COMPILACIÓN Y EXPLOTACIÓN DE UN CORPUS CON FINES ESPECÍFICOS[1]

Inés Lareo Martín

Universidad de A Coruña

RESUMEN*: El propósito de este trabajo es mostrar las posibilidades que ofrecen los corpus electrónicos para las investigaciones lingüísticas y apoyar una de las opciones que este tipo de estudios nos ofrece, la compilación de un corpus específico. La inexistencia o falta de acceso a corpus ya existentes, no debe suponer un problema insalvable para llevar a cabo un estudio determinado. En los diferentes apartados de este artículo detallaremos los pasos seguidos para la compilación de un corpus de textos ingleses de ficción desde 1800 a 1930. Seguidamente se expondrá el método seguido para la extracción y almacenamiento de datos extraídos para un tema concreto de estudio: las colocaciones verbo-nombre en Inglés Moderno Tardío.*

Palabras clave: Diseño de Corpus, corpus lingüístico, compilación, colocación.

ABSTRACT*: The aim of this paper is to show the possibilities that electronic corpora offer for linguistic research and to bear out one of the choices corpus-based studies give, to compile a corpus for a specific purpose. The lack of that type of corpora or the impossibility of accessing them should not be an insuperable problem to achieve our goal. In the Following sections the steps taken for the compilation of a corpus formed by English fiction texts published between 1800 and 1930 will be described. The last section will be devoted to the process of data retrieval and data-base design for a specific study: verb-noun collocations in late Modern English.*

Keywords: Corpora design, linguistic corpora, compilation, collocation

1.- INTRODUCCIÓN

El uso de un corpus es imprescindible a la hora de analizar datos lingüísticos más allá de la pura introspección y, en el caso de un estudio histórico, se convierte en la única fuente de "informantes" a la que el lingüista puede acceder. Los textos, utilizados como fuente de información lingüística, nos pueden revelar peculiaridades o irregularidades de la lengua que los mismos hablantes nativos desconocen. Además, en los estudios diacrónicos, al no disponer de hablantes vivos, se convierten en la fuente central de información. Obviamente las nuevas tecnologías han jugado un papel importante en la transformación de estas herramientas de investigación y los criterios seguidos para la compilación de los actuales corpus computerizados han sido adaptados también a los nuevos tiempos (Johansson 1991; Svartik 1992). El concepto actual de corpus lleva implícita la idea de que éste pueda ser leído o interpretado por una máquina, normalmente un ordenador. Esta característica ha contribuido también a hacer posible la existencia de corpus cada vez más extensos, con el consiguiente aumento de vertientes de trabajo.

Este artículo presentará una posible opción a la hora de enfrentarse a un tema de investigación cuando los corpus existentes no reúnen los requisitos exigidos o no están todavía a disposición del público. La finalidad de este trabajo es presentar los pasos seguidos para la consecución de un fin, centrándonos en los criterios formulados para la compilación de un corpus específico. Además se mostrarán los pasos seguidos para el almacenamiento y

Recibido el 26-11-2006

RæL 5 (2006): 87-106
ISSN 1885-9089

análisis de los datos obtenidos.

En nuestro caso el interés se centró en el Inglés Moderno Tardío, y concretamente, en el tema de estudio de las colocaciones verbo-nombre en las que como variables pudiéramos encontrar *make*, *have*, *take* o *do* (Lareo 2006). Dado que el término *colocación* se presta a multitud de interpretaciones y que el interés de este trabajo no es presentar las diferentes perspectivas o teorías lingüísticas que han tratado este tema, diremos que nuestro concepto de colocación fue entendido desde la perspectiva de la Teoría Sentido-Texto (*TST*), propuesta por el profesor Igor Mel'c‡uk y sus colaboradores[2] (Mel'c‡uk 1995: 177; 1998: 29).

La finalidad de ese trabajo era comprobar si la frecuencia de uso de estas colocaciones continuaba el camino ascendente ya constatado en períodos anteriores de la historia (Brinton y Akimoto 1999; Claridge 2000; Moralejo 2003) y subrayado por otros autores también para los textos posteriores a 1800. Al mismo tiempo, se tuvieron en cuenta otras variables que podían arrojar resultados interesantes en el análisis y que, como explicaremos más adelante, fueron decisivos para la toma de decisiones en el proceso de compilación de nuestro corpus. Concretamente, con dos de ellas, el sexo y la distribución geográfica, esperábamos obtener datos interesantes.

Este artículo seguirá el siguiente esquema: el apartado 2 describirá todos los pasos seguidos para la creación de un corpus específico. Como ya hemos apuntado, el tema de análisis era el uso de las colocaciones verbo-nombre, más concretamente la influencia que ciertas variables extralingüísticas como el sexo del autor, su procedencia geográfica o su edad, podían ejercer sobre su uso. Por este motivo, los subapartados 2.1, 2.2, 2.3 y 2.4 se centrarán en los criterios seguidos para cubrir todas estas variables.

Por último, en el apartado 3 se explicarán los pasos seguidos para la extracción y almacenamiento de datos, dedicando el subapartado 3.1 a las herramientas utilizadas y los problemas que el propio tema de estudio conllevaba. En el 3.2 se mostrará el proceso de almacenamiento de datos para su futura explotación lingüística.

2.- PROCESO DE COMPILACIÓN

Para la realización del estudio sobre las colocaciones verbo-nombre en Inglés Moderno Tardío era necesaria la utilización de un corpus de textos ingleses que nos posibilitara la extracción y posterior análisis de datos procedentes de textos reales, permitiéndonos comprobar la frecuencia de uso, implicaciones extralingüísticas, evolución, primera aparición y, en algunos casos, desaparición de las colocaciones que estábamos investigando. A pesar de que el acceso a varios tipos de corpus es cada día más frecuente, la inexistencia de un corpus computerizado que se adecuara a nuestras necesidades y que abarcara el período entre 1800 y 1930 o la falta de disponibilidad pública de los que se encontraban en proceso de creación[3] motivó el comienzo de la tarea de recopilación. A tal efecto se creó un corpus de textos ingleses pertenecientes al siglo XIX y principios del XX en formato electrónico. De este modo se pudieron aprovechar también las ventajas que este medio aporta a estudios cuantitativos centrados también en aspectos sociolingüísticos (Guzmán-González 2005: 16).

El paso previo a la compilación de textos fue la toma de decisiones y el establecimiento de los criterios tanto de índole lingüística como extralingüística, que afectarían a la selección (Sinclair 2004 y De Smet y Smith 2005). Los aspectos que se tuvieron en cuenta fueron la restricción del período, su división en subperíodos, el tipo de textos, nacionalidad, sexo y edad de los autores y cantidad de palabras por muestra. Los siguientes subapartados se dedicarán a estos puntos cuyo propósito es exponer los parámetros establecidos con el propósito de crear un corpus equilibrado (Moskowich y Crespo 2007).

Recibido el 26-11-2006

RœL 5 (2006): 87-106
ISSN 1885-9089

2.1. Delimitación del período y su división

La elección del período se realizó teniendo en cuenta los estudios sobre colocaciones ya existentes. Consideramos que ciertos aspectos de este tipo de estructuras sintagmáticas han sido ya suficientemente estudiados en épocas anteriores como lo demuestran los trabajos de Akimoto y Brinton (1999), Hiltunen (1999), Matsumoto (1999), Kytö (1999), Claridge (2000), Iglesias Rábade (2001) y Moralejo (2003), entre otros. Sin embargo, el estudio de Minoji Akimoto (1999) puede ser considerado como un primer paso interesante para un análisis más profundo del Inglés Moderno dado que, aunque su investigación abarca los siglos XVIII y XIX, se centra sobre todo en el XVIII. Puesto que Akimoto utiliza también un corpus para su estudio, se decidió que en nuestro material no estuvieran incluidos los autores que ya habían sido seleccionados por Akimoto para su análisis[4], para dar más originalidad al trabajo y poder comparar, si fuera procedente, los datos de ambas investigaciones.

La elección del período (1800-1930) se llevó a cabo teniendo en cuenta la opinión de Görlach (2004: 2) y Beal (2004: 191) sobre la importancia e idoneidad de los siglos XIX y XX para las investigaciones dialectológicas y sociolingüísticas. También se tuvieron en consideración los cambios del siglo XIX (Bailey 1999; Görlach 1999; Beal 2004), su proximidad a lo que podemos considerar inglés actual y los escasos estudios sobre el uso de las colocaciones en esta época. La fecha de comienzo para analizar estas colocaciones en Inglés Moderno Tardío (1800) se estableció teniendo en cuenta las opiniones de Mele y Martín (2001: 582) y la fecha final estipulada por Osselton (1984) para su estudio sobre Inglés Moderno Temprano que abarca de 1500 a 1800. También se tuvieron en cuenta las opiniones de Woolf (1929: 34) sobre el resurgimiento de las mujeres escritoras en la Inglaterra de finales del XVIII y principios del XIX.

