uso de la información contenida en los MRDs plantean, 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<sup>11</sup>.

También hemos de destacar, sin embargo, que no todas las investigaciones realizadas con MRDs han sido infructuosas. Boguraev y Briscoe (1989), 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 (1987) ha diseñado un sistema capaz de construir entradas verbales de forma semiautomática.

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. Como veremos en el apartado siguiente, 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.

Éste, junto con alguno de los problemas que ya hemos señalado anteriormente, ha provocado que en los últimos diez años se haya considerado, en algunos proyectos de enorme magnitud como por ejemplo WordNet™ (Miller et al. 1993), o Cyc (Guha y Lenat 1990), 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 corpórea textuales informatizados como fuentes potenciales para la adquisición de información léxica actualizada.

La mayoría de los experimentos llevados a cabo para la adquisición de información léxica a través de corpórea se hallan aún en fase experimental, por lo que quizás sea pronto para extraer conclusiones definitivas sobre su utilidad<sup>12</sup>. En el momento presente, los corpórea textuales han demostrado ser de gran utilidad en el ámbito de la lexicografía comercial y en otros ámbitos de estudio lingüístico 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 corpórea 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.

Por estas razones un gran número de investigadores (Hindle y Rooth 1991; Boguraev y Pustejovsky 1996) 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.

Un caso interesante de uso conjunto de varias fuentes es la investigación llevada a cabo por Hearst y Schüetze (1996), ya que usan una base de datos construida manualmente, WordNet™, 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 coocurrencia léxica, aumentando y reubicando los elementos del lexicón, y haciéndolos más apropiados

para otras tareas específicas a un dominio determinado (*domain-specific task*), como por ejemplo la recuperación de información (*information retrieval*).

### 3.2. Representación de información léxica

De entre los proyectos orientados a la creación de bases de datos léxicas mediante MRDs a los que hacíamos referencia en el apartado anterior, debemos destacar sin duda Acquilex, cuya finalidad fue la de extraer información léxica, no ya de un solo MRD, sino de varios y de varias lenguas, integrando la información en un único repositorio de información.

El proyecto Acquilex puede servirnos para estudiar el otro gran aspecto de los lexicones computacionales: la *representación* de la información léxica. Ya hemos mencionado el término “LDB”, normalmente empleado como sinónimo de “lexicón computacional”; también hemos mencionado algunas diferencias entre éste y un MRD. Otro término que comenzó a emplearse precisamente en Acquilex fue el de “LKB” (*Lexical Knowledge Base*). En muchas ocasiones nos encontramos con el término “base de conocimiento” usado de una forma muy libre, para hacer referencia a cualquier conjunto de información en formato magnético más o menos compleja. Sin embargo, no es tanto el contenido, sino el contenedor lo que determina el tipo de “base de información”<sup>13</sup>.

En el ámbito de la representación léxica, el término “LKB” hace referencia a un tipo concreto de base de información. Según Ingria et al. (1992: 360), una LKB es “a large-scale depository of lexical information, which incorporates more than just static descriptions of words, e.g., by means of clusters of properties and associated values” y añaden que este repositorio dinámico especifica “(1) constraints on word behavior, (2) dependence of word interpretation, and (3) distribution of linguistic generalizations”.

Como vemos, una LKB se diferencia básicamente de una LDB en su naturaleza dinámica, frente a la representación típicamente estática de una base de datos. El tipo de estructuras de datos que sirven de soporte para la representación de la información también varía de un tipo de sistema a otro. En una LDB, por otra parte, la estructura de datos es el modelo de datos que utilice el DBMS (*Data Base Management System*), normalmente el relacional, por lo que nuestros datos estarán estructurados, como ya hemos mencionado, en tablas (o *relaciones*), registros (o *tuplas*) y campos (o *atributos*). En una LKB, la estructura de datos más utilizada es sin duda la estructura de rasgos (tipificada) –*TFS: (typed) feature structure*– basada en unificación. Ejemplos clásicos de formalismos que emplean estructuras de rasgos con unificación son DATR (Evans y Gazdar 1990) y PATR-II (Shieber 1986, 1992). La figura 7 muestra un ejemplo de estructura de rasgos:

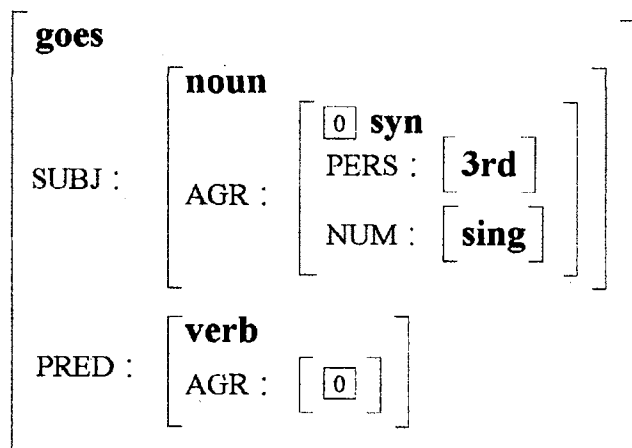


Fig. 7. Estructura de rasgos.

