3

LINGÜÍSTICA COMPUTACIONAL Y LINGÜÍSTICA DE CORPUS. POTENCIALIDADES PARA LA INVESTIGACIÓN TEXTUAL

Chantal Pérez Hernández
(mph@uma.es)
Antonio Moreno Ortiz
(amo@uma.es)
Departamento de Filología Inglesa, Francesa y Alemana
Universidad de Málaga

El concepto de «biblioteca digital» —lo hemos visto en el capítulo de Eva Méndez— implica algo más que una simple colección de objetos digitales. Una de sus características definidoras viene dada, justamente, por la posibilidad que ~~estas~~ ofrecen de optimizar el uso y explotación de sus contenidos digitales utilizando herramientas de carácter computacional. Recordemos, en este sentido, la definición de Lorcan Dempsey (2004): «cualquier colección de recursos digitales gestionados con el principal objetivo de maximizar la utilidad de las colecciones a una comunidad de usuarios». Alejandro ~~Bia~~, en el capítulo precedente, extiende con más detalle este aspecto de las bibliotecas digitales, conocido como «servicios de valor añadido», y destaca el papel que a este respecto desempeña la lingüística computacional.

Efectivamente, una de las posibilidades de explotación más interesante~~s~~ de estos recursos la encontramos en las tecnologías y herramientas que la lingüística computacional desarrolla en el marco de sus investigaciones. La lingüística computacional es una disciplina cuyos orígenes se remontan en el tiempo, y, aunque surgida dentro de un contexto de actuación vinculado a los estudios específicamente lingüísticos, actualmente está viendo redefinidas y actualizadas sus aplicaciones debido a las oportunidades que ofrece de optimizar el uso y explotación de los contenidos textuales digitalizados y de la propia Web.

Teniendo en cuenta su relevancia, y dado que constituye uno de los factores fundamentales en los que se asienta la aproximación a la teoría y literatura artística que el proyecto Atenea propone, consideramos conveniente dedicarle un capítulo específico que sirva como marco general en el que contextualizar la metodología y los resultados que se exponen más adelante. Empezaremos definiendo brevemente algunos conceptos generales que nos sirvan para definir el horizonte epistemológico y metodológico en el que se sitúa esta vertiente de las investigaciones en el ámbito de la teoría y la literatura artística.

1. LINGÜÍSTICA COMPUTACIONAL: ¿QUÉ ES?

1.1. Definición y líneas de actuación

La *lingüística computacional* constituye un campo científico de carácter interdisciplinar, vinculado a la lingüística y a la informática, cuyo fin fundamental es la elaboración de modelos computacionales que reproduzcan distintos aspectos del lenguaje humano y que faciliten el tratamiento informatizado de las lenguas. A este amplísimo campo de estudio también se le denomina *lingüística informática*, *procesamiento del lenguaje natural* (Payrató, 1998: 108) e incluso *ingeniería lingüística*. A pesar de constituir una disciplina relativamente reciente, casi todas las aproximaciones a este campo reconocen dos líneas fundamentales de actuación e investigación: la *lingüística computacional teórica* y la *lingüística computacional aplicada*, a las que Gómez Guinovart (2000*a*: 221) añade una tercera: la *informática aplicada a la lingüística*.

La primera de estas líneas, la *lingüística computacional teórica*, se centra en la consecución de tres objetivos complementarios: la elaboración de modelos lingüísticos en términos formales e implementables; la aplicación de dichos modelos a los diferentes niveles de descripción lingüística; y la comprobación computacional de la congruencia del modelo y de sus predicciones. En segundo lugar, la *lingüística computacional aplicada* supone una orientación más tecnológica de la lingüística computacional, que se centra, a grandes rasgos, en el diseño de sistemas informáticos capaces de gestionar, comprender, producir y traducir enunciados orales y escritos en lenguaje natural, para lo que desarrolla aplicaciones informáticas que pueden agruparse en cuatro grandes categorías: 1) sistemas de consulta a bases de datos a través del lenguaje natural; 2) aplicaciones de las tec-

nologías del habla, como los sistemas de conversión de texto a voz; 3) herramientas para el procesamiento de textos para la elaboración, gestión y revisión de documentos (que se incluyen en la mayoría de los procesadores de textos actuales), o los programas de generación automática de resúmenes, los sistemas de extracción de información y los de catalogación documental automatizada; y, por último, 4) las herramientas orientadas al procesamiento de más de una lengua o de una lengua extranjera, como son las aplicaciones didácticas para la enseñanza de las lenguas, las herramientas de traducción (semi)automática, las memorias de traducción y las bases de datos terminológicas.

Finalmente, el campo de trabajo que se caracteriza por la aplicación de los ordenadores a la investigación lingüística, es decir, el estudio científico del lenguaje, es lo que Gómez Guinovart (2000*a*: 221) denomina *informática aplicada a la lingüística* o *lingüística informática*, expresión que puede aplicarse en un sentido amplio a todas las disciplinas de la lingüística que usan herramientas informáticas para sus estudios, sobre todo en el caso de la lingüística de corpus o la lingüística histórica computacional.

1.2. Breve recorrido histórico

Los orígenes de la lingüística computacional pueden ubicarse a finales de la segunda guerra mundial, cuando distintos equipos científico-técnicos de Estados Unidos y la Unión Soviética comenzaron a trabajar en diversos proyectos para elaborar programas de traducción entre el inglés y el ruso: los servicios de inteligencia y las fuerzas armadas de ambos países tenían un interés especial en esos proyectos y, por ese motivo, fueron los principales impulsores de dichos proyectos durante mucho tiempo.

Durante los años cuarenta y cincuenta se produjeron importantes avances en dos áreas que resultaron claves para las tecnologías de procesamiento de lenguaje natural: la teoría de autómatas, que se originó en los trabajos de Alan Turing (uno de los padres de la informática), y los modelos de teoría de la información, que surgieron de los trabajos de Claude Shannon (1948), quien aplicó la teoría de la probabilidad de procesos de Markov para desarrollar autómatas que procesaran el lenguaje humano. A finales de los años cincuenta, las investigaciones fueron concentrándose en dos campos: el simbólico y el estocástico. Dentro del primero pueden mencionarse dos corrientes importantes: aquella que se interesó

principalmente por el análisis sintáctico, liderada por Noam Chomsky, otros lingüistas formales y científicos computacionales; y la orientada a la inteligencia artificial, en la que destacan Marvin Minsky (1968) y Claude Shannon. El campo estocástico ha estado representado principalmente por los ingenieros informáticos, quienes trabajan mediante estadísticas y probabilidades, y de cuyas investigaciones surgió el método de Bayes para el reconocimiento óptico de caracteres, entre otras muchas aplicaciones de aprendizaje automático. A pesar del interés creciente que suscitaba el desarrollo del estudio computacional del lenguaje y sus aplicaciones prácticas, la traducción automática sufrió un revés importante cuando en 1965 la Academia de Ciencias publicó un informe en el que se describían los escasos resultados obtenidos hasta ese momento (ALPAC, 1964). Como consecuencia, disminuyeron drásticamente los fondos para las investigaciones y la traducción automática se limitó a unos pocos proyectos en Europa y Asia. En las décadas siguientes el interés se dirigió fundamentalmente a la construcción de *corpora* textuales —especialmente en inglés y de muy pequeño tamaño comparados con los que se pueden usar hoy día— y al desarrollo de distintos lenguajes de programación con la ayuda de la lingüística teórica (uno de los más relevantes fue PROLOG) y de distintos programas para el análisis morfológico y sintáctico (lematizadores y desambiguadores).

Indudablemente, a partir de los años noventa la revolución de la World Wide Web tuvo como efecto principal la necesidad de perfeccionar las tecnologías para el procesamiento automático del lenguaje, por lo que en la actualidad numerosas empresas y centros académicos del mundo trabajan afanosa y competitivamente en este campo.

Actualmente, los lingüistas computacionales desarrollan productos informáticos para el análisis automático de la fonética, la fonología, la morfología, la sintaxis y la semántica. Otro campo de relevancia dentro de la lingüística computacional, que entronca con el de la inteligencia artificial, es el de la representación formalizada del conocimiento por medio de la construcción de ontologías, que son representaciones semánticas independientes de la lengua, y que actualmente tiene un amplio campo de aplicación en el desarrollo de la Web 3.0 o Web semántica. En una ontología, definida como la representación explícita de una conceptualización (Gruber, 1995: 908), se especifican un conjunto de tipos de conceptos y sus relaciones, organizados y representados formalmente para su uso computacional, que describen formalmente el conocimiento de un ámbito de especialidad determinado.

En resumen, los lingüistas computacionales intentan elaborar sistemas que hacen posible el diálogo entre personas que hablan lenguas diferentes o de carácter especializado (por ejemplo, traducciones automáticas y bases de datos terminológicas), o entre humanos y máquinas (sistemas expertos, sistemas de diálogo), y, en definitiva, se encargan de producir tecnología y metodologías encaminadas al tratamiento de las diversas lenguas en todas sus facetas. Estas aplicaciones, nacidas de la relación entre lingüística e informática, pueden ordenarse según el grado de complejidad que demandan sus objetivos (Cabré, 1993). En un primer nivel, pueden mencionarse aquellas aplicaciones que se limitan a emplear los datos lingüísticos como meras formas, sin ningún tipo de manipulación, como, por ejemplo, los sistemas de tratamiento de textos, sistemas de edición automática, etcétera. En segundo término, se ubican las herramientas lingüísticas automatizadas que emplean personas relacionadas profesionalmente con el lenguaje y la comunicación; por ejemplo, sistemas de gestión de bases de datos, diccionarios automatizados, sistemas de traducción, redacción, corrección o enseñanza asistidos por ordenador. Otro tipo de aplicaciones está representado por los sistemas automáticos que manipulan los datos para analizarlos o para transformarlos en datos de otras características, como analizadores, verificadores, lematizadores, clasificadores, programas de tratamiento estadístico. Por último, en el nivel más alto de complejidad se sitúan las herramientas avanzadas que actúan, o simulan actuar, con «inteligencia» y son capaces de sustituir en alguna medida la intervención humana, como los generadores de textos, los sistemas de traducción automática o los sistemas de vaciado terminológico de textos.