Con la intención de abrir un futuro camino a la investigación se consideró de interés la inclusión de algunos textos del siglo XX, que permitirían observar si las tendencias del siglo XIX se mantenían en el siglo siguiente. Con esta finalidad se incluyeron textos pertenecientes a la década inmediatamente posterior a la I Guerra Mundial (hasta 1930). Ésta fue considerada como fecha límite de este corpus, siendo conscientes de que el espacio entre guerras fue un período de cambios en todos los aspectos, pudiendo también tomarse este momento como el final del Inglés Moderno Tardío y el principio del inglés actual.

Debido a la amplitud del lapso temporal abarcado por el corpus se consideró oportuno dividirlo en subperíodos de 50 años para poder estudiar mejor los cambios. De esta manera los bloques temporales que componen el corpus son tres: de 1800 a 1850, de 1850 a 1900 y de 1900 a 1930. Dentro de cada bloque se incluyeron dos textos para los primeros 25 años de cada subperíodo y dos textos para los últimos 25 años. Después de revisar algunos de los trabajos sobre la compilación de corpus[5], nos pareció que esta subdivisión era la adecuada para apreciar los cambios que pudieran existir entre los diferentes textos y autores (Tabla 1).

PERÍODO	BLOQUE	FECHA DE PUBLICACIÓN
1º 1800-1850	1	1814
		1812
	2	1840
		1849
2º 1850-1900	3	1860
		1857
	4	1892
		1894
3º	5	1903

1900-1930	6	1902
		1927
		1928

Tabla 1: División temporal

2.2. Tipos de textos

Después de una serie de consideraciones sobre el tipo de textos que deberíamos introducir en nuestro corpus, y dada la naturaleza de los fenómenos lingüísticos que queríamos analizar, decidimos centrarnos en el género de ficción. Ya que somos conscientes de que el uso de etiquetas como género, registro, etc. ha suscitado y todavía suscita mucha polémica (Lee 2001), aclararemos brevemente nuestro concepto de "género de ficción". Con este término nos referiremos a textos literarios en prosa que se encuadren dentro de las novelas, relatos y cuentos. Al seleccionar sólo este género es evidente que no vamos a poder extraer resultados que se puedan hacer extensivos a la lengua en general, pero sí al estilo literario de esta época[6]. Consideramos que éstos son los textos que poseen una mayor variedad de vocabulario y mezcla de estilos debido a la habilidad de los autores al tratar de reproducir la forma de hablar de personajes de distinta clase social e incluir el uso de dialectos y variedades regionales (Bailey 1999; Beal 2004).

La elección de textos de ficción nos permitía también tener acceso a obras escritas por mujeres, menos frecuentes en otras ramas del saber por motivos evidentes. Virginia Woolf (1929: 35) comentó ya en su momento que la ficción y en concreto la novela eran para una mujer del siglo XIX las cosas más sencillas de escribir, pero con este adjetivo no se refería a los contenidos de las obras. Woolf hacía referencia al terreno que las mujeres iban ganando dentro de la sociedad intelectual, que se veía restringido y delimitado por su situación socio-histórica. Esta idea está latente en la siguiente cita de Woolf (1967: 60, citado en Cameron 1990) al referirse a Dorothy Osborne: "Had she been born in 1827, Dorothy Osborne would have written novels; had she been born in 1527, she would never have written at all. But she was born in 1627, and at that date though writing books was ridiculous for a woman there was nothing unseemingly in writing a letter".

Puesto que entre nuestros propósitos se encontraba la comparación de la frecuencia de uso de colocaciones entre ambos sexos, estas palabras apoyaron también nuestra decisión de decantarnos en esta época por las novelas o los cuentos dentro del género de ficción. Esta conjunción de hechos tuvo el peso suficiente como para inclinarnos por este género y este tipo de textos en beneficio de un enriquecimiento de nuestra investigación. En consecuencia, se excluyeron de nuestro corpus las obras de teatro y la poesía, debido también a sus características especiales.

Los textos fueron extraídos de bases de datos electrónicas. Las obras pertenecen en su mayoría a la *Chadwyck-Healey Collection*[7]. En los casos en los que no fue posible localizar en esta fuente textos que cumplieran nuestros requisitos, tanto desde el punto de vista del autor como de la fecha, se recurrió a otras fuentes como el *Project Gutenberg*[8], *wordtheque.com*[9] y *classicreader.com*[10].

Aunque la localización de todas las obras en formato electrónico no ha sido una tarea fácil, conseguimos recopilar textos que cumplieran nuestros requisitos hasta 1930. En algunos casos, las decisiones de selección de textos y autores estuvieron sujetas a la propia accesibilidad y disponibilidad de los textos.

2.3. Los autores

Para la selección de los autores se tuvieron en cuenta las ideas de Warner (1961: 80) al

hablar de la importancia que tienen todos los escritores para el estudio de un estilo desde el punto de vista histórico. Es más, en su opinión, la fama del autor no tiene que ser una garantía de su fidelidad a las normas de la lengua, sino más bien todo lo contrario. Esta afirmación la argumenta hablando de las licencias que se suelen permitir los escritores muy famosos, mientras que los menos conocidos son, por lo general, mucho más fieles a las reglas de la lengua. Apoyándonos en esta opinión procuramos que la selección de autores abarcara tanto novelistas muy conocidos como menos conocidos, pero siendo siempre conscientes y realistas en cuanto a la posibilidad de acceder a los textos. Como ya se ha dicho, antes de realizar la selección se descartó a los escritores ya incluidos por Akimoto en su corpus, eligiendo después nuestros autores basándonos en dos criterios principales: la nacionalidad, entendida como lugar de origen y el sexo. En cuanto al origen se eligieron escritores nativos británicos por dos motivos principales. En un principio, porque, como ya expuso Algeo (1995: 213), la mayor parte de estas colocaciones es considerablemente más utilizada en inglés británico que en inglés americano. Por otra parte, porque como constataron Hoffmann y Lehmann (2000), la habilidad y la intuición que muestran los hablantes nativos en relación a las colocaciones son muy superiores a las de los que utilizan el inglés como una segunda lengua aprendida posteriormente.

Estas decisiones se tomaron también para darle mayor coherencia al corpus y para dejar abierta otra vertiente de la investigación que posibilitaría la futura ampliación del corpus con escritores americanos. Los resultados de este estudio paralelo podrían compararse fácilmente con los nuestros, buscando las tendencias de las dos variedades diatópicas en el tema de las colocaciones.

Otra de las variables importantes a la hora de compilar nuestro corpus fue el sexo. Nos referiremos a la variable "sexo" sin entrar en la polémica suscitada entre los términos "género" y "sexo" (Bing y Bergvall 1996; Eckert y McConnell-Ginet 2003; Bratt y Tucker (Eds.) 2003). Es necesario, de todas maneras, aclarar que con este término no queremos hacer sólo una diferenciación biológica de los autores, sino que entendemos que incluye también otras diferencias como un determinado tipo de educación y pautas sociales. Dado que no pretendíamos hacer un estudio pormenorizado del grado de masculinidad y femineidad que encierran hombres y mujeres como muy bien señalan Eckert y McConnell-Ginet (2003: 251-53), sino que nuestro objetivo se centraba en comprobar si se podían establecer diferentes pautas en la frecuencia de utilización de nuestras colocaciones en los siglos XIX y principios del XX, dependiendo del sexo de los escritores, al igual que se ha hecho en trabajos sobre otros temas como Nurmi (1999) y Olsen (2005).

Conviene exponer las ideas que nos llevaron a seguir con la introducción de esta variable. En primer lugar partimos de las propuestas de Holmes (1999: 462). En este trabajo la autora extrae el siguiente universal sociolingüístico: "Women and men develop different patterns of language use". Ya Coates (1986; 1990: 65) había vislumbrado esta máxima cuando subrayó: "the fact that women and men differ in terms of their communicative behaviour is now a established sociolinguistic fact". Más tarde Milroy *et al.* (1994) y Milroy y Gordon (2004) coincidieron con las autoras anteriores al resaltar que el sexo del hablante provoca, en muchos casos, más diferencias lingüísticas que la clase social a la que pertenezcan. Esta idea fue puesta de manifiesto también por Romaine (2003: 109) al hablar de la situación de las mujeres británicas en períodos anteriores de la historia en los que incluso las mujeres de clase social alta no tenían acceso a las normas de la lengua escrita. Con esta situación se encontró Tanabe (1999) en su estudio de las cartas escritas por las mujeres y los hombres Paston, debido a la imposibilidad de comparar en esa época la lengua de hombres y mujeres aunque la clase social fuera la misma (Montoya 2003).