Una característica fundamental de este tipo de formalismos es el uso de la herencia de propiedades como mecanismo fundamental de representación. En el ejemplo anterior, el atributo AGR(ement) de PRED(icate) tiene como valor 0 , por lo que hereda los valores de AGR para SUBJ(ect). Como podemos adivinar, las estructuras de rasgos no sólo sirven para almacenar información léxica estática en el lexicón, sino que también pueden dar soporte a las representaciones gramaticales intermedias que se generan en tiempo de ejecución en las aplicaciones de NLP (parsing, etc.). Esta flexibilidad y potencia ha hecho de estos formalismos los más populares en la construcción de lexicones computacionales enfocados a aplicaciones de NLP. La figura 8 muestra una entrada léxica del lexicón de Aquilex.

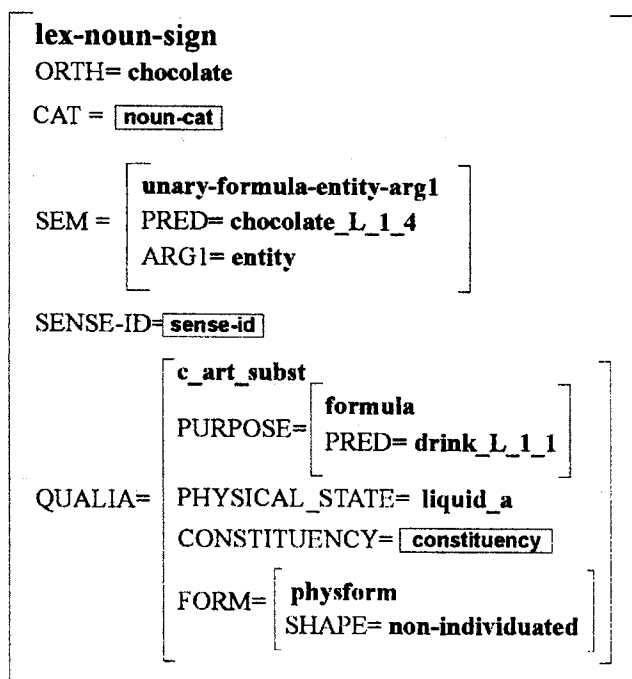


Figura 8. Entrada léxica del Aquilex (Copestake y Sanfilippo 1993).

Pudiera parecer que una LKB, utilizando un formalismo basado en unificación, es en cualquier circunstancia más apropiada que una LDB, sin embargo, no todo son ventajas, ya que como suele ser normal, flexibilidad va unida a complejidad. Para empezar, una LKB requiere una implementación compleja, normalmente en Lisp (listas anidadas que se corresponden con los corchetes de las figuras 7 y 8), no es fácil implementar un interfaz adecuado y se hace muy difícil de manejar cuando la LKB crece (control de cambios, seguridad, integridad referencial, etc.); tampoco es fácil convertir una LKB en un diccionario tradicional, es decir conseguir un *output* aceptable para su utilización por humanos a partir de la información contenida en la LKB. En definitiva, no se cuenta con un gestor que facilite todas estas tareas y otras muchas que un DBMS sí aporta. Por tanto, no es el tipo de representación adecuada para tareas lexicográficas, donde una LDB bien diseñada puede suplir algunas deficiencias que estos sistemas muestran con respecto a los formalismos basados en unificación (herencia) y además facilita su utilización a personas no especialistas en informática, es decir, el equipo de lexicógrafos. Además, los gestores de bases de datos con arquitectura cliente/servidor permiten su utilización en red de un modo transparente para el usuario (lexicógrafo), facilitando su labor con *vistas* personalizadas de los datos relevantes a cada uno.

El modo habitual de trabajo con una LKB es utilizando la línea de comandos y ficheros de texto, mientras que lo habitual en una LDB es utilizar un interfaz gráfico de usuario. La figura 9 muestra un interfaz de este tipo<sup>14</sup>.