2. LINGÜÍSTICA COMPUTACIONAL Y LINGÜÍSTICA DE CORPUS

Un área básica de actuación de la lingüística computacional y cuyo desarrollo es requisito básico para que las demás aplicaciones computacionales que se diseñen para el procesamiento del lenguaje natural sean útiles, es precisamente la del estudio del uso y la estructura lingüística, lo que Gómez Guinovart (2000*a*: 201) denomina *lingüística informática* y cuyo máximo exponente es la *lingüística de corpus*, entendida como el estudio empírico de la lengua a partir de los datos que proporciona el análisis de ejemplos reales de producciones lingüísticas (orales o escritas) almacenadas en un ordenador. Cuanto mayor sea el conocimiento que tengamos de la estructura lingüística (a nivel morfológico, sintáctico, semántico

y pragmático) y del uso que los humanos hacemos de la lengua, mejores serán las aplicaciones informáticas que podamos diseñar para comunicarnos, para manejar, almacenar o extraer información o para interactuar con los ordenadores.

La idea de estudiar la lengua a partir de ejemplos reales de uso no es en absoluto nueva, aunque los años noventa trajeron un resurgimiento de los métodos empíricos y estadísticos de análisis lingüístico típicos de la década de los cincuenta (Church y Mercer, 1993). En aquellos años ya era práctica común el estudio de las unidades léxicas basándose en la coaparición con otras palabras.

También en los años cincuenta, J. R. Firth, figura eminente dentro de la tradición lingüística británica, publicaba *Papers in Linguistics*, donde este enfoque del 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 resurgimiento de los estudios de corte empírico es la disponibilidad creciente de cantidades masivas de datos en formato magnético. Hasta hace solo diez años, el corpus de un millón de palabras creado por Francis y Ku era en la Universidad de Brown parecía enorme. Hoy por hoy, muchos centros de investigación poseen *corpora* que contienen cientos o incluso miles de millones de palabras, y muchos otros de similar tamaño son accesibles a través de Internet.

2.1. El concepto de «corpus» y su definición

Definir *corpus*, tal y como se usa hoy en día en el ámbito de la lingüística o lexicografía de corpus, o en la lingüística computacional en general, no es tan sencillo como podría parecer a primera vista.

En principio, se puede llamar *corpus* a cualquier colección que contenga más de un texto (*corpus* como *cuerpo textual*). Sin embargo, existe cierto consenso en el seno de la comunidad científica relativo al hecho de que un corpus no solo ofrece información sobre sí mismo, es decir, sobre lo que contiene, sino que «representa» una sección más amplia de la lengua seleccionada de acuerdo a una tipología específica. La noción de «representatividad» aparece en otras definiciones recogidas, por ejemplo, en Tognini-Bonelli (1996: 45), en la que un corpus se

define como una colección de textos escogidos para caracterizar un estado o una variedad de una lengua.

La definición que ofrecen Atkins, Clear y Ostler (1992: 1) añade otro aspecto esencial en la creación de un corpus: este debe ser construido de acuerdo a una serie de criterios explícitos; y McEnery (2003) alude al hecho de que exista una coherencia en el proceso de recopilación y diseño:

> The term corpus should properly only be applied to a well-organized collection of data, collected within the boundaries of a sampling frame designed to allow the exploration of a certain linguistic feature (or set of features) via the collected data.

Siguiendo la definición de Santalla (2005: 45-46), podemos definir un corpus como un conjunto de textos de lenguaje natural e irrestricto, almacenados en un formato electrónico homogéneo, y seleccionados y ordenados de acuerdo con criterios explícitos para ser utilizados como modelo de un estado o nivel de lengua determinado, en estudios o aplicaciones relacionados en mayor o menor medida con el análisis lingüístico.

2.1.1. Representatividad del corpus: recopilación y diseño

Uno de los mayores caballos de batalla en lo que se refiere a la creación de un corpus son los criterios que deben guiar su diseño para que sea realmente representativo de la lengua que, valga la redundancia, representa. ¿Qué variedades de uso de la lengua debe incluir? ¿En qué proporción? ¿Cuál debe ser el tamaño de un corpus para que, realmente, represente una lengua, o, mejor dicho, el uso que sus hablantes hacen de ella? Este tipo de consideraciones son las que deben guiar los criterios de recopilación de los textos incluidos en el corpus. Aunque la literatura sobre este campo es extensa, la realidad es que hasta la fecha casi todos los *corpora* se han diseñado con criterios internos al proyecto en cuestión, y solo en determinados casos (British National Corpus, Birmingham Collection of English Text, corpus Cumbre, corpus ARTHUS) se han hecho públicos los criterios de selección de los textos incluidos en el corpus.

Representatividad, estandarización y tipología de los *corpora* han sido tres de los temas más debatidos entre la comunidad científica, con opiniones diversas recogidas en varios artículos y propuestas, algunas de ellas hechas en el seno de importantes proyectos europeos (Atkins, Clear y Ostler, 1992; Biber, 1993;

Quirk, 1992; EAGLES, 1996*a*, 1996*b*). En EAGLES (1996*a*: 4), por ejemplo, Sinclair define unos criterios mínimos que deben cumplirse para que un conjunto de textos en formato electrónico pueda ser considerado un corpus (cantidad, calidad, simplicidad y documentación), y clasifica los diferentes tipos de *corpora* que pueden existir, para así diferenciarlos de las colecciones de textos o los archivos (*archives*), ya que estos últimos no cumplen alguna de ellas.

Biber (1993: 243) define la representatividad como el grado en el que una muestra incluye la variabilidad de toda una población e identifica a continuación una serie de criterios externos e internos para la compilación de un corpus. Los primeros (criterios externos) son esencialmente criterios no lingüísticos que determinan el tipo de género, modalidad, origen y finalidad de los textos que han de incluirse. Estos pueden identificarse con relativa facilidad antes de la construcción del corpus y son los que se encargan de que el corpus represente una variedad suficiente de contextos situacionales, por lo que se consideran externos. Un segundo grupo de criterios identifica tipos diferentes de textos, de acuerdo con una serie de categorías lingüísticas (distribución de pronombres, proposiciones o tiempos verbales, por ejemplo). Estos criterios son internos a los textos que componen el corpus. El proceso de compilación, además, debe, según Biber (1993: 256), ser cíclico, de forma que primero se debe construir un corpus «piloto» para estudiar su composición y decidir qué parámetros del diseño deben ser modificados. Esta misma concepción cíclica en la compilación del corpus aparece reflejada en el trabajo de Tognini-Bonelli (1996*b*: 73), en el que señala que el diseño del corpus debe ser revisado continuamente así como los resultados del análisis de los datos evaluados, de forma que puedan modificarse algunos de los criterios de diseño, si el lingüista lo considera necesario.

2.1.2. TIPOLOGÍA DE LOS *CORPORA*

Un aspecto íntimamente relacionado con el de la representatividad es el de la tipología de los *corpora*. Atkins, Clear y Ostler (1992: 1) distinguen cuatro tipos fundamentales de lo que ellos denominan genéricamente *colecciones textuales* (*text collections*):

- archivos (*archives*): un repositorio de textos en formato magnético en el que los textos no están relacionados ni coordinados de forma alguna, como, por ejemplo, el Oxford Text Archive;

- bibliotecas de texto en formato magnético (*ETL*, *electronic text library*): una colección de textos en formato magnético que poseen un formato estandarizado y siguen ciertas convenciones en cuanto al contenido, pero sin rigurosas limitaciones de selección;
- corpus: una sección de una *ETL*, creada siguiendo unos criterios de selección explícitos y con un propósito específico; por ejemplo, el corpus de Cobuild o el de Longman/Lancaster;
- subcorpus: una porción de un corpus, ya sea un componente estático de un corpus mayor o más complejo, o una selección que se haga de forma dinámica *online* (mientras se está consultando el corpus).

La distinción propuesta por Atkins, Clear y Ostler puede usarse para diferenciar un corpus de otras colecciones genéricas de texto en formato magnético, basándonos en la aplicación de unos criterios rigurosos en la selección de los textos. Aun así, esta definición no puede usarse para clasificar la gran variedad de *corpora* existentes, variedad que en la mayoría de las ocasiones viene dada por otra de las características que ellos asignan al corpus: este ha de ser creado con una finalidad precisa.