Basándonos en las premisas anteriores y puesto que la situación de las mujeres británicas de nuestro período histórico había cambiado ya con relación a siglos anteriores, se

estableció también una equiparación entre el número de escritores y escritoras. Opinamos que, en concreto en el círculo de la creación literaria, las mujeres se podían considerar equiparables social e intelectualmente a los hombres, lo que permite un estudio comparativo. Esta comparación se verá enriquecida, al mismo tiempo, por las diferencias en la manera de expresarse, el estilo y las vivencias específicas de las escritoras, resaltadas por Woolf (1929).

Siguiendo a Chambers (2003: 7-8) tuvimos en cuenta, además del sexo, otro de los factores que ejercen una influencia directa en nuestro comportamiento lingüístico: la edad. En su trabajo se reconocen tres etapas en las que la edad tiene una relación directa con nuestra manera de hablar: la niñez, la adolescencia y la madurez. La estabilidad lingüística se alcanza a partir de la edad adulta (Chambers 2003: 202) y las características propias de cada individuo se mantienen durante el resto de su vida. Estas premisas son las que se siguieron para delimitar la edad de los escritores, que, como se extrae de la fecha de nacimiento y la de publicación, se encontraban entre los 29 y 68 años en el momento de escribir la obra objeto de análisis.

El conjunto de autores seleccionados es por tanto un grupo compacto que puede representar perfectamente una parte de la sociedad de ese momento dado que, como puntualizan Tannen (2003) y Milroy y Gordon (2004), comparten un mismo bagaje cultural. Esta base cultural común no se refiere sólo al país de origen y lengua madre, sino también al sexo, edad y distribución geográfica. De este modo nuestro corpus nos permitiría investigar las posibles diferencias de uso de las colocaciones entre ambos sexos, edades y zonas de Gran Bretaña.

En cada uno de los subperíodos se delimitaron los dos bloques, incluyendo para cada uno de ellos dos textos que no estuvieran muy alejados en el tiempo tal como se dijo más arriba (ver Tabla 1), pero que hubieran sido escritos por un hombre y por una mujer que hubieran nacido y vivido la mayor parte del tiempo en las Islas Británicas. Aunque podemos ver en la Tabla 2 que dos de los escritores nacieron en Corfú y Calcuta, respectivamente, su educación y su vida adulta transcurrieron en ambos casos en las Islas Británicas, quedando su lugar de nacimiento como una mera anécdota. Del mismo modo, las estancias en el extranjero de algunos de los escritores no fueron consideradas relevantes por haber tenido lugar después de su período de formación. Por tanto, teniendo en cuenta los comentarios anteriores, la Tabla 2 recoge algunos datos biográficos básicos de los doce autores estudiados para la investigación como el lugar y año de nacimiento. Cuando el lugar se consigna entre paréntesis significa que no ha sido juzgado relevante para su posterior desarrollo lingüístico. En estos casos, la zona de mayor influencia aparece después de la fecha de nacimiento.

AUTORES			
MUJERES		HOMBRES	
NOMBRE	NACIMIENTO	NOMBRE	NACIMIENTO
Edgeworth, Maria	(Oxfordshire), 1767, Ireland	Scott, Sir Walter	Edinburgh, 1771
Gaskell, Elizabeth	London, 1810	Thackeray, William M.	(Calcuta), 1811, Cambridge
Brontë, Charlotte	Yorkshire, 1816	Collins, Wilkie	London, 1824
Somerville, Edith OE. y Martin Ross	(Corfu), 1858, Cork, Galway, 1862	Doyle, Sir A. Conan	Edinburgh, 1859
Nesbit, Edith Anne OEnone	London, 1858	Butler, Samuel	Nottinghamshire, 1835
Woolf, Virginia	London, 1882	Lawrence, David Herbert	Nottinghamshire, 1885

Tabla 2: Datos biográficos de los autores

Debemos aclarar que aunque la novela *The Real Charlotte* fue escrita por dos autores, Edith Somerville y Martin Ross[11], el hecho de que ambas fueran mujeres e irlandesas, nos animó a incluirla en nuestro corpus. El cumplimiento de estas dos condiciones hacia perfectamente posible el análisis bajo los parámetros preestablecidos. Así pues, las referencias correspondientes se harán utilizando sólo en adelante el nombre de Somerville.

2.4. Tamaño de la muestra

Para decidir la cantidad de palabras que debían componer cada muestra se consideró el tema objeto de estudio, las colocaciones verbo-nombre. Para realizar un análisis de unidades fraseológicas y de variables diatópicas era necesaria una muestra representativa de cada autor. Con esta finalidad, el volumen de palabras de cada texto que se estipuló para la creación del corpus fue de unas 75.000 por muestra. En los casos en los que los textos no tenían la extensión necesaria se eligió un extracto de otro texto del mismo autor y de una fecha próxima al texto ya seleccionado[12]. De este modo se equiparaba el número de palabras utilizadas de cada autor, permitiendo el estudio de idiosincrasias del propio autor y de la repercusión de algunos factores extralingüísticos.

El volumen total de palabras del corpus final asciende a 900.000 repartidas de la siguiente manera: 449.825 (49,98%) en textos de escritores y 450.175 (50,02%) en textos de escritoras. Antes de realizar el cómputo de palabras de cada muestra se eliminaron del texto los fragmentos que pudieran aparecer en medio de las obras y que no estuvieran en prosa, como las canciones o los poemas. Todos estos pasos se siguieron considerando que el período seleccionado era suficientemente amplio y con un volumen de texto razonable como para que un estudio sobre la evolución y frecuencia de uso de estas colocaciones pudiera aportar datos concluyentes.

En la Tabla 3 se encuentra toda la información sobre este corpus, a saber, la especificación de los 3 subperíodos, el sexo y nombre de los autores, fecha de publicación de la obra, el título de la misma y la cantidad de palabras incluidas para cada muestra.

CORPUS 1800-1930 total **900.000** palabras					
Per.	**Sexo**	**Autor**	**Publi**	**Obra**	**Palabras**
1800 - 1850	H	Scott, Sir Walter	1814	*Waverley Vol.I*	82.362
		Thackeray, William M.	1840	*Catherine*	67.700
	M	Edgeworth, Maria	1812	*Absentee*	75.741
		Gaskell, Elizabeth	1849	*Mary Barton*	74.700
1850 - 1900	H	Collins, Wilkie	1860	*The Woman in White*	74.990
		Doyle, Sir A. Conan	1892	*The Adventures of Sherlok Holmes*	75.064
	M	Brontë, Charlotte	1857	*The Professor*	75.190
		Somerville, Edith y Martin Ross	1894	*The Real Charlotte*	75.000
1900 - 1930	H	Butler, Samuel	1903	*The Way of All Flesh*	74.270
		Lawrence, David Herbert	1928	*Lady Chaterley's Lover*	75.439
	M	Nesbit, Edith A. OEnone	1901 1902	*The Wouldbegoods* *Five Children and It*	44.400 30.591
		Woolf, Virginia	1927	*To the Lighthouse*	71.053

Tabla 3: El Corpus

3. METODOLOGÍA

Una vez compilado el corpus se procedió a la extracción, almacenamiento y procesado de datos, siguiendo una serie de pautas y utilizando la ayuda de algunos programas informáticos que detallaremos a continuación. También fue necesario consultar la información de los diccionarios *The BBI Dictionary of English Word Combinations* (*BBI*) (1997), *Oxford English Dictionary* (*OED*) (1994), *Oxford English Dictionary* (*OED2*) (2002) y *Oxford Collocations Dictionary* (*OCD*) (2002) para la identificación de los lexemas[13] y para observar la evolución que algunas colocaciones pudieran presentar.

3.1.- Herramientas de búsqueda

El tema de estudio, las colocaciones verbo-nombre, supuso un problema a la hora de utilizar alguna de las herramientas de análisis que se encuentran en el mercado. Las colocaciones, al ser construcciones multiléxicas formadas, en nuestro caso, por dos elementos (verbo y nombre), ofrecen la posibilidad de incluir otros elementos léxicos entre ambas (artículos, adjetivos u otro tipo de modificadores). Esta característica, resaltada ya por Greenbaum (1988: 114) "two items are collocates of each other if they belong to a single remembered set, no matter how far apart they may be in a strech of language", se convirtió en un problema a la hora de proceder a un análisis automatizado de los datos. Los programas de búsqueda suelen ofrecer unas herramientas de concordancias, pero éstas basan sus resultados en el concepto de colocación defendido por Jones y Sinclair (1974) y Sinclair (1991), entre otros, que fundamentaron sus investigaciones en la frecuencia de aparición, dándole a los resultados un carácter estadístico y formal. Para este sector de los lingüistas, el concepto de colocación parece reducirse a la frecuencia de coaparición lineal de los elementos léxicos en el discurso, cuya combinación se produzca en un espacio de texto inferior a cuatro palabras[14]. Sin embargo, como se puede comprobar en los ejemplos siguientes, la colocación no siempre se encuentra dentro de un espectro de cuatro palabras como presupone Sinclair (1991).