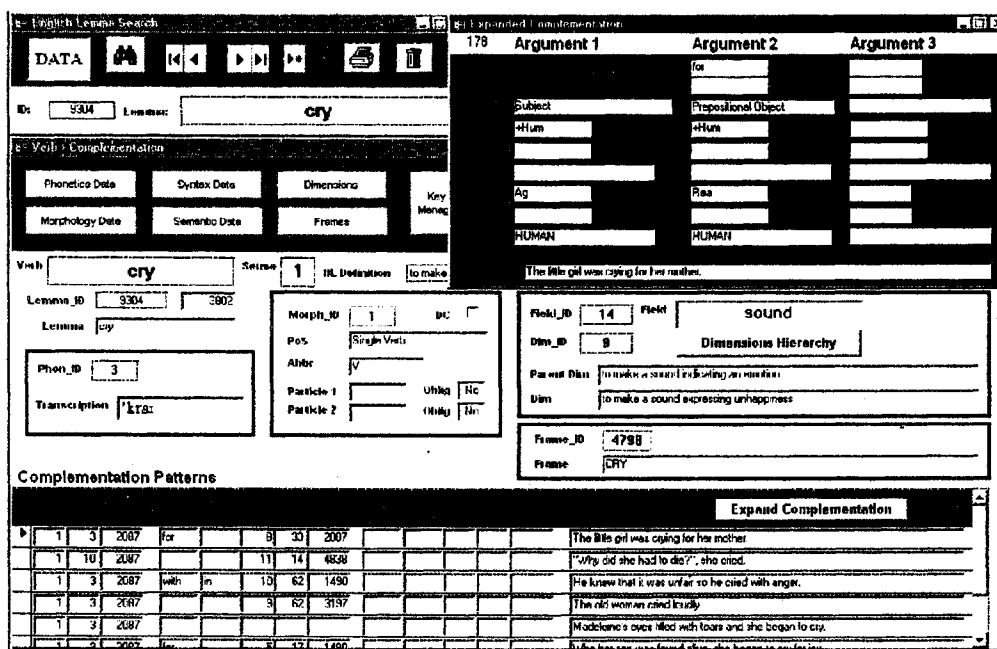


Fig. 9. Interfaz Gráfico de Usuario de una LDB

Los DBMS actuales mantienen un gran nivel de abstracción entre los niveles externos de la base de datos y el nivel conceptual de la misma. El usuario final no

tiene por qué conocer este esquema conceptual, ni siquiera el emplazamiento físico de los datos, ya que las vistas externas, junto con el interfaz se encargan de mostrar sólo aquello que le es relevante para su trabajo. Esto es válido para todos los usuarios, tanto para los lexicógrafos encargados de la edición como de los usuarios que consultan la base de datos. Un conocido ejemplo de una LDB que proporciona este tipo de funcionalidad es CELEX, que desde su servidor Web (<http://www.kun.nl/celex>) proporciona vistas personalizadas, o *subdiccionarios*, de la base de datos principal.

Las más complejas LKBs proporcionan, como ya hemos visto, funcionalidades distintas y se suelen utilizar como fuentes de información léxica para aplicaciones de NLP, pero, en aquellas empresas que contemplen la introducción manual de datos como fuente principal, el empleo de una LDB es sin duda lo más apropiado, así como en los casos en los que los receptores principales de la información léxica sean usuarios humanos. Esto no implica que nuestro lexicón se halle restringido a este formato para siempre. Un correcto diseño nos permitirá reutilizar nuestra información del modo más apropiado dependiendo de las nuevas necesidades. De hecho, éste fue el procedimiento en el ya mencionado proyecto Acquilex, donde las distintas LDBs monolingües fueron fusionadas en la LKB multilingüe final, que utilizaba un formalismo basado en unificación mediante TFS específico para este proyecto llamado LRL (*Lexical Representation Language*) (Calzolari et al. 1993).

Algo parecido ha ocurrido con otra conocida LDB, WordNet™ (Miller et al. 1993). Este lexicón fue inicialmente concebido como una base de datos de motivación psicolingüística y con una gran carga de información semántica, pero probablemente ha sido el recurso léxico que más se ha reutilizado en la historia de los lexicones computacionales<sup>15</sup>.

En definitiva, el mejor tipo de sistema de información dependerá en cualquier caso de la aplicación que pretendamos darle a nuestro lexicón, pero también debemos tener en cuenta posibles futuras aplicaciones, por lo que se ha de hallar un compromiso entre estos dos factores de tal modo que una inversión de estas proporciones pueda ser explotada al máximo.

#### 4. CONCLUSIÓN

En este trabajo hemos querido exponer una serie de metodologías que aprovechan las posibilidades de los ordenadores digitales modernos para llevar a cabo una serie de tareas relacionadas con el análisis lingüístico, la lexicografía y la lexicología. Como hemos visto, estas metodologías se han configurado como campos de estudio en sí mismos: la lexicografía y lexicología computacionales y la lingüística y lexicografía de corpus. Signos claros del asentamiento de estas disciplinas son no sólo el gran número de publicaciones y proyectos dedicados a las mismas, sino

también el creciente número de herramientas, incluso comerciales, que han sido desarrolladas.