En EAGLES (1996*a*) encontramos una tipología de *corpora* más específica,[52] entre los que se encuentran los *corpora* de referencia (*reference corpus*): creados para que sean una muestra representativa de las variedades más importantes de una lengua, así como de sus estructuras y vocabulario generales, de forma que ofrezcan información lo más amplia posible sobre una lengua y puedan servir de base en la construcción de gramáticas, diccionarios y obras de referencia. El British National Corpus,[53] el Bank of English[54] y el CREA[55] son ejemplos de *corpora* de referencia. Un concepto diferente es el de «corpus monitor» (*monitor corpus*); en él se propone la creación de un corpus con un tamaño constante, en el que se van añadiendo continuamente materiales nuevos a la vez que se van eliminando cantidades equivalentes de material antiguo para ofrecer así al lingüista la posibilidad de observar cambios recientes en el uso de la lengua. Con el aumento en la capacidad de los ordenadores, la idea de «flujo de circulación» (*rate of flow*)

[52] Véase Pérez Hernández (2002) para un repaso de los proyectos más importantes de creación de *corpora* en lengua inglesa y española y la página creada por David Lee: <http://devoted.to/corpora>, donde se recogen enlaces a casi todos los proyectos existentes hoy día.

[53] <http://www.natcorp.ox.ac.uk/>.

[54] <http://www.collins.co.uk/Corpus/CorpusSearch.aspx>.

[55] <http://corpus.rae.es/creanet.html>.

ha ido tomando forma, y en la actualidad no se considera necesario poner límite al tamaño del corpus, siempre que crezca con una constitución que pueda considerarse equivalente a la de estadios anteriores y posteriores.

Otro tipo que suele distinguirse es el corpus oral (como el corpus MICASE[56] o corpus C-Oral-Rom),[57] aunque el informe EAGLES pone de manifiesto que no existe consenso sobre lo que debe considerarse un corpus oral. En algunos casos, se usa para hacer referencia a un corpus en el que se recogen conversaciones informales y espontáneas, que han tenido lugar sin la intervención de ningún medio de comunicación. En otros casos, el uso del término se amplía para referirse a cualquier tipo de lengua en la que los hablantes se comportan de forma oral, como, por ejemplo, en los textos escritos para ser hablados.

También suele hacerse una distinción entre *corpora* especiales, especializados y *corpora* diseñados con fines especiales. Pearson (1998: 45) pone de manifiesto que en ocasiones los términos *corpus especial* y *corpus especializado* se usan indistintamente, o no se hace distinción entre *corpus especializado* y *subcorpus*, aunque es necesario hacer una distinción clara entre estos tipos de *corpora*. Los *corpora* especializados constituyen un tipo de corpus especial, ya que este es el término con el que se suele hacer referencia a los *corpora* que se construyen para que sean representativos de una variedad lingüística específica o de algún tipo de sublenguaje o lengua especializada. Los *corpora* especializados que se crean para el estudio de la lengua usada para fines específicos y de la terminología usada en sublenguajes poseen características similares a las de los *corpora* de referencia (en cuanto a cantidad, calidad, simplicidad y documentación), aunque es indudable que el criterio de representatividad debe restringirse al del dominio de estudio específico para el que son creados, mientras que un subcorpus es cualquier porción seleccionada de un corpus mayor, sea este del tipo que sea. La expresión *corpus especial* suele usarse para describir *corpora* (normalmente pequeños) que han sido diseñados con algún propósito específico, como, por ejemplo, los que contienen lenguaje infantil o de hablantes no nativos.

También suele hacerse una distinción entre dos tipos de *corpora* bilingües (o multilingües): los *corpora* paralelos y *corpora* comparables. Los *corpora* paralelos (también llamados en ocasiones *bi-texts*) están compuestos por un texto y su traducción a una o varias lenguas, mientras que los *corpora* comparables (tam-

[56] <http://lw.lsa.umich.edu/eli/micase/index.htm>.

[57] <http://lablita.dit.unifi.it/cromdemo/>.

bién denominados *paired texts*) son aquellos que poseen características y composiciones similares, es decir, tipos similares de textos en más de una lengua, de forma que es posible establecer comparaciones interlingüísticas.[58]

En la última década, la capacidad de acceso a sitios web remotos (y de los documentos que en ellos se encuentran) a través de Internet ha hecho posible que existan dos tipos de iniciativas que tienen que ver con el avance de la www. Por un lado, es posible acceder y consultar algunos de los proyectos de creación de grandes *corpora*, con casi la misma facilidad que si tuviéramos los textos instalados en nuestro ordenador y los procesáramos con un programa de manejo de corpus. Es el caso de, por ejemplo, Corpus del Español,[59] un corpus de más de cien millones de palabras procedentes de 20 000 textos del español de los siglos XIII al XX, al que se puede acceder para hacer varios tipos de búsquedas a través de Internet, o el Banctrad (Badia y otros, 2002), un proyecto que proporciona, vía web, acceso a *corpora* alineados en varias lenguas europeas.

Otro tipo de iniciativas ofrecen igualmente acceso a través de Internet a un corpus textual, pero en este caso, en vez de ser un corpus creado para tal propósito, es toda la información que existe en la www en formato textual la que se toma como base para las búsquedas. Es lo que se conoce como *la Web como corpus* (*Web as a corpus*; v., por ejemplo, Morley, 2006, donde se detalla la iniciativa denominada Webcorp,[60] que permite realizar búsquedas de textos disponibles en la www). Como nos cuenta Eva Méndez en su capítulo, existen tendencias similares que conciben la Web como la gran biblioteca virtual. Sin embargo, y a pesar de la inmensa riqueza y cantidad de textos a los que es posible acceder a través de iniciativas de este tipo, presentan, por supuesto, bastantes problemas, como el de la calidad y relevancia de los ejemplos que se pueden extraer (v. Kilgariff y Grefenstette, 2003).

Según vemos, la variedad de *corpora* existentes se debe en la mayoría de los casos a una de las características que Atkins, Clear y Ostler (1992) señalaban

[58] Los *corpora* paralelos más usados hoy día proceden de organismos oficiales de comunidades bilingües, donde gran parte de los documentos publicados deben aparecer en todas las lenguas oficiales de la comunidad, como es el caso del Parlamento canadiense, donde, por ley, las intervenciones de los representantes pueden hacerse indistintamente en inglés o en francés, pero las transcripciones de las sesiones (*Canadian Hansards*) han de conservarse en ambas lenguas, de modo que un equipo de traductores se encarga al final de cada sesión de traducir las intervenciones de uno a otro idioma. Un ejemplo de corpus comparable puede encontrarse en el proyecto NERC (Network of European Reference Corpora), una iniciativa europea de construcción de *corpora* de idénticas características y composición en todas las lenguas de la Unión Europea.

[59] <http://www.corpusdelespanol.org/>.

[60] <http://www.webcorp.org.uk/>.

como fundamentales en un corpus: el hecho de que son creados para un propósito específico, ya sea este de carácter general (*corpora* de referencia) o mucho más restringido (*corpora* especial; por ejemplo, del lenguaje de los afásicos).

2.2. Lingüística computacional y análisis de *corpora*

Los estudios lingüísticos basados en corpus han hecho que algunos postulados lingüísticos que se habían sostenido durante años empiecen a cuestionarse. Sinclair (1991), por ejemplo, demuestra de forma muy convincente que en ocasiones formas diferentes de un mismo lema deben considerarse como unidades léxicas independientes, ya que su comportamiento sintáctico o su significado es diferente. La tradicional noción de «forma canónica» a la que se asigna un significado (o significados) para todas sus formas posibles no se corresponde a veces con la frecuencia y distribución que se encuentra en un corpus.

También se hace patente que no es posible separar el estudio léxico del estudio gramatical, ya que en la mayoría de los casos las estructuras sintácticas y las léxicas son interdependientes, de forma que no es posible separar el estudio léxico del sintáctico. En esta línea son de especial importancia los trabajos realizados sobre marcos colocacionales (*collocational frameworks*), concepto propuesto originalmente en Renouf y Sinclair (1991) y ampliamente desarrollado y aplicado al español por Butler (1998*a*; 1998*b*). Por otra parte, los estudios realizados por Stubbs (2001) sobre prosodias semánticas (*semantic prosodies*) indican que también es necesario replantearnos el uso tradicional de la palabra como unidad básica de significado.

En el estudio de la gramática de las lenguas, los trabajos del denominado *corpus-driven approach to grammar* (v. Hunston y Francis, 1999; Sinclair y Mauraren, 2006) ofrecen una perspectiva bastante diferente a la que se encuentra en las gramáticas tradicionales, sobre todo en lo referente al estudio léxico, su interrelación con los patrones sintácticos y la fraseología. Siguiendo esta misma orientación, la editorial Longman ha publicado una serie de gramáticas, como la *Longman Grammar of Spoken and Written English* (*LGSWE*), en la que, según se puede leer en la introducción, se hace una descripción detallada del «actual use of grammatical features in different varieties of English» (*LGSWE*, 1998: 4).

Los trabajos sobre equivalencia de traducción contenidos en Sinclair, Payne y Pérez (1996) y en Tognini-Bonelli (1996) muestran que, con más frecuencia de lo

que un diccionario bilingüe parece indicar, no es posible asignar un equivalente de traducción apropiado sin tener en cuenta el contexto situacional y el cotexto lingüístico en el que las palabras aparecen, por lo que es necesario ampliar la noción tradicional de equivalencia de traducción.

En el ámbito de la teoría y la praxis lexicográfica, el uso de los *corpora* informatizados cuenta ya con dos décadas de historia, comenzando con el proyecto Collins Cobuild, llevado a cabo de forma conjunta por la editorial Harper Collins y la Universidad de Birmingham. Su uso se ha extendido de tal forma que hoy por hoy casi todas las editoriales se han implicado activamente en la creación y uso de *corpora* con fines lexicográficos. De hecho, en el ámbito de la lexicografía es importante destacar ahora que los *corpora* se han convertido en una herramienta lexicográfica fundamental para el estudio de las diferentes acepciones de las entradas léxicas y para el estudio de las colocaciones y la fraseología (Baugh, Harley y Jellis, 1996; Sánchez y otros, 1995). También ofrecen información decisiva sobre las diferencias de uso entre la lengua oral y la escrita, los rasgos prosódicos y la frecuencia relativa de uso, tanto de determinadas palabras como de determinados significados de una palabra, información clave para la inclusión (o exclusión) de una acepción o una palabra en un diccionario.