No se puede olvidar tampoco que las posibilidades de búsqueda que ofrecen los programas en este campo son limitadas. Algunos de los casos conflictivos para una búsqueda automática se recogen en los siguientes ejemplos. Concretamente en (1) y (2) se han incluido dos muestras de los abundantes casos en los que entre el verbo y el nombre pueden haberse introducido una serie de elementos que los distancien considerablemente en la oración.

> (1) So she wandered about, too restless to *take* her usual heavy morning's *sleep*, up and down the streets, greedily listening to every word of the passers-by, and loitering near each group of talkers, anxious to scrape together every morsel of information (Gaskell 1849)

> (2) Shortly after, Mr. Bradwardine remitted from Scotland a sum in reimbursement of expenses incurred in the King's High Court of Westminster, which, although not quite so formidable when reduced to the English denomination, *had*, in its original form of Scotch pounds, shillings, and pence, such a formidable *effect* upon the frame of Duncan Macwheeble, the laird's (Scott 1814)

También se encontraron ejemplos en los que se coordinan varias bases (nombres) con un mismo verbo:

> (3) He *has* that quiet *deference*, that *look* of pleased, attentive interest in listening to a woman, and that secret *gentleness* in his voice (Collins 1860)

Es obvio que estos casos sólo pueden ser localizados manualmente debido al material

léxico que separa al verbo de las variadas bases. Una circunstancia similar se produjo con las cláusulas de relativo, debido a la separación entre el verbo y la base, como en (4).

(4) Many other Trades' Unions, connected with different branches of business, supported with money, countenance, and encouragement of every kind, the *stand* which the Manchester power-loom weavers were *making* against their masters. (Gaskell 1849)

La labor del investigador se vio también complicada con la contracción de los verbos como en el siguiente ejemplo.

(5) As to her, I'll come to that directly; but first I*'ve* a word for you. I see you are a scoundrel; you*'ve* no business to be promenading about with another man's wife. I thought you had sounder sense than to get mixed up in foreign hodge-podge of this sort. (Brontë 1857)

Estos casos dificultan la búsqueda porque aumentan considerablemente los resultados de un rastreo al poder corresponder también a su función como verbo auxiliar.

Otros de los problemas surgieron del empleo por parte de los autores de formas arcaicas o dialectales de los verbos, o bien de las representaciones gráficas que simulen una determinada manera de hablar de un personaje, como en (6).

(6) Howe'er, I *han* no objection, if so be there's an opening, to speak up for what yo say; (Gaskell 1849)

Al mismo tiempo y dado, como hemos apuntado, que la búsqueda debía ser realizada a través de los verbos seleccionados para el estudio (*have, take, do* y *make*), los programas de exploración no podían localizar los casos en los que dichos verbos funcionen sólo como verbo de apoyo y descartar sus apariciones como verbos plenos o auxiliares. Por consiguiente, el análisis de los textos se realizó con la ayuda de un programa de búsqueda *Text Search* versión 2.4 (Alcott 2003) y la lectura minuciosa de los párrafos en los que se localizaron posibles casos. Este programa permite analizar los diferentes archivos del corpus a la vez, ofreciendo la posibilidad de ver, abrir y editar los archivos originales en sus aplicaciones asociadas.

Aunque el material seleccionado por el programa se vio engrosado por todas las apariciones de estos verbos en función de auxiliar o verbo pleno, la ventaja que ofrecía su utilización era resaltar la localización exacta del elemento buscado en todos los archivos del corpus. Una vez finalizada la búsqueda el programa presenta una pantalla en la que incluye los nombres de los archivos en los que se encuentran esas formas, además del número de veces y el contexto en el que aparecen. Así, nuestro siguiente paso consistió en analizar todas las extracciones que había realizado el programa y decidir cuáles eran colocaciones, según nuestro enfoque teórico, y cuáles eran verbos plenos o auxiliares. Como se puede apreciar, la ayuda del programa consistió en centrar nuestra atención en los casos en los que los escritores utilizaban alguna de las formas o funciones de estos cuatro verbos, pero en ningún caso podríamos haber hecho la extracción de los datos de manera automática.

Una vez localizados y determinados los momentos y contextos en los que estos verbos funcionaban como colocativos, los datos fueron archivados en un fichero específico para cada autor. Estos ficheros contenían parte de la información que se incluyó más tarde en la base da datos.

La primera tarea realizada con los ejemplos extraídos, antes de introducirlos en la base de datos fue diferenciar los lexemas ante los que nos encontrábamos. Dado que nuestro

vocabulario se situaba en el siglo XIX y principios del XX recurrimos al *Oxford English Diction*ary (1994, 2002) (*OED*) (Simpson, Weiner y Durkin 2004) para efectuar la distinción de los diferentes lexemas en ese momento concreto.

La identificación de los diferentes lexemas es un paso muy importante, aunque complicado. Debemos tener en cuenta que la información colocacional con la que contábamos, recogida en los diccionarios de colocaciones *BBI* y *OCD*, se refiere al inglés actual, mientras que el corpus abarca un período anterior. La falta de coincidencia entre la información de los diccionarios y los datos recogidos de los textos puede tener varias explicaciones. Por un lado, encontramos lexemas que no habían sido incluidos por los lexicógrafos en sus obras, bien porque están en desuso, bien por omisión. Por otro lado, obtuvimos lexemas que aparecen en los diccionarios con un posible verbo de apoyo o verbo colocativo, mientras que en los textos se utilizaba otro diferente. Estos últimos casos pueden proporcionarnos una base para el un futuro estudio de la evolución y cambios experimentados por estos verbos en combinación con ciertos lexemas desde el siglo XIX hasta hoy en día.

Por su parte estos diccionarios de colocaciones presentan la información de la siguiente manera. Dentro de una misma entrada, dedicada a un vocablo, encontramos diferentes lexemas que pueden formar colocación con uno u otro de los verbos. Como hemos comentado anteriormente, el sentido es el que determina la aparición de un verbo concreto. Como ejemplo recordaremos, en la Tabla 4, los datos del *OCD* sobre el nombre *chance* unido a los verbos centrales de este estudio.

CHANCE *noun*			
LEXEMA	**SENTIDO**	**COLOCATIVO**	**EJEMPLO**
1	possibility	*give, have*	The doctors gave him little chance of surviving the night.
2	opportunity	*have,* *take*	I finally had the chance to meet my hero. Take every chance that comes your way
3	risk	*take*	The guide book didn't mention the hotel, but we decided to take a chance.
4	luck/fortune	ninguno	

Tabla 4: Descripción de *chance* en el *Oxford Collocations Dictionary* (*OCD*)

En este diccionario se recogen 4 sentidos diferentes, aunque para el cuarto no se incluye ninguno de los verbos que analizábamos. Es importante resaltar que no todos los lexemas seleccionan los mismos verbos colocativos, como se puede observar en esta tabla. Por este motivo nos pareció imprescindible especificar los sentidos que podemos encontrar dentro de una palabra polisémica y detallar los verbos correspondientes para cada uno. Podríamos decir que el nombre *chance* puede formar colocación con los verbos *give, have* y *take*, pero esta información no sería totalmente cierta para ninguno de los cuatro lexemas. Si tomamos un ejemplo de alguno de los trabajos publicados sobre este tema, como el de Live (1973), podemos comprobar que, a pesar de que esta autora también resalta la importancia de diferenciar los sentidos, adjunta una lista de nombres en la que se reflejan todas las combinaciones verbales posibles de éstos con sus denominados *light verbs*. Si fijamos nuestra atención en *approach* comprobaremos que, según los datos recogidos en la lista, se combina con los verbos *have, take* y *make*. Sin embargo, la información de los diccionarios de colocaciones utilizados no coincide con esta aseveración. El *OCD* distingue tres lexemas de *approach*, que representaremos mediante subíndices:

1- Way of dealing with sb/sth: have**.** *Some teachers have a more formal approach to*

teaching. Take**.** *We need to take a more pragmatic approach.*

2- Act of coming nearer: make**.** *The aircraft had to make a steep approach to the landing strip.*

3- Discussion with sb in order to ask them for sth: make/have**.** *We'll have to make an approach to the managing director*.

De la lista de Live se infiere que cualquiera de estos tres lexemas puede combinarse con los tres verbos; no obstante, las opciones colocativas son, según este diccionario, *have* con $APPROACH_1$ y $APPROACH_3$; *make* con $APPROACH_2$ y $APPROACH_3$; *take* sólo con $APPROACH_1$.