Durante nuestra exposición hemos intentado dar una visión general de los diferentes aspectos que estos términos implican y las actividades que comúnmente se desarrollan en éstos ámbitos de investigación ya que, debido al alto grado de interconexión que existe entre ellas, son términos que en ocasiones se emplean indistintamente.

Hemos visto cómo la lexicografía de corpus (siguiendo los postulados de la lingüística de corpus) enfatiza la necesidad de utilizar evidencia lingüística proveniente de textos reales, en lugar de confiar únicamente en las fuentes tradicionales de conocimiento léxico: diccionarios y materiales existentes y el conocimiento e intuición del lingüista o lexicógrafo. En la práctica del análisis lingüístico, no se trata de sustituir las metodologías tradicionales por las nuevas, sino que éstas se ven reforzadas y apoyadas por evidencia mensurable y accesible mediante diversas herramientas informáticas. La lexicografía de corpus, por tanto, se basa, en mayor o menor medida según el enfoque adoptado, en emplear recursos textuales para obtener la información léxica que incluirá en sus diccionarios y lexicones, haciéndolos mejores y más completos.

También hemos hecho referencia a otra fuente de información léxica para la construcción de lexicones computacionales de forma automática o semiautomática: los MRDs, o diccionarios en formato magnético, haciendo un repaso de las ventajas y dificultades que plantean. Finalmente, no hemos querido dejar de lado la otra gran faceta de la lexicografía y lexicología computacionales: la construcción de los lexicones computacionales en sí. Éste es un vasto campo de investigación y desarrollo en el que se integran metodologías provenientes de muy distintas disciplinas, como son la representación del conocimiento, las bases de datos y la lógica, entre otras, por lo que tan sólo hemos pretendido atisbar algunos de los aspectos interesantes y relevantes para el lingüista o el lexicógrafo de este prometedor campo de la lingüística informática o computacional.

## NOTAS

1. Por ejemplo, V. Ooi, en un libro de muy reciente publicación define la *lexicografía computacional* como “either using the computer to achieve the goal of fully automating lexicographic tasks or utilising machine-readable versions of commercial dictionaries in a format explicit enough for computational linguistic systems” (Ooi 1998:1).
2. 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. (1996:2) 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...”.



3. Tanto McEnery y Wilson (1997) como Tognini-Bonelli (1996) hacen un repaso extenso de los estudios de carácter empírico realizados desde finales del siglo XIX hasta los años cincuenta, en lo que se conoce como "Early Corpus Linguistics". Estos estudios se encuadran en áreas tales como la adquisición del lenguaje, la lingüística comparada e histórica, la dialectología o la enseñanza de la lengua. Dentro de esta tendencia empiricista pre-chomskiana destacan los trabajos realizados por lingüistas de la talla de Z. Harris, A. Hill o C. Fries, para los que el uso de un corpus (es decir, una colección lo suficientemente amplia de texto producido de forma espontánea) era condición suficiente y necesaria para el estudio lingüístico.
4. Véase, por ejemplo, los numerosos proyectos de investigación e iniciativas conjuntas que aparecen reflejadas en las actas de las conferencias anuales llevadas a cabo en el *University of Waterloo, Centre for the New OED and Text Research*, o los artículos e informes contenidos en Walker, Zampolli y Calzolari (1995), Wilks et al. (1996), Boguraev y Briscoe (1989) y Kiefer, Kiss y Pajzs (1992).
5. El *Collins Cobuild Corpus* fue creado con el propósito de que fuera una muestra representativa del inglés británico moderno, por lo que contenía textos tanto provenientes de variedades regionales como de lenguaje general con una gran difusión entre los hablantes. Hoy por hoy, casi veinte años después, este corpus, desarrollado en el denominado *Bank of English*, cuenta con 320 millones de palabras y ofrece diversos servicios (como el acceso directo a parte del corpus) a través de su servidor de internet (<http://titania.cobuild.co.uk>).
6. El BNC es un corpus creado bajo la dirección de Sir Randolph Quirk, con la colaboración de las editoriales Oxford University Press, Addison-Wesley Longman y Larousse Kingsfisher Chambers, la *British Library* y las Universidades de Oxford y Lancaster, con la finalidad de que sea una muestra representativa (de 100 millones de palabras) del mayor número posible de estilos y variedades de la lengua inglesa actual y que ofrezca a la comunidad científica y también a la industria un corpus representativo que pueda ser usado en una amplia variedad de tareas en el ámbito del procesamiento del lenguaje natural y de las industrias de la lengua. Se distribuye en formato CD-ROM y puede ser adquirido directamente a través de su servidor de internet en la dirección <http://info.ox.ac.uk/bnc>.
7. Bruno M. Schulze y Ulrich Heid realizaron en 1994 un estudio comparando los programas de concordancias más destacados, tanto comerciales como académicos. Este informe puede obtenerse a través del servidor de EURALEX en [http://www.ims.uni\\_stuttgart.de/euralex](http://www.ims.uni_stuttgart.de/euralex).
8. La noción misma de *colocación* ha sido entendida y definida de formas diferentes por diferentes autores. En términos generales, suele entenderse la concurrencia (aparición simultánea) de dos o más palabras en un segmento de texto en el que la distancia entre los elementos de la colocación no sobrepase las cuatro o cinco palabras.
9. Existen excepciones a esta definición. Algunas editoriales, como Oxford con su OALD y Longman con el LDOCE han publicado versiones en formato magnético de sus diccionarios que contienen más información que sus correspondientes en papel. Por ejemplo, la versión magnética del LDOCE contiene códigos que aportan información de carácter semántico sobre sustantivos y argumentos verbales.
10. Más adelante explicamos lo que se entiende por este término y otros parecidos.
11. 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.
12. Boguraev y Pustejovsky (1996) ofrecen una colección muy ilustrativa de proyectos que se están llevando a cabo actualmente en esta línea.
13. Este término lo proponen Brodie y Mylopoulos (1986) y debería ser empleado para hacer referencia de forma genérica a cualquier repositorio de información estructurada que cumplan los mínimos requisitos en cuanto a consulta y actualización de datos. El ambicioso término de "base de conocimiento" debería utilizarse con cautela, porque ni siquiera los investigadores en Inteligencia Artificial consiguen ponerse de acuerdo en cuáles son las características que diferencian a una base de datos de una base de conocimiento (véase, por ejemplo, Brodie y Mylopoulos 1986, Bubenko y Orci 1989).