Toda esta problemática lingüística a la que la lingüística de corpus trata de dar respuesta es perfectamente extrapolable al ámbito de los vocabularios específicos, que tienen los mismos problemas de desambiguación, delimitación de significados, asignación adecuada de equivalentes, etcétera, con las consiguientes dificultades interpretativas que se derivan de ello. De ahí la relevancia del uso de *corpora* en las investigaciones terminológico-lingüísticas desde el punto de vista de las propias disciplinas especializadas, y no solo como vertiente de los estudios lingüísticos centrados en lenguajes de especialidad o tecnolectos.

De hecho, la investigación basada en corpus ha supuesto el nacimiento de nuevos métodos de estudio en áreas tan diversas como la adquisición de conocimiento léxico, la elaboración de gramáticas, los estudios socioculturales, la estilística, la traducción automática, el reconocimiento del habla, la obtención de información, la lexicografía monolingüe y bilingüe, la construcción de diccionarios electrónicos o la compilación de lexicones computacionales y repositorios de información terminológica. Existen otras importantes áreas de estudio en las que la investigación basada en corpus está ofreciendo nuevas perspectivas y resultados prometedores; por ejemplo, los estudios sociolingüísticos y culturales (Leech y Fallon, 1992; Adel y Reppen, 2008) y las diversas aplicaciones hechas a la enseñanza de

la lengua (v. Sánchez y otros, 1995; Scott y Tribble, 2006). Asimismo, y como los estudios asociados al proyecto Atenea demuestran, las posibilidades de aplicación no son conclusas, sino que las herramientas computacionales y los enfoques conceptuales pueden reutilizarse para múltiples finalidades, como, por ejemplo, auxiliar la interpretación textual, la delimitación de categorías semántico-nocionales, diferencias conceptuales mediante terminología comparada, etcétera.[61]

2.3. Análisis cualitativo y cuantitativo de corpus. Algunos ejemplos

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 estos pueden servir para construir modelos estadísticos más complejos, que expliquen la evidencia hallada en el texto.

Estos dos tipos de análisis no deben considerarse excluyentes, sino más bien complementarios, ya que el análisis cualitativo, por un lado, ofrece una gran riqueza y precisión en las observaciones realizadas; los fenómenos poco frecuentes pueden recibir igual atención que los muy frecuentes. Por otro lado, el análisis cuantitativo puede ofrecer 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.

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 (Butler, 1985; Fielding y Lee, 1991; Charniak, 1993; Oakes, 1998). Existe también en el mercado un importante número de programas (tanto comerciales como gratuitos para fines académicos) con interfaces de usuario fáciles de manejar y a la vez muy versátiles y sofisticados, aunque la mayoría de las grandes editoriales y centros de investigación han desarrollado herramientas de análisis específicas para el corpus que poseen y que por tanto se adaptan perfectamente a cualquier tipo de información metatextual que se haya añadido a su corpus (información sintáctica y

[61] Véanse capítulos 12 y 13.

sobre la clase morfológica de las palabras, identificación del texto y especificaciones sobre su procedencia, tipo o variedad lingüística a la que pertenece, etcétera), y además suelen adaptarse y desarrollarse para satisfacer las necesidades específicas de los investigadores, ya sean lingüistas, lexicógrafos o terminólogos.[62]

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, y TACT, del departamento 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 conjunto de herramientas para el manejo de corpus desarrollado por Michael Scott para Oxford University Press, conocido como Wordsmith Tools,[63] aunque existen una infinidad de opciones.

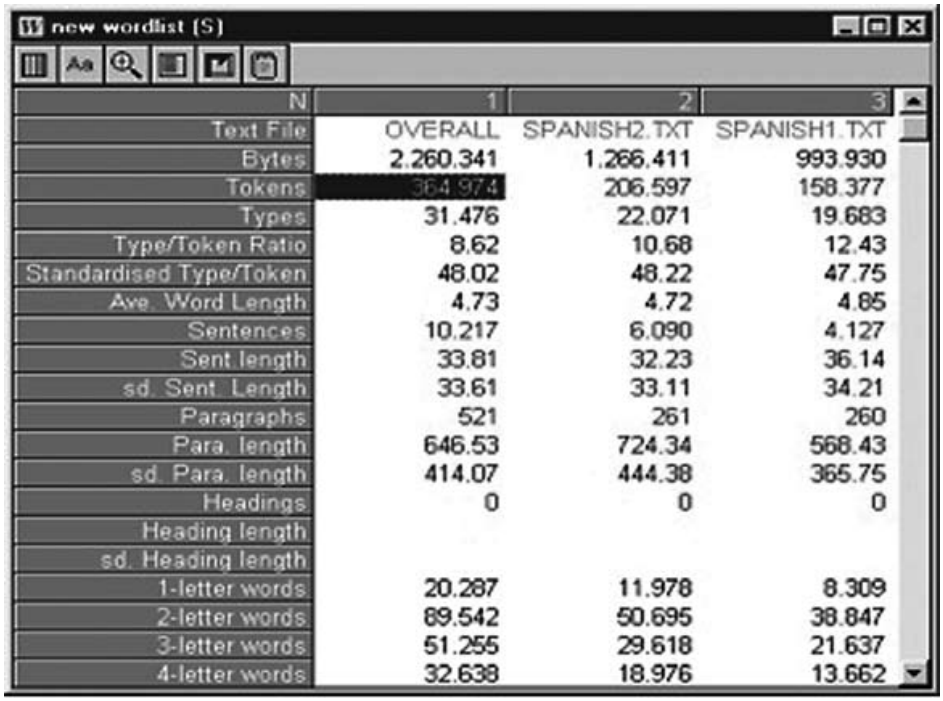
Casi todos los programas mencionados nos ofrecen las herramientas básicas de manejo de corpus, como, por ejemplo, la capacidad de realizar listas de las formas (*types*) que aparecen en un corpus, ordenadas de diferentes maneras, ya sea por orden alfabético, frecuencia, o en algunos casos por orden alfabético inverso; e índices estadísticos sobre el número de palabras, oraciones o párrafos y la longitud de estos.

Estas listas pueden ser de gran utilidad, por ejemplo en el ámbito de la lexicografía o en el estudio de los lenguajes de especialidad, ya que ayudan a decidir la lista de voces que han de incluirse en un diccionario, teniendo en cuenta su frecuencia de uso o para averiguar qué palabras son las más frecuentes en un dominio de especialidad. También pueden ofrecernos índices de frecuencia en los que muestre la ratio palabras/formas (*type/token*), es decir, el número total de palabras de un texto frente al número de palabras diferentes que aparecen en él, lo que es indicativo de la mayor o menor riqueza léxica de un texto; o comparar los índices en varios ficheros de texto. En la figura 1 mostramos una captación de pantalla tomada del programa Wordsmith Tools en la que se compara la lista de palabras y la ratio palabra/forma de dos ficheros de texto diferentes.[64]

[62] Véase la ya citada página de David Lee para una lista de los programas de manejo de corpus más usados: <http://devoted.to/corpora>.

[63] <http://www.lexically.net/wordsmith/index.html>.

[64] Este tipo de cálculo puede ser fundamental 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 manera que este puede subdividirse en seccio-



new wordlist (S)

N	1	2	3
Text File	OVERALL	SPANISH2.TXT	SPANISH1.TXT
Bytes	2.260.341	1.266.411	993.930
Tokens	364.974	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

FIGURA 1. Índices de frecuencia de dos ficheros realizados con Wordsmith Tools

Una aplicación concreta de este tipo de estudios se expone en el capítulo 12 en relación con textos de fray José de Sigüenza (1605) y fray Francisco de los Santos (1657).

Tanto Wordsmith Tools como TACT cuentan con una serie de herramientas para preprocesar el texto antes del análisis. Estas herramientas nos permiten 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 (Stopword Lists).

Otra de las herramientas de manejo de corpus más importantes y versátiles para el estudio lingüístico son los programas que 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

nes más pequeñas o *subcorpora*, 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.