Uno de los últimos puntos que se tuvo en cuenta para la recogida de datos hace referencia a los ejemplos en los que el lexema aparecía pronominalizado o había sido elidido, como en los ejemplos siguientes.

(7) Ernest's father and mother *have* no *interest*, and if they *had* […] they would not use it. (Butler 1903)

(8) She *had* no *contact* with them and intended to *have none* (Lawrence 1928)

(9) If you want *money* you must *make it* for yourselves as I did (Butler 1903)

Nuestra postura ante estos casos fue la siguiente. Puesto que nuestro estudio giraba en torno al uso y frecuencia de las colocaciones, en estas muestras se recogieron dos casos para los lexemas correspondientes a INTEREST1.7A y CONTACT1.1A, aunque el lexema apareciera escrito sólo una vez. En el último ejemplo se contabilizó un caso de *make money*, ya que el pronombre *it,* en esta muestra, substituye al lexema MONEY1.1A. La identificación numérica de los lexemas se explica detalladamente a continuación.

3.2. Clasificación y almacenamiento de datos

Para almacenar y posteriormente analizar toda la información se utilizó la base de datos *Access* 2000. Mediante este programa se creó una ficha para cada lexema en la que se incluyeron los siguientes diecisiete campos : "lexema", "significado", "verbo apoyo", "autor", "período", "BBI", "Oxford", "OED", "1ª aparición", "Artículo", "Justificación", "plural", "modificador", "negación", "determinante", "comentario" y "ejemplo". El contenido de cada campo se detallará a continuación, partiendo de una de las fichas de la base de datos reflejada en la Ilustración 1.

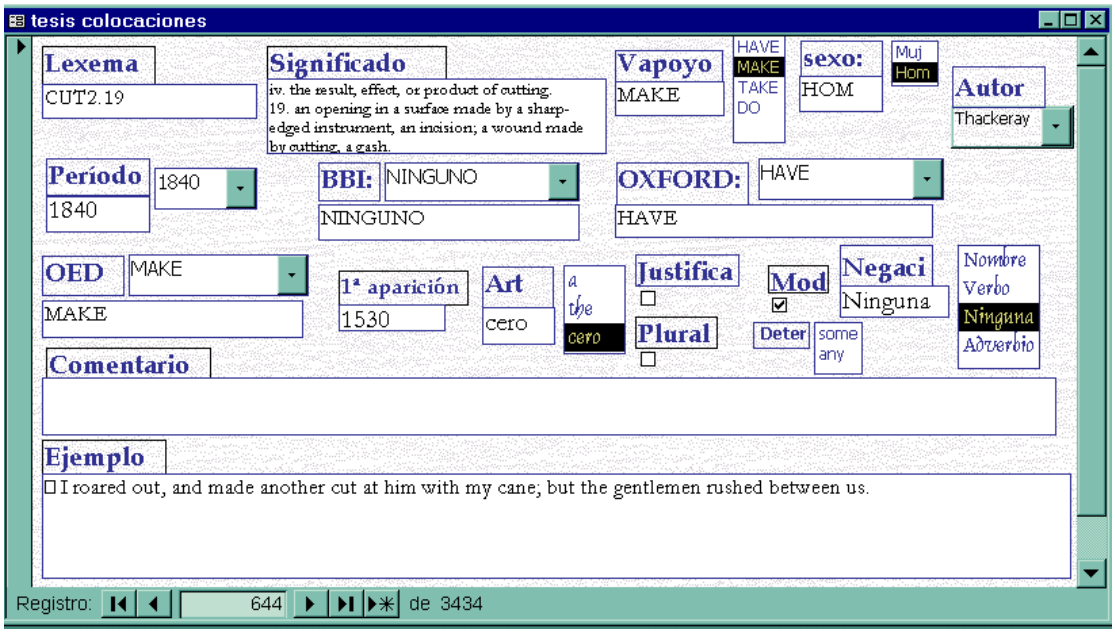

Ilustración 1: Ficha de la base de datos

La información que recoge el campo "Lexema" se extrajo de los resultados ofrecidos por el programa de búsqueda. Como ya hemos puntualizado los parámetros de localización de las colocaciones son los verbos *have, take, make* y *do* en todas sus formas y variantes. Así, en el campo "lexema" se archivó el nombre que acompaña a uno de estos verbos formando una colocación.

Asimismo, ya hemos mencionado que para etiquetar los diferentes lexemas entre los ejemplos extraídos del corpus se utilizó el *OED* como fuente principal de información. Una vez descifrado el sentido de un determinado nombre en su contexto, se buscó en el diccionario la definición que correspondía a este lexema concreto. Para ello se revisaron todos los sentidos que incluía la entrada del *OED* para ese nombre. Por ejemplo, para etiquetar la colocación que aparece en la ficha de *cut* (*I roared out, and made another cut at him with my cane; but the gentlemen rushed between us)* (Ilustración 1) se siguió el procedimiento detallado a continuación.

Como se aprecia en la Ilustración 1, el lexema va seguido de unos números y letras. Esta nomenclatura utilizada para su etiquetado hace referencia a las diferentes columnas del *OED* empezando de izquierda a derecha (Ilustración 2). El número que sigue al lexema identifica el lugar que el lexema ocupa en la columna de los resultados de la izquierda (Ilustración 2: *word list*). Si sólo aparecía una vez como nombre, el número que seguirá al lexema será el 1, pero si lo hacía más veces, el número dependerá de cuál de ellos sea el que incluyera el sentido que estábamos buscando.

El siguiente identificador se tomó de la pantalla de la derecha. Normalmente cada sentido está precedido de un número e incluso de una letra. Éstos son los siguientes identificadores del lexema. Como podemos comprobar en la imagen siguiente (Ilustración 2) la identificación del lexema *cut* es CUT2.19.

Recibido el 26-11-2006

RæL 5 (2006): 87-106
ISSN 1885-9089

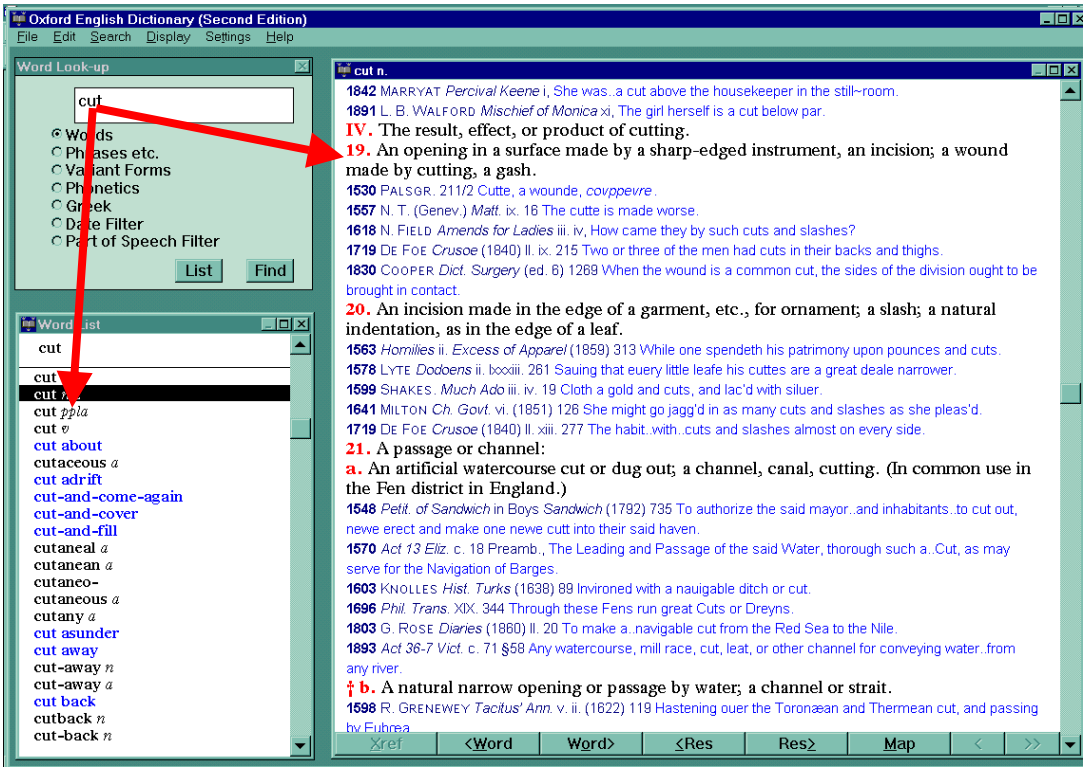

Ilustración 2: Resultado de la búsqueda de *CUT* en el *OED*

En la misma pantalla de resultados se localizaba la información que se consignaría en la ficha de la base de datos bajo los campos "Significado", "OED" y "1ª aparición". Dentro del primer campo se copió la definición 19: *an opening in a surface made by a sharp-edged instrument, an incision; a wound made by cutting.* En el campo "OED" se consignaron los verbos colocativos que incluyeran los ejemplos del diccionario (en este caso tendríamos HAVE) y en el tercero, la fecha de la primera aparición de ese lexema (para nuestro ejemplo es 1530).