14. Esta captura de pantalla pertenece a la LDB creada dentro del proyecto DGYCIT PB94-0437. Los detalles sobre este lexicón se encuentran en Moreno Ortiz (1998).
15. Véase, por ejemplo, Vossen (1997), Viegas et al. (1998), Artale et al. (1998), por mencionar algunas de las más recientes aplicaciones.

## REFERENCIAS

- Aarts, J, P. de Haan y N. Oostdijk, eds. 1993. *English Language Corpora: Design, Analysis and Exploitation. Papers from the thirteenth International Conference on English Language Research on Computerized Corpora. Neijmen 1992*. Amsterdam: Rodopi.
- Aarts, J. 1991. "Intuition-Based and Observation-Based Grammars". Eds. Aijmer & B. Altenberg.
- Aijmer K. y B. Altenberg, eds. 1991 *English Corpus Linguistics*. London: Longman.
- Alvar Ezquerro, M. 1983. *Lexicología y Lexicografía. Guía Bibliográfica*. Salamanca: Almar.
- Alvar Ezquerro, M. y J. A. Villena Ponsoda. 1994. *Estudios para un Corpus del Español*. Grafur: Universidad de Málaga. Anejo número 7 de Analecta Malacitana, Revista de la Sección de Filología de la Facultad de Filosofía y Letras de Málaga.
- Artale, A. et al. 1998. "Coping with WordNet Sense Proliferation". *Proceedings of the First International Conference on Language Resources and Evaluation (ELRA)*. Granada, 28-30 May. 97-104.
- Atkins, B. 1987. "Semantic ID tags: Corpus evidence for dictionary senses". *The Uses of Large Text Databases. Proceedings of the 3rd Annual Conference of the UW Centre for the New OED*. Waterloo, Ontario: Oxford U. P. 17-36.
- Atkins, B. 1990. "Corpus Lexicography: the bilingual dimension". Eds. L. Cignoni & C. Peters. *Linguistica Computazionale. Computational Lexicology and Lexicography*. Special issue dedicated to B. Quemada, I. Pisa: Giardini Editori e Stampatori. 43-64.
- Atkins, B. 1991. "Building a Lexicon: The Contribution of Lexicography". Ed. B. Boguraev.
- Atkins, B. 1992. "Tools for computer-aided Corpus Lexicography: The Hector project". Eds. F. Kiefer, G. Kiss & J. Pajzs.
- Atkins, B. 1993. "Theoretical lexicography and its relation to dictionary-making". Ed. W. Chisholm.
- Atkins, B., J. Kegl y B. Levin. 1986. "Explicit and implicit information in dictionaries". *Advances in Lexicology*. Proceedings of the 2nd Annual Conference for the New OED. Waterloo, Ontario: OUP. 45-63.