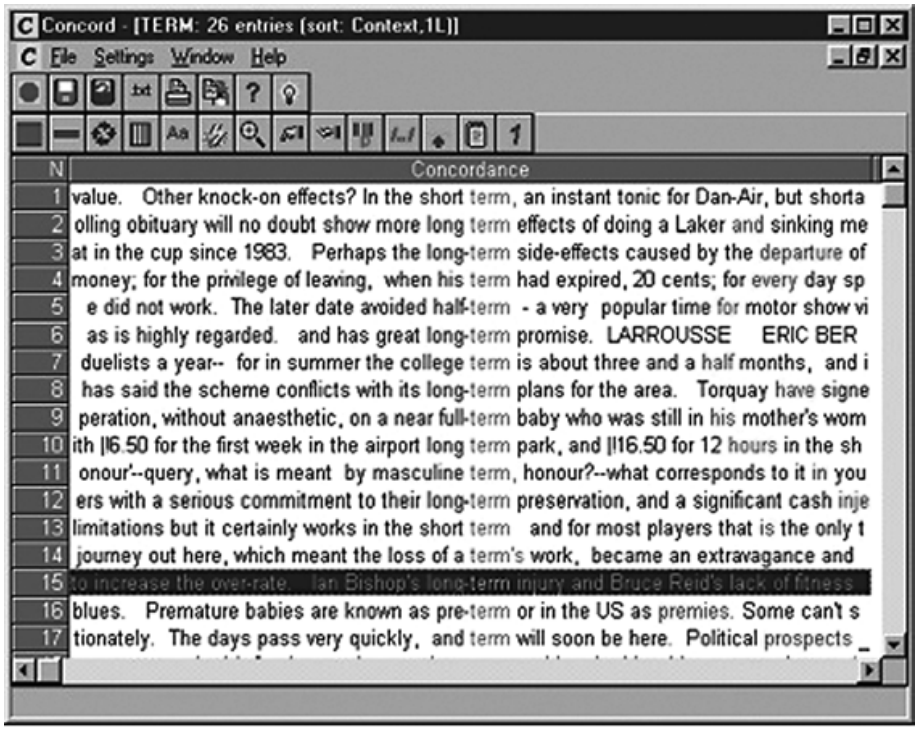
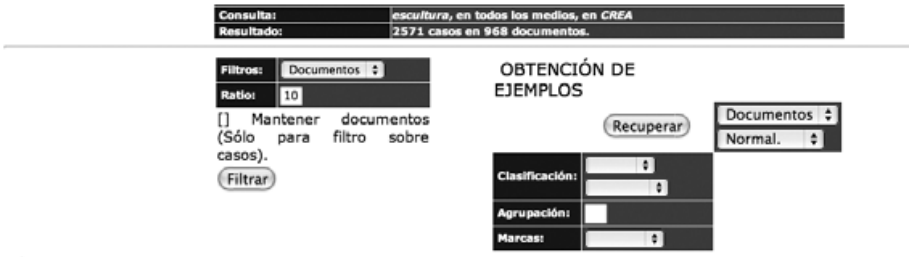


FIGURA 2. Líneas de concordancia extraídas con la utilidad Concord, de Wordsmith Tools

determinado (normalmente por el lexicógrafo) de caracteres de cotexto 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 es posible visualizar 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 permiten obtener un número determinado de líneas (cien, doscientas, 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, etcétera, 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).

La figura 2 es una captación 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. Estos diferentes tipos de ordenación permiten centrar nuestra atención en



Estadísticas

Año	%	Casos
1997	17.59	385
1996	16.91	370
1991	8.04	176
1992	6.35	139
1986	6.21	136
2000	5.53	121
2001	5.02	110
1989	4.34	95
1995	4.20	92
Otros	25.77	564

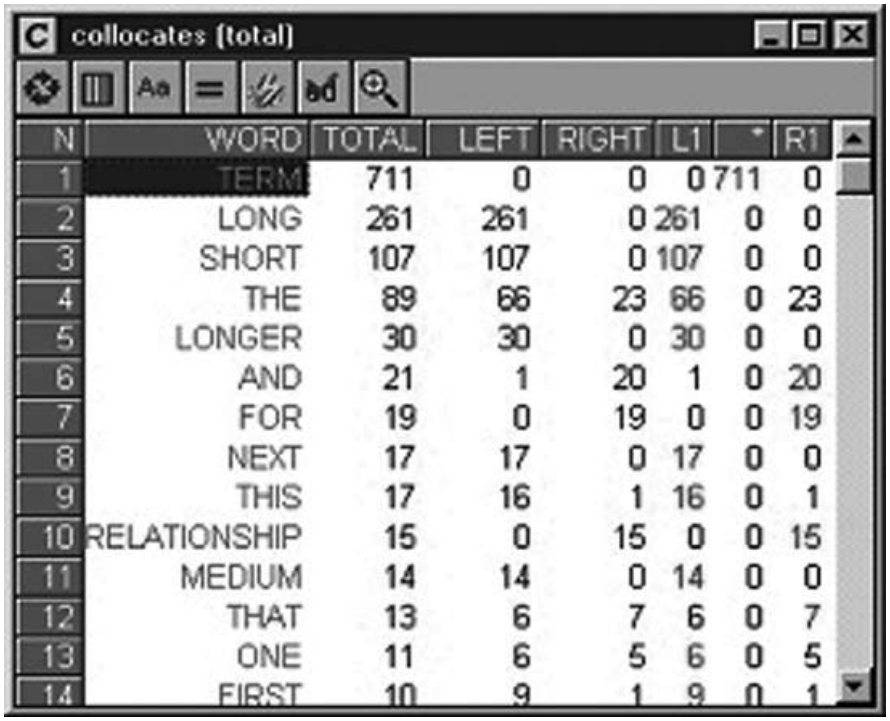
País	%	Casos
ESPAÑA	69.59	1756
ARGENTINA	9.27	234
MÉXICO	7.17	181
PERÚ	2.49	63
CHILE	1.94	49
ECUADOR	1.70	43
COLOMBIA	1.22	31
COSTA RICA	1.18	30
VENEZUELA	1.18	30
Otros	4.20	106

Tema	%	Casos
4.- Artes.	64.17	1650
2.- Ciencias sociales, creencias y pensamiento.	8.71	224
7.- Ficción.	8.51	219
5.- Ocio, vida cotidiana.	6.37	164
8.- Miscelánea.	6.26	161
3.- Política, economía, comercio y finanzas.	1.98	51
9.- Oral.	1.86	48
1.- Ciencia y Tecnología.	1.55	40
6.- Salud.	0.54	14

FIGURA 3. Estadísticas de aparición de *escultura* obtenidas del CREA

el cotexto 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 lleva un sustantivo determinado (por ejemplo, podríamos obtener listas de los distintos calificativos que determinan el sustantivo *manera* en un texto o en un conjunto de ellos); o, al revés, el tipo de sustantivos a los que acompaña un adjetivo determinado (por ejemplo, podríamos obtener los distintos sustantivos —y por tanto conceptos— que son determinados por los adjetivos *gracioso* o *caprichoso*). Muchos de estos programas permiten el uso de caracteres comodín (*wildcards*), con los que se pueden buscar diferentes formas de una misma palabra o realizar búsquedas difusas, múltiples y de frases idiomáticas con cierto grado de variación.

Con la mayoría de los programas que existen en el mercado también podemos identificar la fuente original de una línea de concordancia determinada, ampliar el cotexto 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. El CREA (Corpus del



collocates (total)

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

FIGURA 4. Colocaciones más frecuentes de la palabra *term* extraídas con Concord, de Wordsmith Tools

Español Actual),[65] por ejemplo, ofrece (en su acceso a través de la web) la posibilidad de ver los índices de aparición de una palabra determinada en las diferentes variedades del español y de tipos de textos que contiene. Para la palabra de búsqueda *escultura*, por ejemplo, ofrece la siguiente estadística:

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 obtención de información.

El estudio de los hábitos colocacionales de las palabras es uno de los caballos de batalla de las actividades relacionadas con la enseñanza y aprendizaje de la lengua, la traducción automática y la lexicografía, tanto monolingüe como bilingüe, y es especialmente relevante para la discriminación del significado en las

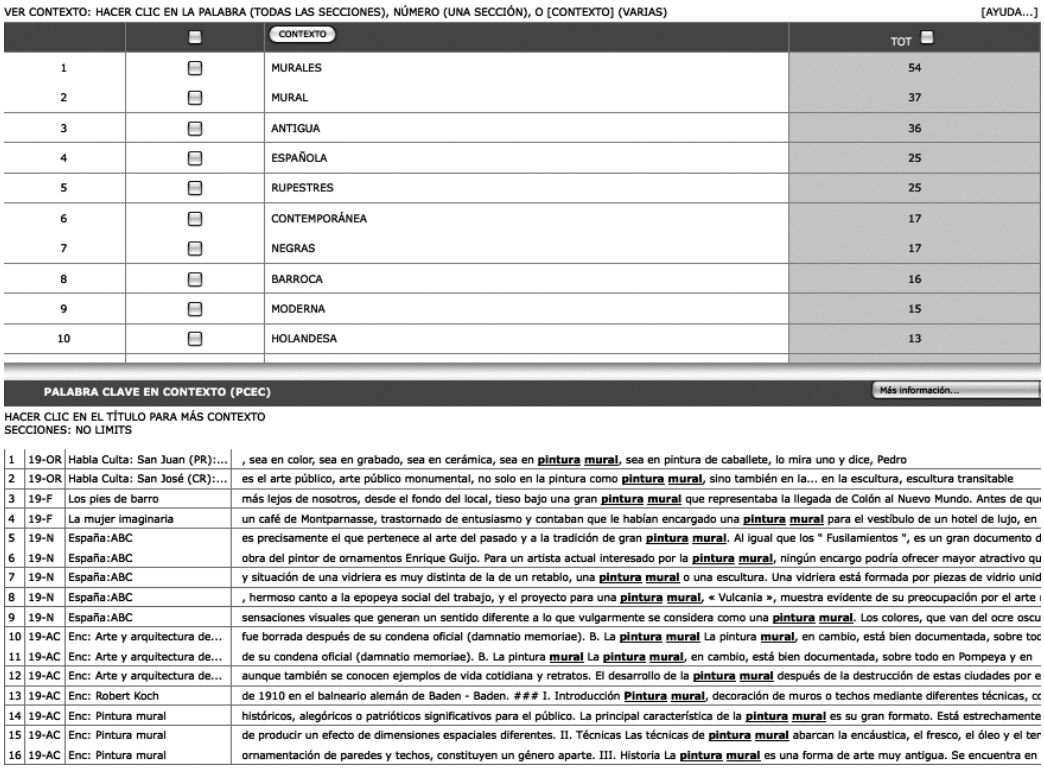
[65] <http://www.rae.es>.

palabras polisémicas.[66] Sin embargo, es una de las áreas en las que los estudiantes y 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. También es de gran utilidad en el campo de los lenguajes especializados, ya que en muchos casos determinadas palabras coaparecen con una serie muy diferente de palabras dependiendo de que se usen con su acepción en lengua común o en un discurso especializado.