Dentro del campo "Vapoyo" se introdujo el verbo colocativo que aparecía en nuestro corpus, para este ejemplo MAKE. Seguidamente se señalaba el "sexo" del autor con las variables HOM para escritores y MUJ para escritoras. El campo "autor" mostraba el apellido del autor del texto (Thackeray). Dentro del campo "Período" se incluía la fecha de publicación del texto (1840). A continuación, en los campos "BBI" y "OXFORD", se señalaban los verbos colocativos que recogían los dos diccionarios actuales de colocaciones con los que estamos trabajando. Las posibilidades que se consideraron en estos campos son las siguientes:

- ♦NO. Para los casos en los que el nombre no estuviera recogido en el diccionario.
- ♦NO L. Siempre que sí hubiera una entrada para el nombre pero no se incluyera el lexema concreto que aparece en nuestro corpus.
- ♦NINGUNO. Cuando se localizara el lexema pero la información no reflejara explícitamente nuestros verbos colocativos.
- ♦Uno de los cuatro verbos o las combinaciones posibles que incluyan los diccionarios (*have*, *make*, *take* y *do*).

En nuestro ejemplo tenemos "BBI" como NINGUNO (porque este diccionario no incluye ninguno de los verbos que estamos analizando) y en "OXFORD" los colocativos HAVE y MAKE (porque son los dos verbos de apoyo que figuran en el diccionario de

colocaciones *OCD*).

La última parte de la ficha recoge información de la realización concreta de la colocación en el corpus referida a las posibilidades de modificación, a la negación, etc. Por un lado tenemos "ART", donde se muestran tres variables (THE, A, CERO). En este ejemplo marcaríamos la casilla de "cero", pues éste es el artículo utilizado. A continuación se muestra un campo en el que teníamos que verificar el uso del artículo determinado *the* o los demostrativos *this/that*[15], para las muestras que lo contuvieran. Se marcaría la casilla de verificación con una cruz para los casos en los que el artículo fuera exigido por el propio discurso, como en los usos anafóricos,

(10) 'To make *the experiment* whether he may not be brought to communicate to me some circumstances which may hereafter be useful to alleviate, if not to exculpate, his conduct.' (Scott 1814)

(11) When *the engagement* was made for me by my father, with my own consent? (Collins 1860)

o los casos en los que existía postmodificación (Quirk *et al.* 1985: 286), como

(12) Let me have *the pleasure of* telling him that he's a coward and a liar; (Thackeray 1840)

(13) Many other Trades' Unions, connected with different branches of business, supported with money, countenance, and encouragement of every kind, *the stand which* the Manchester power-loom weavers were making against their masters. (Gaskell 1849)

o cuando lo exigía la premodificación, como en

(14) Lord Colambre had *the greatest dread* of marrying any woman whose mother had conducted herself ill. (Edgeworth 1812)

(15) We were engaged after *the first walk* that we took (Doyle 1892).

El siguiente campo, "Plural", recoge información sobre la flexión del lexema. Aquí marcamos la casilla sólo si el lexema está en plural. Este campo no se marcaría en las muestras en las que el plural forma un nuevo lexema. Estaríamos en estos casos ante lexemas con forma de plural, pero sin singular, estando incluidos en los diccionarios de colocaciones ya en plural. En la Tabla 5 se detallan los ejemplos del corpus en los que encontramos estos lexemas, incluyendo su definición en el *OED* y el verbo con el que forman colocación en el texto.

LEXEMA	DESCRIPCIÓN DEL (*OED*)	VERBO	#
HONOURS1.5C	5c. pl. civilities or courtesies rendered, as at an entertainment:	*Do*	2
GROUNDS1.5C	5c. reason, motive.	*Have*	1
LIBERTIES BBI	undue familiarity	*Take*	1
MANNERS1.4C	4. collect. pl. c. the modes of life, customary rules of behaviour, conditions of society, prevailing in a people.	*Have*	3
MEASURES1.21A	21. a plan or course of action intended to attain some object.	*Take*	5
PAINS1.6A	6. a. pl. trouble taken in accomplishing or attempting something; labour, toil, exertions, or efforts, accompanied with care and attention, to secure a good or satisfactory result. most freq. in phr. to take pains,	*Take*	15

ORDERS1.6B	6b. the rank, status, or position of a clergyman or ordained minister of the church. now always pl.,	*Take*	1
SCRUPLES2.1A	1. a thought or circumstance that troubles the mind or conscience; a doubt, uncertainty or hesitation in regard to right and wrong, duty, propriety, etc.;	*Have*	1
PREPARATIONS1.1B	1b. usually in pl.: things done by way of making ready for something; preparatory actions, proceedings, or measures.	*Make*	4

Tabla 5: Plurales como lexemas

Los tres últimos campos del registro se concentran en la modificación del nombre. Por un lado tenemos el campo "Mod", que marcaremos si el nombre está modificado por algún adjetivo. Debajo encontramos la casilla "Deter", que incluye "some" y "any" como variables.

Seguidamente tenemos el campo "Negaci", donde se especifica el tipo de negación que se utiliza en el ejemplo. Las posibles opciones son: nombre, verbo, ninguna, adverbio. En la Ilustración 1 vemos que, al no estar negada la oración del ejemplo, marcaremos la casilla "ninguna". Por último, creamos un campo para los posibles comentarios que surgieran del análisis y otro campo para incluir el ejemplo localizado en nuestro corpus.

Este largo proceso fue seguido con todas y cada una de las colocaciones extraídas del corpus. Así, las 900.000 palabras de las muestras dieron como resultado final 3.434 registros como el de la Ilustración 1, que acabamos de describir. Estos registros representaban todos los casos o ejemplos de colocaciones localizadas en el corpus, pudiéndose aislar y etiquetar de entre ellos 957 lexemas diferentes. Recordaremos que no se utilizó la frecuencia de aparición como motivo de exclusión de ninguno de los casos localizados porque en nuestro concepto de colocación, enmarcado dentro de la Teoría Sentido-Texto (*TST*), la frecuencia no tiene ningún valor para catalogar el estatus de una combinación de lexemas.

4. CONCLUSIÓN

Como se refleja en este artículo, a pesar del actual auge de los corpus computerizados, no siempre es posible el acceso a un corpus específico que sea capaz de cumplir las expectativas de una investigación determinada. Sin embargo, éste no es siempre motivo suficiente para cambiar la línea de nuestro trabajo. Por este motivo, aunque la compilación de un corpus para una investigación conlleva un trabajo añadido, siempre que se sigan unos criterios estables y razonados, es posible extraer datos que nos permitan llegar a conclusiones interesantes.

Recordaremos que nuestro interés se centraba en la evolución y comportamiento de las colocaciones (V+N) en un determinado momento de la historia de la lengua inglesa y que nuestras pesquisas se dirigían a comprobar si ciertos factores extralingüísticos como el sexo, la edad o la procedencia geográfica podían influir de alguna manera en el uso de este tipo de colocaciones por parte de la comunidad literaria de la época en cuestión. Puesto que el análisis minucioso de los datos extraídos de este corpus no forma parte directa del tema de este artículo, no mencionaremos la tendencia que se observaba en los datos.

La presentación y análisis de los datos extraídos de este corpus tanto desde el punto de vista de los factores extralingüísticos como de la evolución lingüística observada serán el tema de futuros trabajos. Sin embargo, la fiabilidad del corpus cuya creación hemos descrito, ha sido ya probada (Lareo 2006).

Recibido el 26-11-2006

RæL 5 (2006): 87-106
ISSN 1885-9089

[1]. Este trabajo ha sido realizado en parte gracias a las ayudas recibidas de la Xunta de Galicia y de la Universidade da Coruña, a través del Vicerrectorado de Investigación y de Tercer Ciclo. Asimismo se agradece también a la Xunta de Galicia, concretamente a su Dirección Xeral de Investigación y Desenvolvemento, el apoyo mostrado al proyecto PGIDIT03PXIB10402PR (dirigido por Isabel Moskowich).

[2]. A finales de los 60 A. Zholkovsky, I. Mel'□uk y J. Apresjan forman el *Círculo Semántico de Moscú*. Actualmente el profesor I. Mel'□uk trabaja en la Universidad de Montreal desarrollando diferentes líneas de investigación ([Documento de Internet disponible en http://www.olst.umontreal.ca/ melcuk/]).