- Atkins, B., J. Keggl y B. Levin. 1988. "Anatomy of a Verb Entry: from Linguistic Theory to Lexicographic Practice", *International Journal of Lexicography*. 1 (2): 84-125.
- Baker, M., G. Francis y E. Tognini-Bonelli, eds. 1993. *Text and Technology. In Honour of John Sinclair*. Philadelphia/Amsterdam: John Benjamins Publishing Company.
- Baugh, S., A. Harley y S. Jellis. 1996. "The Role of Corpora in Compiling the Cambridge Dictionary of English", *International Journal of Corpus Linguistics*. 1(1): 39-60.
- Biber, D. et al. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Bennet, P.A. et al. 1986. *Multilingual Aspects of Information Technology*. Hants: Gower.
- Bibliograf SA, ed. 1992. *EURALEX '90 Proceedings* Barcelona: Bibliograf /VOX.
- Boguraev, B. y T. Briscoe, eds. 1989. *Computational Lexicography for NLP*. London and New York: Longman.
- Boguraev, B. 1991. "Building a Lexicon: The Contribution of Computers". Ed. B. Boguraev. *Building a Lexicon*. Special Issue. *International Journal of Lexicography*. Vol. 4 (3), 1991.
- Boguraev, B. y J. Pustejovsk. 1996. *Corpus Processing for Lexical Acquisition*. Cambridge, Mass: The M Press.
- Brodie, M. L. y J. Mylopoulos, eds. 1986. *On Knowledge Base Management Systems. Integrating Artificial Intelligence and Database Technologies*. New York: Springen-Verlag
- Brown, F. Peter et al. 1990. "A statistical approach to machine translation", *Computational Linguistics*. 16 (2), June 1990. 79-85.
- Bubenko, J. A. y I. P. Orci. 1989. "Knowledge Base Management Systems: A Database View". Eds. J. W. Schmidt, y C. Thanos, 363-368.
- Bunge, M. 1995. "Quality, Quantity, Pseudoquantity and Measurement in Social Science". *Journal of Quantitative Linguistics* 2 (1): 1-10.
- Butler, C. 1985. *Statistics in Linguistics*. Oxford: Blackwell.
- Calzolari, N. et al. 1993. "Encoding Lexicographic Definitions as Typed Feature Structures". *Theorie und Praxis des Lexikons*. Eds. F. Beckmann y G. Heyer, Walter de Gruyter: Berlin.
- Calzolari, N. 1994. "Issues for Lexicon Building". Eds. A. Zampolli, N. Calzolari & M. Palmer. *Current Issues in Computational Linguistics: In Honour of Don Walker*. Liguistica Computazionale Vol. IX-X, Pisa: Giardini Editori e Stampatori.
- Charniak, E. 1993. *Statistical Language Learning*. The MIT Press: Cambridge, Mass.
- Chrisholm, W., ed. 1993. *Dictionaries. Journal of The Dictionary Society of North América*, 14. 1992/1993. Cleveland: Cleveland State University.

- Church, K. y P. Hanks. 1990. "Word Association Norms, Mutual Information and Lexicography". *Computational Linguistics* 16(1): 22-29.
- Church, K. y R. Mercer. 1993. "Introduction to the special issue on Computational Linguistics using large corpora". *Computational Linguistics* 19 (1). ACL. 1-24.
- Clari, M. 1994. "Compilation of Entries in Bilingual Dictionaries. Sense Discrimination and the Problem of Presentation of Translation Equivalents in Current Bilingual Lexicography". Ponencia presentada en el *Malvern Seminar*, Malvern, Mayo 1994.
- Clear, J. 1993. "From Firth principles: Computational Tools for the Study of Collocation". Eds. M. Baker, G. Francis, & E. Tognini-Bonelli. 271-292.
- Clear, J. 1994. "I Can't See the Sense in a Large Corpus". Eds. F. Kiefer, G. Kiss, J. Pajzs.
- Clear, J. 1996. "Technical Implications of Multilingual Corpus Lexicography", *International Journal of Lexicography* 9(3).
- Copetake, A. y A. Sanfilippo. 1993. "Multilingual Lexical Representation ". Ed. B. Dorr. *Building Lexicons for Machine Translation*. Proceedings of the AAAI Spring Symposium, Stanford, California.
- Evans, R. y G. Gazdar. 1990. *The DATR Papers*. Technical Report CSRP 139, School of Cognitive and Computing Sciences, University of Sussex, Falmer: Sussex.
- Fielding, N. G. y M. G. Lee, eds. 1991 *Using Computers in Qualitative Research*. SAGE.
- Firth, J.R. 1957. "A Synopsis of Linguistic Theory, 1930-1955". *Studies in Linguistic Analysis*, Special Volume, Philological Society, 1-32.
- Fox, G. 1987. "The Case for Examples". Ed. J. M. Sinclair. 137-149.
- Gale, W. y K. Church. 1993. "A program for aligning sentences in bilingual corpora" *Computational Linguistics* 19 (1): 75-101.
- Grishman, R. 1986. *Computational Linguistics*. Cambridge: Cambridge University Press. Traducción de Antonio Moreno Sandoval (1991) *Introducción a la Lingüística Computacional*. Madrid: Visor.
- Guha, R. V. y D. B. Lenat. 1990. "CyC: a Midterm Report", *Artificial Intelligence* 11: 32-59.
- Haensch, G. et al. 1988. *La Lexicografía. De la Lingüística Teórica a la Lexicografía Práctica*. Madrid: Gredos.
- Hanks, P. 1987. "Definitions and explanations". Ed. J. M. Sinclair, 116-136.
- Hanks, P. 1993. "Lexicography: Theory and Practice". Ed. W. Chirsholm, 97-111.
- Hanks, P. 1996. "Contextual Dependency and Lexical Sets", *International Journal of Corpus Linguistics*. 1 (1): 75-98