Por esta razón, es muy útil contar con herramientas computacionales que ofrezcan listas 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 posición inmediatamente posterior y anterior), en relación a las líneas de concordancia que habíamos extraído anteriormente.

Algunos de estos cálculos estadísticos son muy útiles para el estudio de las colocaciones, 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 compara 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. Si lo consideramos desde el punto de vista semántico, la asociación fuerte de dos palabras puede darnos a entender que semánticamente se encuentran vinculadas o existe una relación consistente entre ellas. Si tenemos en cuenta que en el ámbito de los vocabularios especializados el plano semántico está constituido por los conceptos específicos de las disciplinas, la fuerza asociativa de dos palabras podría ser síntoma

[66] La noción misma de «colocación» ha sido entendida y definida de formas diferentes por diferentes autores. En términos generales, suele entenderse la coaparición (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. Corpas Pastor (1996: 53) define las colocaciones como «unidades fraseológicas que, desde el punto de vista del sistema de la lengua, son sintagmas completamente libres [...] pero que, al mismo tiempo, presentan cierto grado de restricción combinatoria determinada por el uso (cierta fijación interna)».



VER CONTEXTO: HACER CLIC EN LA PALABRA (TODAS LAS SECCIONES), NÚMERO (UNA SECCIÓN), O [CONTEXTO] (VARIAS) [AYUDA...]

		CONTEXTO	TOT
1		MURALES	54
2		MURAL	37
3		ANTIGUA	36
4		ESPAÑOLA	25
5		RUPESTRES	25
6		CONTEMPORÁNEA	17
7		NEGRAS	17
8		BARROCA	16
9		MODERNA	15
10		HOLANDESA	13

PALABRA CLAVE EN CONTEXTO (PCEC) Más información...

HACER CLIC EN EL TÍTULO PARA MÁS CONTEXTO
SECCIONES: NO LIMITS

1	19-OR	Habla Culta: San Juan (PR):...	, sea en color, sea en grabado, sea en cerámica, sea en **pintura mural**, sea en pintura de caballete, lo mira uno y dice, Pedro
2	19-OR	Habla Culta: San José (CR):...	es el arte público, arte público monumental, no solo en la pintura como **pintura mural**, sino también en la... en la escultura, escultura transitable
3	19-F	Los pies de barro	más lejos de nosotros, desde el fondo del local, tieso bajo una gran **pintura mural** que representaba la llegada de Colón al Nuevo Mundo. Antes de qu
4	19-F	La mujer imaginaria	un café de Montparnasse, trastornado de entusiasmo y contaban que le habían encargado una **pintura mural** para el vestíbulo de un hotel de lujo, en
5	19-N	España:ABC	es precisamente el que pertenece al arte del pasado y a la tradición de gran **pintura mural**. Al igual que los " Fusilamientos ", es un gran documento d
6	19-N	España:ABC	obra del pintor de ornamentos Enrique Guijo. Para un artista actual interesado por la **pintura mural**, ningún encargo podría ofrecer mayor atractivo qu
7	19-N	España:ABC	y situación de una vidriera es muy distinta de la de un retablo, una **pintura mural** o una escultura. Una vidriera está formada por piezas de vidrio unid
8	19-N	España:ABC	, hermoso canto a la epopeya social del trabajo, y el proyecto para una **pintura mural**, « Vulcania », muestra evidente de su preocupación por el arte
9	19-N	España:ABC	sensaciones visuales que generan un sentido diferente a lo que vulgarmente se considera como una **pintura mural**. Los colores, que van del ocre oscu
10	19-AC	Enc: Arte y arquitectura de...	fue borrada después de su condena oficial (damnatio memoriae). B. La **pintura mural** La pintura **mural**, en cambio, está bien documentada, sobre toc
11	19-AC	Enc: Arte y arquitectura de...	de su condena oficial (damnatio memoriae). B. La pintura **mural** La **pintura mural**, en cambio, está bien documentada, sobre todo en Pompeya y en
12	19-AC	Enc: Arte y arquitectura de...	aunque también se conocen ejemplos de vida cotidiana y retratos. El desarrollo de la **pintura mural** después de la destrucción de estas ciudades por e
13	19-AC	Enc: Robert Koch	de 1910 en el balneario alemán de Baden - Baden. ### I. Introducción **Pintura mural**, decoración de muros o techos mediante diferentes técnicas, co
14	19-AC	Enc: Pintura mural	históricos, alegóricos o patrióticos significativos para el público. La principal característica de la **pintura mural** es su gran formato. Está estrechamente
15	19-AC	Enc: Pintura mural	de producir un efecto de dimensiones espaciales diferentes. II. Técnicas Las técnicas de **pintura mural** abarcan la encáustica, el fresco, el óleo y el ter
16	19-AC	Enc: Pintura mural	ornamentación de paredes y techos, constituyen un género aparte. III. Historia La **pintura mural** es una forma de arte muy antigua. Se encuentra en

FIGURA 5. Lista de los adjetivos que muestran mayor frecuencia de coocurrencia con la palabra *pintura* (Corpus del Español)

de la posible vinculación o asociación de los conceptos que ellas designan. Así, por ejemplo, midiendo el grado de coocurrencia de las palabras *manera* y *gracia* —utilizadas con un sentido artístico—, podríamos deducir su mayor o menor vinculación conceptual.

Hay otros índices, como, por ejemplo, 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. La figura 5, por ejemplo, nos muestra una captación de pantalla

extraída del Corpus del Español donde se muestran los adjetivos que más frecuentemente coocurren con la palabra *pintura*, en posición +1. Como puede apreciarse, en la parte inferior de la pantalla es posible tener acceso a las líneas de concordancia que corresponden a cada combinación de «*pintura* + adj.»:[67]

Como vemos, existen infinidad de posibilidades en el análisis lingüístico de corpus y de programas para llevarlos a cabo. De los resultados de dichos análisis depende el éxito que tienen (y tendrán en el futuro) algunas de las aplicaciones que la lingüística computacional tiene en el análisis textual, de las que nos ocupamos brevemente a continuación.

3. LA LINGÜÍSTICA COMPUTACIONAL EN LA ERA DE LA INFORMACIÓN Y EL CONOCIMIENTO

El recurso más importante que posee la especie humana es el conocimiento, y su gestión y diseminación dependen del uso eficiente de la información necesaria para crearlo. Ambos términos, *información* y *conocimiento*, suelen usarse como sinónimos, aunque no lo son. La información es una de las características de la sociedad actual, y por ello se habla de que vivimos en medio de una «revolución de la información». La información se genera a partir de los datos, que deben procesarse para que su valor transcienda al del insumo parcial y quede establecida toda su significación (Cabrera Cortés, 2003). La información, por tanto, no es un fin en sí misma, sino que la cuestión clave es qué hacer con ella para transformarla en conocimiento. Esta transformación es un proceso humano de construcción, en el que se debe posibilitar la adquisición, selección, actualización y utilización de la información para tal fin.

En la época actual, del manejo eficiente de la información depende el uso de todos los otros recursos naturales, industriales y humanos. Durante la historia de la humanidad, la mayor parte del conocimiento se ha comunicado, guardado y manejado en la forma de lenguaje natural. La actualidad no es una excepción: el conocimiento sigue existiendo y creándose en forma de documentos, libros, artículos —aunque todos estos ahora se puedan guardar también en formato electrónico, o sea, digital—. Este es, precisamente, el gran avance: el que los ordenadores se hayan convertido en una ayuda enorme para el manejo de cantidades ingentes de

[67] Acceso a través de <http://www.corpusdelespanol.org> (consulta: diciembre del 2008).

texto, imágenes y otros formatos en los que almacenamos la información, esa que es la base de lo que ahora denominamos *la sociedad del conocimiento*. Sin embargo, lo que es conocimiento para nosotros, los seres humanos, no lo es para los ordenadores. Para ellos no son más que archivos, secuencias de caracteres que se pueden copiar, enviar, guardar o borrar (Gelbukh y Sidorow, 2006).

Ahí es donde la lingüística computacional ofrece su potencial servicio y se convierte en herramienta fundamental para el desarrollo y sustento de la sociedad de la información y el conocimiento basada en las telecomunicaciones (Internet) y condicionada por influyentes intereses culturales y comerciales. Mencionamos a continuación algunos de los ámbitos de actuación en los que la lingüística computacional es de mayor utilidad en el manejo de la información contenida en textos.

3.1. Búsqueda de la información y gestión de documentos

La aplicación de procesamiento de lenguaje natural más obvia y quizá más importante en la actualidad es la búsqueda y obtención de información (Gelbukh y Sidorow, 2006). Por un lado, Internet y las bibliotecas digitales contienen una cantidad enorme de conocimiento que puede dar respuestas a muchísimas preguntas que tenemos, pero la cantidad de información es tan grande que deja de ser útil al no poder ser encontrada fácilmente, y en muchas ocasiones la tarea consiste en decidir el grado de relevancia que un documento tiene para el usuario y ordenarlo según un criterio determinado. Por añadidura, el problema más complejo es entender la necesidad real del usuario, por qué formula su búsqueda y qué espera encontrar al formularla.