[3]. Los corpus localizados como *A Corpus of Late Modern English Prose*, compilado por David Denison, se limita sólo a cartas informales escritas entre 1861 y 1919 ([Documento de Internet disponible en http://www.art.man.ac.uk/english/staff/dd/lmodeprs.htm] 2002). Otros proyectos de más reciente aparición ya mencionados, pero no accesibles en el momento en el que se realizó este estudio, son *The Corpus of Late Modern English Texts* (*CLMET*) y el *Corpus of Nineteenth Century English* (*CONCE*). El primero, dirigido por Hendrik de Smet ([Documento de Internet disponible en http://perswww.kuleuven.ac.be /~u0044428/] 2004) abarcará el período de 1710-1920 y recogerá una amplia variedad de textos y géneros. Por su parte el *CONCE*, dirigido por Merja Kytö y Juhani Rudanko, incluye diferentes géneros y autores, y estará también en un futuro a disposición del público (véase Kytö, M., J. Rudanko y E. Smitterberg 2000).

[4]. Thomas Carlyle, Charles Darwin, Thomas DeQuincey, Charles Lamb, Stuart Mill, Mary Shelley y Oscar Wilde.

[5]. A la hora de delimitar los subperíodos se tuvieron en cuenta trabajos anteriores como Nevalainen y Raumolin-Brunberg (1990), sobre el *Corpus of Early Modern Standard English*; Kytö (1991), sobre el *Helsinki Corpus of English Texts. Diachronic and Dialectal;* Taavitsainen y Pahta (1997), sobre el *Corpus of Early English Medical Writing;* Kytö *et al.* (2000), sobre el *Corpus of 19th-century English* y De Smet (2004), sobre *the Corpus of Late Modern English Texts.* En el primero se encontraron subperíodos de 70 años, en el segundo de entre 70 y 100 años, en el tercero entre 75 y 100, en el cuarto de 30 años y en el último de 70 años.

[6]. En el estudio de Claridge (2000: 187) sobre el uso de colocaciones V+N en el Inglés Moderno Temprano en el *Lampeter Corpus*, se concluye que no se puede hablar de un perfil de texto concreto más o menos proclive a este fenómeno. La autora llega a esta conclusión al observar los porcentajes tan similares alcanzados por los diferentes tipos de textos utilizados.

[7]. Esta base de datos ofrece más de 350.000 obras americanas e inglesas que abarcan desde el siglo VIII hasta la primera mitad del siglo XX. Nuestros textos forman parte de la sección que comprende *Nineteenth-century Fiction.* Las obras que contiene esta sección están comprendidas entre 1781 y 1903.

[8]. El *Project Gutenberg* pone al servicio público una serie de textos electrónicos que incluyen obras de todos los géneros y épocas, así como de la mayoría de autores y países. Las obras son distribuidas a través de la asociación *Gutenberg* en la Universidad de Carnegie-Mellon por el catedrático Michael S. Hart.

[9]. Esta biblioteca ha sido creada por *Logos*, una compañía internacional de traducción.

[10]. Contiene más de 3.000 textos de los siglos XVIII y XIX, puesto a disposición en la Red por Stephane Theroux. Los textos están publicados "as they were originally written".

[11]. Este pseudónimo fue utilizado por la prima de Somerville, Violet Florence Martin. Como se puede apreciar en la siguiente cita, la obra seleccionada es una de las más importantes novelas del siglo XIX que describe la sociedad aristocrática irlandesa: "*The Real Charlotte* is a vivid account of Ireland's aristocratic social life. It tells of the importance of marriage for social status, and the antics and schemes devised by various women to achieve this patriarchal, materialistic society's goal of an adequate spouse. The tale turns to revenge when the main character, Charlotte Mullen, is spited by the man she desires because of her double curse; no money, and excessive homeliness" (English Department of the United States Naval Academy [Documento de Internet disponible en http://www. usna.edu/EnglishDept/ilv/sandr.htm] (2005)).

[12]. Como se puede comprobar en la Tabla 3, esta tarea sólo fue necesaria en el caso de los cuentos de Edith Nesbit.

[13]. Utilizaremos el término *lexema* en el sentido de Mel'čuk (1998: 168): "*a lexeme is a word taken in one, well-specified sense".*

[14]. Sinclair introduce en 1966 los términos de *node* (núcleo), *collocates* (colocados) y *span* (espacio de texto). El autor define la colocación (1991: 170) como "the occurrence of two or more words within a short space of each other in a text".

[15]. Los demostrativos también pueden ser exigidos por el propio contexto, como apunta Akimoto (1999: 213) cuando habla de su función de relacionar los contenidos previos con los que le siguen en el discurso.

RERERENCIAS

Fuentes Primarias

Brontë, Ch. 1857. *The Professor*. London: Smith, Elder & Co. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Butler, S. 1903. *The Way of all Flesh*. London: Grant Richards. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Collins, W. 1860. *The Woman in White*. London: Sampsonlow, son & Co. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Doyle, Sir A. C. 1892. *The Adventures of Sherlock Holmes*. (25-10-2003). [Documento de Internet disponible en http://promo.net/pg].
Edgeworth, M. 1812. *The Absentee*. London: Johnson & Co. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Gaskell. E. 1849. *Mary Barton: A Tale of Manchester*. 3ª edición. London: Chapman and Hall. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Lawrence, D. H. 1928. *Lady Chaterley's Lover*. (5-11-2003). [Documento de Internet disponible en http:// wordtheque.com].
Nesbit, E. A. OE. 1901. *The Wouldbegoods*. (29-10-2003). [Documento de Internet disponible en http://www. classicreader.com].
Nesbit, E. A. OE. 1902. *Five Children and It*. (29-10-2003). [Documento de Internet disponible en http://www. classic.reader.com].
Scott, Sir W. 1814. *Waverley*. Edinburg: Robert Cadell. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Somerville, E. OE. y M. Ross. 1894. *The Real Charlotte*. London, Ward & Downey Ltd. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http:// collections. chadwyck. co.uk].
Thackeray, W. M. 1840. *Catherine*. London: James Fraser. Nineteenth-century Fiction. (20-10-2003). [Documento de Internet disponible en http://collections. chadwyck. co.uk].
Woolf, V. 1927. *To the Lighthouse*. (25-10-2003). [Documento de Internet disponible en http://promo. net/pg].

Fuentes Secundarias

Aijmer, K. y B. Altenberg, eds. 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.
Akimoto, M. 1999. "Collocations and Idioms in Late Modern English". *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Eds. L. Brinton y M. Akimoto. Amsterdam/ Philadelphia: John Benjamins Publishing. 207-238.
Akimoto, M. y L. Brinton. 1999. "The Origin of the Composite Predicates in Old English". *Collocational and Idiomatic Aspects of Composite Predicates in the History of English*. Eds. L. Brinton y M. Akimoto. Amsterdam/ Philadelphia: John Benjamins Publishing. 21-58.
Alcott, G. 2003. *Super Text Search*. Ver. 2.4. (11-4-2003). [Documento de Internet disponible en http:// www.galcott.com/].
Algeo, J. 1995. "Having a Look at the Expanded Predicate". *The Verb in Contemporary English: Theory and Description*. Eds. B. Aarts, y Ch. F. Meyer. Cambridge: Cambridge University Press. 203-217.
Bailey, R. W. 1999. *Nineteenth-century English*. Ann Arbor: The University of Michigan Press.
Beal, J. C. 2004. *English in Modern Times: 1700-1945*. London: Arnold.