- Hartmann, R.R.K. 1983. *Lexicography. Principles and Practice*. London/New York: Academic Press.
- Hausmann, E. J. *et al.*, eds. 1989, 1990, 1991. *Dictionaries. International Encyclopædia of Lexicography*. Berlin: Walter de Gruyter. [Vol. I on Lexicographic Theory, Vol. II on Dictionaries Types and History, Vol. III on Lexicographic Traditions and Interlingual Dictionaries.]
- Hearst, M. y H. Schüetze. 1996. "Customising a Lexicon to Better Suit a Computational Task". Eds. B. Boguraev y J. Pustejovsky.
- Hindle, D. y M. Rooth. 1991. "Structural Ambiguity and Lexical Relations". *Proceedings of the 29<sup>th</sup> Annual Meeting of the Association for Computational Linguistics*. ACL.
- Hoey, M., ed. 1993. *Data, Description, Discourse. Papers on the English Language in Honour of John Sinclair*. London: Harper Collins Publishers.
- Ingria, R., B. Boguraev y J. Pustejovsky. 1992. "Dictionary/Lexicon". Ed. Shapiro. *Encyclopedia of Artificial Intelligence* (2<sup>nd</sup> ed.). New York: John Wiley. 341-65.
- Instituto Cervantes. 1996. *Report on Linguistic Resources for Spanish II. Written and Spoken Corpora Available or in Progress in Spain*. Observatorio Español de Industrias de la Lengua. Instituto Cervantes. Alcalá de Henares, 1996.
- Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London and New York: Longman.
- Kiefer, F., G. Kiss y J. Pajzs, eds. 1992. *Papers in Computational Lexicography. COMPLEX' 92*. Budapest: Linguistic Institute Hungarian Academy of Science.
- Klavans, J. L. y E. Tzoukermann. 1990. "Combining Lexical Information from Bilingual Corpora and Machine-Readable Dictionaries". *Proceedings of the 13th International Conference on Computational Linguistics: COLING*. Helsinki, Finland.
- Lager, T. 1995. *A Logical Approach to Computational Corpus Linguistics*. Tesis Doctoral. Gothenburg Monographs in Linguistics 14. Department of Linguistics, Göteborg University, Sweden.
- Leech, G. 1992. "Corpora and theories of linguistic performance". Ed. J. Svartvik. 105-134.
- Levin, B. 1991. "Building a Lexicon: The Contribution of Linguistics". Ed. B. Boguraev. *Building a Lexicon*. Special Issue. *International Journal of Lexicography* 4 (3).
- Lipka, L. 1990. *An Outline of English Lexicology. Lexical Structure, Words Semantics and Word Formation*. Tübingen: M. Niemeyer.
- McEnery, T. y A. Wilson. 1996. *Corpus Linguistics*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edingurgh University Press.

- Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. y Teng, R. 1993. *Five Papers on WordNet™*. CSL Report 43, July 1990. Revisión de Marzo 1993.
- Moreno Ortiz, A. J. 1998. *Diseño e Implementación de un Lexicón Computacional para Lexicografía y Traducción Automática*. Tesis doctoral. Universidad de Córdoba.
- Moon, R. 1987. "The analysis of meaning". Ed. J. M. Sinclair. 86-103.
- Moon, R. 1994. "The Analysis of Fixed Expressions in Text". Ed. Couthard, *Advances in Written Text Analysis*. London: Routledge. 117-135.
- Ooi, V.B.Y. 1998. *Computer Corpus Lexicography*. Edinburgh Textbooks in Empirical Linguistics. Edinburgh: Edinburgh University Press.
- Oostdijk, N y P. de Haan, eds. 1994. *Corpus-based Research into Language*. In *Honour of Jan Aarts. Language and Computers: Studies in Practical Linguistics*. No 12. Amsterdam: Rodopi.
- Oostdijk, N. 1991. *Corpus Linguistics and the Automatic Analysis of English*. *Language and Computers: Studies in Practical Linguistics*. No 6. Amsterdam: Rodopi.
- Pustejovsky, J. 1987. "On the Acquisition of Lexical Entries: the Perceptual Origin of Thematic Relations", *Proceedings of the 25<sup>th</sup> Conference of the Association for Computational Linguistics*, 172-178.
- Pustejovsky, J. 1991. "The Generative Lexicon". *Computational Linguistics*. 17 (4).
- Rayson, P., G. Leech y M. Hodges. 1997. "Social Differentiation in the Use of English Vocabulary: Some Analysis of the Conversational Component of the British National Corpus". *International Journal of Corpus Linguistics*. 2 (1): 133-152.
- Renouf, A., y Sinclair, J. M. 1991. "Collocational frameworks in English". Eds. K. Aijmer y B. Altenberg, 128-143.
- Sánchez, A. et al. 1995. *CUMBRE: Corpus Lingüístico del Español Contemporáneo. Fundamentos, Metodología y Aplicaciones*. Madrid: SGEL.
- Sánchez, A. y P. Cantos. 1997. "Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus: An 8-Million-Word Corpus of Contemporary Spanish". *International Journal of Corpus Linguistics* 2 (2): 259-280.
- Shieber, S. M. 1986. *An Introduction to Unification-Based Approaches to Grammar*, CSLI Lecture Notes. Vol. 4. Chicago: Chicago University Press.
- Shieber, S. M. 1992. *Constraint-Based Grammar Formalisms*. Cambridge, Mass: The MIT Press.
- Sinclair, J. M., ed. 1987a. *Collins Cobuild English Language Dictionary*. London: Harper Collins.
- Sinclair, J. M., ed. 1987b. *Looking Up: an Account of the COBUILD Project in Lexical Computing*. London: Collins.

- Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Sinclair J. M. 1992a. "Trust the text". Eds. M. Davies y L. Ravelli. *Advances in Systemic Linguistics: Recent Theory and Practice*. London: Pinter. 5-19.
- Sinclair, J. M. 1992b. "The automatic analysis of corpora". Ed. J. Svartvik. 379-398.
- Sinclair, J. M. 1993. "Lexicographers' needs". *Zeitschrift für Anglistik and Amerikanistik*. Berlin: Langenscheidt.
- Sinclair, J. M. 1996. "The Empty Lexicon". *International Journal of Corpus Linguistics* 1 (1): 99-119.
- Sinclair, J. M. y D. M. Kirby. 1990. "Progress in Computational Lexicography". *World Englishes* 9, No 1. Oxford: Pergamon.
- Sinclair, J. M., J. Payne y Ch. Pérez, eds. 1996. *Corpus to Corpus: A Study of Translation Equivalence*. *International Journal of Lexicography* 9 (3).
- Snell-Hornby, M. 1984. "The bilingual dictionary - help or hindrance?". Ed. R.K.K. Hartmann, 274-282.
- Stubbs, M. 1993. "British Traditions in Text Analysis". Eds. M. Baker, G. Francis & E. Tognini-Bonelli, 1-33.
- Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.
- Svartvik, J., ed. 1992. *Directions in Corpus Linguistics. Proceeding of Nobel Symposium 82, Stockholm, 4-8 August 1991*. Berlin/New York: Mouton de Gruyter.
- Tognini-Bonelli, E. 1996. *Corpus Theory and Practice*. Birmingham: TWC.
- Tomaszczyk, J. y B. Lewandowska. 1990. *Meaning and Lexicography*. Amsterdam: John Benjamins.
- Viegas, E., A. Ruelas, S. Beale y S. Nirenburg. 1998. "Extending a Core Lexicon Using On-Line Language Resources with Savoir-Faire". *Proceedings of the First International Conference on Language Resources and Evaluation (ELRA)*. Granada, 28-30 May. 97-104.
- Vossen, P. 1997. "EuroWordNet: a Multilingual Database for Information Retrieval". *Proceedings of the DELOS Workshop on Cross-language Information Retrieval*, March 5-7, Zurich.
- Walker, D.E., A. Zampolli, y N. Calzolari. 1995. *Automating the Lexicon I: Research and Practice in a Multilingual Environment*. Oxford: OUP.
- Walker, D. E. 1993. "The Ecology of Language". Eds. A. Zampolli, N. Clazolari y M. Palmer. 359-376.
- Wilks, Y. et al. 1996. *Dictionaries, Computers and Meanings*. Cambridge, Mass: The MIT Press.
- Zampolli, A., N. Calzolari y M. Palmer, eds. 1994. *Current Issues in Computational Linguistics: In Honour of Don Walker*. Giardiani Editori e Stampatori. Pisa.