Las técnicas más usadas actualmente para la obtención de información implican la búsqueda por palabras clave: se buscan los archivos que contienen las palabras que el usuario teclea. Es decir, la representación formal usada es el conjunto de las cadenas de letras (palabras), usualmente junto con sus frecuencias en el texto (número de ocurrencias). Sin embargo, estos métodos son muy limitados y mejoran sus resultados complementados por medio de los siguientes métodos adicionales (Gelbukh, 2003):

- coincidencia de las formas morfológicas de palabras: buscando *hacer*, encontrar *hecho*, con la ayuda de generadores morfológicos;

- coincidencia de los sinónimos, conceptos más generales y más específicos: buscando *perro*, encontrar *chucho*, *mascota*, *animal*, etcétera, usando recursos léxicos jerarquizados semánticamente, aunque uno de los problemas es el de medir las distancias en esas jerarquías para establecer la distancia semántica que separa dos palabras;
- tener en cuenta las relaciones entre las palabras en la petición del usuario y en el documento: buscando *estudio de planes*, rechazar como no relevante *planes de estudio*. Para lograr este grado de calidad, se necesita reconocer (automáticamente) la estructura del texto y representarla en forma que permita la comparación necesaria, por ejemplo, en la forma de grafos conceptuales (Montes y Gómez y otros, 2001).

3.2. Generación de resúmenes y minería de textos (text mining)

Otro modo de filtrar la información relevante es la presentación resumida de un documento por medio de la generación automática de resúmenes de textos o colecciones de textos. Se trata de analizar un texto completo (o una colección grande de textos) y generar un informe corto de todo lo relevante que dicen estos textos. Así se da al lector una idea de su contenido sin la necesidad de que él tenga que leerlos completos.

Existen diferentes variantes de la tarea de generación de resúmenes. Por ejemplo, se puede buscar la opinión prevaleciente (más común) sobre un tema determinado. Digamos, hay muchos artículos sobre el procesamiento de lenguaje natural, pero ¿cuáles son los problemas que más se discuten?, ¿cuáles son las soluciones que más frecuentemente se proponen? Por supuesto, esto lo podemos extrapolar a cualquier ámbito del conocimiento especializado, como es la historia y la teoría del arte, por indicar un ejemplo que cuadra con el discurso de este libro. Una variante de la generación de resúmenes es la generación de resúmenes temáticos del texto: presenta un breve informe sobre los temas (aunque no las ideas) que se discuten en un texto dado (Gelbukh y otros, 1999). Por ejemplo: un texto habla sobre guerra, política y narcotráfico; otro texto habla sobre ciencia, tecnología y transporte. A pesar de la menor riqueza de esta presentación —en comparación con los resúmenes completos—, tiene algunas ventajas: es más simple de obtener y, como consecuencia, da resultados más seguros y estables; además, permite realizar operaciones matemáticas con los conjuntos (vectores) de temas obtenidos (Gelbukh y otros, 1999).

La minería de textos (*text mining*) es otra de las áreas de la lingüística computacional que están siendo de gran utilidad para facilitar el procesamiento automático de la semántica del lenguaje natural. Es una disciplina englobada en el ámbito de las técnicas de acceso, obtención y organización de información y consiste en un conjunto de técnicas que nos permiten extraer información relevante y desconocida de manera automática dentro de grandes volúmenes de información textual, normalmente en lenguaje natural y por lo general no estructurada. Lo más destacado es que con ello se permite tener acceso a conocimiento que no existía explícitamente en ningún texto de la colección, pero que surge de relacionar el contenido de varios de ellos (Hearst, 1999; Kodratoff, 1999).

Su objetivo es, por tanto, descubrir tendencias, desviaciones y asociaciones entre una gran cantidad de información textual. Esto nos permite encontrar conocimiento significativo a partir de datos textuales sin estructurar. La minería de textos extrae información nueva, por lo que es algo totalmente distinto a una simple búsqueda de información, en la cual se busca información ya existente, no se extrae información nueva. Constituye, por tanto, una herramienta de gran utilidad, ya que alrededor de un ochenta por ciento de la información de las organizaciones está almacenada en forma de texto no estructurado.

Una de las principales características de la minería de textos es que, por lo general, la información no está estructurada, al contrario de lo que ocurre en la minería de datos (*data mining*), en la que la información suele extraerse de una base de datos, por lo que sí está estructurada. Esto hace que la extracción de información de una base de datos sea más sencilla, ya que las bases de datos están diseñadas para que sea posible el tratamiento automático de la información.

Este proceso consiste en dos etapas principales: una etapa de *preprocesamiento* y una etapa de *descubrimiento* (Tan y otros, 2006). En la primera etapa, los textos se transforman en algún tipo de representación estructurada o semiestructurada que facilite su posterior análisis, mientras que en la segunda etapa las representaciones intermedias se analizan con el objetivo de descubrir en ellas algunos patrones interesantes o nuevos conocimientos. Dependiendo del tipo de métodos usados en la etapa de preprocesamiento, es el tipo de representación del contenido de los textos construida; y dependiendo de esta representación, es el tipo de patrones descubiertos.

Montes y Gómez (2001) resume los tres tipos de estrategias empleadas en los actuales sistemas de minería de texto (tabla 1).

ETAPA DE PREPROCESAMIENTO	TIPO DE REPRESENTACIÓN	TIPO DE DESCUBRIMIENTOS
Categorización	Vector de temas	Nivel temático
Texto completo	Secuencia de palabras	Patrones de lenguaje
Extracción de información	Tabla de datos	Relaciones entre entidades

Tabla 1

Estos métodos limitan a un nivel temático o de entidad sus resultados, haciendo imposible descubrir cosas más detalladas como:

- consensos, que, por ejemplo, respondan a preguntas como: «¿Cuál es la opinión mayoritaria de los españoles sobre la guerra de Irak?»;
- tendencias, que indiquen, por ejemplo, si han existido variaciones en la postura de Rodríguez Zapatero con respecto a la educación;
- desviaciones, que identifiquen, por ejemplo, opiniones «raras» con respecto a la victoria de la selección española de fútbol en la Eurocopa del 2008.

Para mejorar la expresividad y diversidad de los descubrimientos de los sistemas de minería de textos se han usado, entre otros métodos, los grafos conceptuales (Sowa, 1984; 1999) para conseguir una mejor representación del contenido de los textos. La transformación de textos en grafos conceptuales es una tarea compleja vinculada al análisis sintáctico y semántico de los textos; algunos ejemplos de textos transformados automáticamente en grafos conceptuales incluyen partes de artículos científicos (Myeng y Khoo, 1994; Montes y Gómez y otros, 1999), de expedientes médicos (Baud y otros, 1998) o de casos legales (Chaudhary y otros, 2006), aunque hoy día también existen una gran cantidad de herramientas y aplicaciones comercializadas por empresas dedicadas al análisis de contenidos web, como, por ejemplo, el Nstein Technologies,[68] Polyanalyst[69] o IBM Intelligent Miner for Text.[70]

[68] <http://www.nstein.com/en/>.
[69] <http://www.megaputer.com/polyanalyst.php>.
[70] <http://www-01.ibm.com/software/data/iminer/>.

4. CONCLUSIÓN

En este brevísimo repaso de un campo tan amplio e interdisciplinar como es la lingüística computacional, hemos intentado mostrar la forma en que sus diferentes ámbitos de actuación pueden facilitar el tratamiento informatizado de las lenguas, algo de vital importancia para el desarrollo y avance de la sociedad de la información y el conocimiento basada en las telecomunicaciones. Como hemos visto, la aplicación de procesamiento de lenguaje natural más obvia y quizá más importante en la actualidad es la búsqueda y obtención de información, ya que tanto Internet como las bibliotecas digitales contienen una cantidad enorme de conocimiento a la que es preciso que tengamos acceso de forma eficiente. Para que dichas aplicaciones ofrezcan a sus potenciales usuarios los resultados esperados, es necesario que tengamos un profundo conocimiento de la estructura lingüística de las lenguas naturales y es en este sentido en el que las técnicas de explotación de los corpus textuales informatizados están suponiendo un gran avance para los sistemas de búsqueda documental y de minería de textos. Como hemos mostrado, cuanto mayor sea el conocimiento que tengamos de la estructura lingüística (a nivel morfológico, sintáctico, semántico y pragmático) y del uso que los humanos hacemos de la lengua, mejores serán las aplicaciones informáticas que podamos diseñar para comunicarnos, para manejar, almacenar o extraer información o para interactuar con los ordenadores.

Asimismo, a lo largo de estas páginas hemos querido demostrar cómo la lingüística computacional y la lingüística de corpus ofrecen herramientas de análisis y enfoques metodológicos inestimables para optimizar la investigación de los contenidos —lingüístico-terminológicos y semántico-nocionales— presentes en los textos, perfectamente reutilizables por las disciplinas especializadas que tienen en el texto y en sus informaciones su objeto de estudio primario.

BIBLIOGRAFÍA

Ädel, A.; R. Randi (dirs.) (2008): *Corpora and Discourse. The Challenges of Different Settings,* Amsterdam: John Benjamins.

alpac (1964): informe, en S. Nirenburg, H. Somers, Y. Wilks (dirs.) (2003): *Readings in Machine Translation,* Cambridge (Mass.): mit, 131-135.

Atkins, B.; J. Clear, N. Ostler (1992): «Corpus Design Criteria», *Literary and Linguistic Computing,* 7, núm. 1, 1-16.

Badia, T.; G. Boleda, C. Colominas, A. González, M. Garmendia, M. Quixal (2002): «BancTrad: a Web Interface for Integrated Access to Parallel Annotated Corpora», en

Proceedings of the LREC'02 Workshop on Language Resources for Translation Work and Research, Las Palmas, 28 de mayo.

BAUD, R., y otros (1998): *Extracting Linguistic Knowledge from an International Classification,* Nashville: Division of Biomedical Informatics, Vanderbilt University.

BAUGH, S.; A. HARLEY, S. JELLIS (1996): «The Role of Corpora in Compiling the Cambridge Dictionary of English», *International Journal of Corpus Linguistics,* 1, núm. 1, 39-60.

BIBER, D. (1993): «Representativeness in Corpus Design», *Literary and Linguistic Computing,* 8, núm. 4, 243-257.

BUTLER, C. (1985): *Statistics in Linguistics,* Oxford: Blackwell.

— (1998*a*): «Collocational Frameworks in Spanish», *International Journal of Corpus Linguistics,* 3, núm. 1, 1-32.

— (1998*b*): «Multi-word Lexical Phenomena in Functional Grammar», *Revista Canaria de Estudios Ingleses,* 36, 13-36.

BUTLER, Christopher (1990): «Language and Computation», en *An Encyclopaedia of Language,* Londres/Nueva York: Routledge.

CABRÉ, M. T. (1993): *La terminologia. Teoria, metodologia i aplicacions,* Barcelona: Empúries. (Trad. al castellano, Barcelona: Antártida, 1993.)

CABRERA CORTÉS, I. A. (2003): «El procesamiento humano de la información: en busca de una explicación», *Acimed. Revista Cubana de Profesionales de la Información y la Comunicación en Salud,* 11, núm. 6; <http://bvs.sld.cu/revistas/aci/vol11_6_03/aci05603.htm>.

CHARNIAK, E. (1993): *Statistical Language Learning,* Cambridge (Mass.): MIT.

CHAUDHARY, M.; C. DOZIER, G. ATKINSON, G. BEROSIK, X. GUO, S. SAMLER (2006): «Mining Legal Text to Create a Litigation History Database», en *Proceedings of Law and Technology,* Cambridge (Mass.), 208-217.

CHURCH, K.; P. HANKS (1990): «Word Association Norms, Mutual Information and Lexicography», *Computational Linguistics,* 16, núm. 1, 22-29.

CHURCH, K.; R. MERCER (1993): «Introduction to the Special Issue on Computational Linguistics Using Large Corpora», *Computational Linguistics,* 19, núm. 1, 1-24.

CORPAS PASTOR, G. (1996): *Manual de fraseología española,* Madrid: Gredos.

DOMÍNGUEZ BURGOS, A. (2002): «Lingüística computacional: un esbozo», *Boletín de Lingüística,* 18, 104-119.

EAGLES (1996*a*): «Preliminary Recommendations on Corpus Typology», documento EAGLES (Expert Advisory Group on Language Engineering) EAG-TCWG-CTYP/P.

— (1996*b*): «Text Corpora Working Group Reading Guide», documento EAGLES (Expert Advisory Group on Language Engineering) EAG-TCWG-FR-2.

FIELDING, N. G.; M. G. LEE (dirs.) (1991): *Using Computers in Qualitative Research,* SAGE.

FIRTH, J. R. (1957): «A Synopsis of Linguistic Theory, 1930-1955», *Studies in Linguistic Analysis* (Philological Society), vol. especial, 1-32.

GELBUKH, A. (2003): «A Data Structure for Prefix Search under Access Locality Requirements and its Application to Spelling Correction: Computación y Sistemas», *Revista Iberoamericana de Computación,* 6, núm. 3, 167-182.

— y G. SIDOROV (2006): *Procesamiento automático del español con enfoque en recursos léxicos grandes,* México: Centro de Investigación en Computación, Instituto Politécnico Nacional.

— y G. SIDOROV, A. GUZMÁN-ARENAS (1999): «A Method of Describing Document Contents through Topic Selection», en *SPIRE'99, International Symposium on String Processing and Information Retrieval* (Cancún, México, 22-24 de septiembre), IEEE Computer Society Press, 73-80.

GÓMEZ GUINOVART, J. (2000*a*): «Lingüística computacional», en F. G. Ramallo, E. X. Rei-Doval, P. Rodríguez (dirs.): *Manual de ciencias da linguaxe,* Vigo: Xerais, 221-268.

— (2000*b*): «Perspectivas de la lingüística computacional», *Novatica,* mayo-junio.

GRUBER, T. (2008): «Ontology», en Ling Liu, M. Tamer Özsu (dirs.): *Encyclopedia of Database Systems,* Springer.

GRUBER, T. R. (1995): «Toward Principles for

the Design of Ontologies used for Knowledge Sharing», *International Journal of Human and Computer Studies,* 43, núm. 5-6, 907-928.

Hearst, M. (1999): «Untangling Text Data Mining», en *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics,* 20-26.

Hunston, S.; G. Francis (1999): *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English,* Amsterdam/Filadelfia: John Benjamins.

Kehoe, A.; A. Renouf (2002): «WebCorp: Applying the Web to Linguistics and Linguistics to the Web», en *Proceedings of the WWW 2002 Conference,* Honolulú.

Kilgarriff, A.; K. Grefenstette (2003): «Introduction to the Special Issue on the Web as Corpus», *Computational Linguistics,* 29, núm. 3.

Kodratoff, Y. (1999): «Knowledge Discovery in Texts: a Definition, and Applications», en *Proccedings of the 11th International Symposium on Foundations of Intelligent Systems (ISMIS-99).*

Leech, G.; R. Fallon (1992): «Computer Corpora. What do they tell us about Culture?», *ICAME Journal,* 16, 29-50.

Marvin M. (1968): *Semantic Information Processing,* MIT Press.

Mcenery, T. (2003): «Corpus Linguistics», en R. Mitkov (dir.): *The Oxford Handbook of Computational Linguistics,* Oxford: Oxford University, 448-463.

— y A. Wilson (1996): *Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics,* Edimburgo: Edinburgh University.

Montes y Gómez, M. (2001): *Minería de texto. Un nuevo reto computacional. Memoria del 3.er Taller Internacional de Minería de Datos (MINDAT-2001),* México: Universidad Panamericana.

Morley, B. (2006): «WebCorp: a Tool for Online Linguistic Information Retrieval and Analysis», en A. Renouf, A. Kehoe (dirs.): *The Changing Face of Corpus Linguistics,* Amsterdam: Rodopi.

Myaeng, F.; H. Khoo (1994): *Linguistic Processing of Text for a Large-scale Conceptual Information Retrieval System, Lecture Notes in Artificial Intelligence 835,* Springer Verlag.

Oakes, M. P. (1998): *Statistics for Corpus Linguistics. Edinburgh Textbooks in Empirical Linguistics,* Edimburgo: Edinburgh University.

Payrató Giménez, L. (1998): *De profesión, lingüista: panorama de la lingüística aplicada,* Madrid: Ariel.

Pearson, J. (1998): *Terms in Context. Studies in Corpus Linguistics,* Amsterdam/Filadelfia: John Benjamins; vol. 1.

Pérez Hernández, M. Chantal (2002): *Explotación de los «corpora» textuales informatizados para la creación de bases de datos terminológicas basadas en el conocimiento,* Madrid: ELIES/Red Iris (Estudios de Lingüística del Español, 18).

Quirk, R. (1992): «On Corpus Principles and Design», en J. Svartvik (dir.): *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82* (Estocolmo, 4-8 de agosto de 1991), Berlín/Nueva York: Mouton de Gruyter, 457-469.

Renouf, A.; J. M. Sinclair (1991): «Collocational Frameworks in English», en K. Aijmer, B. Altenberg (dirs.): *English Corpus Linguistics,* Londres: Longman, 128-143.

Sánchez, A. (1995): *Cumbre: Corpus Lingüístico del Español Contemporáneo. Fundamentos, metodología y aplicaciones,* Madrid: SGEL.

— 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, núm. 2, 259-280.

Santalla del Río, M. P. (2005): «La elaboración de corpus lingüísticos», en M. Cal, P. Núñez, I. M. Palacios (dirs.): *Nuevas tecnologías en lingüística, traducción y enseñanza de lenguas,* Santiago de Compostela: Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 45-63.

Scott, M.; C. Tribble (2006): *Textual Patterns: Key Words and Corpus Analysis in Language Education,* Amsterdam: John Benjamins.

Shannon, C. E. (1948): «A Mathematical Theory of Communication», *Bell System Te-*

chnical Journal, 27, julio y octubre, 379-423 y 623-656.

SINCLAIR, J. M. (1991): *Corpus, Concordance, Collocation,* Oxford: Oxford University.

— y A. MAURANEN (2006): *Linear Unit Grammar: Integrating Speech and Writing,* Amsterdam: John Benjamins.

— y J. PAYNE, C. PÉREZ (dirs.) (1996): «Corpus to Corpus. A Study of Translation Equivalence», *International Journal of Lexicography,* 9, núm. 3.

SOWA, J. (1984): *Conceptual Structures: Information Processing in Mind and Machine,* Reading (Mass.): Addison-Wesley.

— (1999): *Knowledge Representation: Logical, Philosophical and Computational Foundations,* 1.ª ed., Thomson Learning.

STUBBS, M. (2001): *Words and Phrases. Corpus Studies of Lexical Semantics,* Oxford: Blackwell.

TAN, P.; M. STEINBACH, V. KUMAR (2006): *Introduction to Data Mining,* Addison-Wesley.

TOGNINI-BONELLI, E. (1996): *Corpus Theory and Practice,* Birmingham: TWC.