Recibido el 26-11-2006

RæL 5 (2006): 87-106
ISSN 1885-9089

Benson, M., E. Benson y R. Ilson. 1997. *The BBI Dictionary of English Word Combinations* (*BBI*). Amsterdam/Philadelphia: John Benjamins.
Bing, J. M. y V. L. Bergvall. 1996. "The Question of Questions: Beyond Binary Thinking". *Rethinking Language and Gender Research: Theory and Practice.* Eds. V. L. Bergvall, J. M. Bing y A. F. Freed. London/New York: Longman. 1-30.
Bratt Paulston, Ch. y G. R. Tucker, eds. 2003. *Sociolinguistics. The Essential Readings*. Oxford: Blackwell Publishers.
Cameron, D., ed. 1990. *Feminist Critique of Language. A Reader.* London/New York: Routledge.
Chambers, J.K. 2003. *Sociolinguistic Theory: Linguistic Variation and its Social Significance*. Malden /Oxford/ Victoria: Blackwell Publishing.
Claridge, C. 2000. *Multiword Verbs in Early Modern English*: *A Corpus-based Study*. Amsterdam/Atlanta: Rodopi.
Coates, J. 1986. *Women, Men and Language*. London: Longman.
Coates, J. 1990. "Language and Sex in Connected Speech. Introduction". *Women in Their Speech Communities. New Perspectives on Language and Sex.* Eds. J. Coates y D. Cameron. London/New York: Longman. 63-73.
De Smet, H. *The Corpus of Late Modern English Texts (CLMET)*. (7-6-2004). [Documento de Internet disponible en http://perswww.kuleuven.ac.be/~u0044428/].
De Smet, H. 2005. "A Corpus of Late Modern English Texts". *ICAME Journal* 29: 69-82.
Eckert, P. y S. McConnell-Ginet. 2003. *Language and Gender*. Cambridge: Cambridge University Press.
Görlach, M. 1999. *English in Nineteenth Century England. An Introduction*. Cambridge: Cambridge University Press.
Görlach, M. 2004. *Trends in Linguistics: Text Types and the History of English.* Berlin/New York: Mouton de Gruyter.
Greenbaum, S. 1988. *Good English and the Grammarian*. New York: Longman.
Guzmán-González, T. 2005. "Out of the Past: A Walk with Labels and Concepts, Raiders of the Lost Evidence, and a Vindication of the Role of Writing". *International Journal of English Studies* 5: 13-31.
Hiltunen, R. 1999. "Verbal Phrases and Phrasal Verbs in Early Modern English". *Collocational and Idiomatic Aspects of Composite Predicates in the History of English.* Eds. L. Brinton y M. Akimoto. Amsterdam/ Philadelphia: John Benjamins Publishing. 133-166.
Hoffmann, S. y H. M. Lehmann. 2000. "Collocational Evidence from the *BNC*". *Corpora Galore: Analyses and Techiques in Describing English*. Ed. J. M. Kirk. Amsterdam/Atlanta: Rodopi. 17-32.
Holmes, J. 1999. "Women´s Talk: The Question of Sociolinguistic Universals". *Language and Gender: A Reader.* Ed. J. Coates. Oxford/Malden: Blackwell Publishers. 461-483.
Iglesias Rábade, L. 2001. "Composite Predicates in Middle English with the Verbs *Nimen* and *Taken*". *Studia Neophilologica* 73: 143-163.
Johansson, S. 1991. "Times change and so do corpora". *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Eds. K. Aijmer y B. Altenberg. London: Longman. 305-314.
Jones, S. y Sinclair, J. M. 1974. "English Lexical Collocations. A Study in Computational Linguistics". *Cahiers de Lexicologie* 24: 15-61.
Kytö, M. 1991. *Manual to the Diachronic Part of the Helsinki Corpus of English Texts. Coding Conventions and Lists of Source Texts*. Helsinki: University of Helsinki.

Kytö, M. 1999. “Collocational and Idiomatic Aspects of Verbs in Early Modern English”. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English.* Eds. L. Brinton y M. Akimoto. Amsterdam/ Philadelphia: John Benjamins Publishing. 167-206.

Kytö, M., J. Rudanko y E. Smitterberg. 2000. “Building a bridge between the present and the past: A Corpus of 19th-century English”. *ICAME Journal* 24: 85-98.

Lareo, I. 2006. *Las colocaciones verbo-nombre en Inglés Moderno Tardío. Algunos Aspectos de su Comportamiento y Evolución.* Tesis doctoral presentada en la Universidad de A Coruña.

Lareo, I. En prensa. “Make-Collocations in Nineteenth-Century Scientific English”. *Studia neophilologica.*

Lee, D. 2001. "Genres, Registers, Text Types, Domains, and Styles: Clarifying the Concepts and Navigating a Path through the *BNC* Jungle". *Language Learning & Technology* 5: 37-72.

Live, A. H. 1973. “The *TAKE-HAVE* Phrasal in English”. *Linguistics* 95: 31-50

Matsumoto, M. 1999. “Composite Predicates in Middle English”. *Collocational and Idiomatic Aspects of Composite Predicates in the History of English.* Eds. L. Brinton y M. Akimoto. Amsterdam/ Philadelphia: John Benjamins Publishing. 59-96.

Mel’c‡uk, I. 1995. “Phrasemes in Languages and Phraseology in Linguistics”. *Idioms: Structural, and Psychological Perspectives*. Eds. M. E. Everaert, A. Schenk van der Linden y R. Schreuder. New Jersey: Lawrence Erlbamm Associates, Inc. 167-232.

Mel’c‡uk, I. 1998. “Collocations and Lexical Functions”. *Phraseology: Theory, Analysis and Applications*. Ed. A. P. Cowie. Oxford/New York: Oxford University Press. 23-53.

Mele, M. y Mª A. Martín. 2001. "Formación y Desarrollo del Inglés Moderno". *Lingüística Histórica Inglesa*. Eds. I. de la Cruz y J. Martín. Barcelona: Ariel. 573-623.

Milroy, L., L. Milroy y S. Hartley. 1994. “Local and Supra-local Change in British English: The Case of Glottalisation”. *English World-Wide* 15: 1-33.

Milroy, L. y M. Gordon. 2004. *Sociolinguistics: Method and Interpretation.* Malden/Oxford/Victoria: Blackwell.

Montoya Reyes, A. 2003. *La Difusión de la variedad escrita del inglés estándar en la correspondencia y documentos de la familia Paston.* Tesis Doctoral. A Coruña: Universidade da Coruña.

Moralejo, T. 2003. *Composite Predicates in Middle English*. München: LINCOM.

Moskowich, I y B. Crespo. 2007. “Presenting the Coruña Corpus: A Collection of Samples for the Historical Study of English Scientific Writing”. *'Of Varying Language and Opposing Creed': New Insights into Late Modern English*. Eds. J. Pérez Guerra, J. L. Bueno Alonso, D. González-Álvarez y E. Rama Martínez. Bern: Peter Lang (Linguistic Insights Series).

Nevalainen, T. y H. Raumolin-Brunberg. 1990. “A Corpus of Early Modern Standard English”. *Neuphilologische Mitteilungen* 90: 67-110.

Nurmi, A. 1999. *A Social History of Periphrastic Do.* Helsinki: Ramoulin-Brunberg. 111-139.

Olsen, M. 2005. "*Écriture féminine:* Searching for an indefinable Practice?". *Literary and Linguistic Computing* 20 Supplementary Issue: 147-164.

Osselton, N. E. 1984. “Informal Spelling Systems in Early Modern English: 1500-1800". *English Historical Linguistics: Studies in Development*. Eds. N. F. Blake y Ch. Jones. Sheffield: University of Sheffield. 123-137.

Oxford English Dictionary (*OED*). 1994. Ver. 1.13. Oxford: Oxford University Press.

Oxford English Dictionary (*OED2*). 2002. Ver. 3.00. Oxford: Oxford University Press.
Oxford Collocations Dictionary (*OCD*). 2002. Oxford: Oxford University Press.
Quirk, R., S. Greenbaum, G. Leech, J. Svartvik. 1985. *A Comprehensive Grammar of the English Language*. New York: Longman.
Romaine, S. 2003. "Variation in Language Gender". *The handbook of language and gender*. Eds. J. Holmes y M. Meyerhoff. Malden: Blackwell. 91-118.
Simpson, J., E. Weiner y P. Durkin. 2004. "The Oxford English Dictionary Today". *Transactions of the Philological Society* 102: 335-381.
Sinclair, J. McH. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
Sinclair, J. 2004. "Corpus and Text: Basic Principles" [Documento de Internet disponible en http://ahds.ac.uk/creating/guides/linguistic-corpora/index.htm].
Svartik J., ed. 1992. *Directions in Corpus Linguistics*. Berlin: Mouton de Gruyter. 335-73.
Taavitsainen, I. y P. Pahta. 1997. "Corpus of Early English Medical Writing 1375-1750". *ICAME Journal* 21: 71-78.
Tanabe, H. 1999. "Composite Predicates and Phrasal Verbs in *The Paston Letters*". *Collocational and Idiomatic Aspects of Composite Predicates in the History of English.* Eds. L. Brinton y M. Akimoto. Amsterdam/ Philadelphia: John Benjamins Publishing. 97-132.
Tannen, D. 2003. "The Relativity of Linguistic Strategies: Rethinking Power and Solidarity in Gender Dominance". En Ch. Bratt Paulston y G. R. Tuckers, eds. *Sociolinguistics. The Essential Readings*. Oxford: Blackwell Publishers. 208-229.
Warner, A. 1961. *A Short Guide to English Style*. London: Oxford University Press.
Woolf, V. 1929. "Women and Fiction". *Feminist Critique of Language. A Reader.* Ed. D. Cameron. London/New York: Routledge. 33-40.
Woolf, V. 1967. "Dorothy Osborne Letters". *Collected Essays* Vol. 3. Ed. L. Woolf. London: Chatto & Windus.

Recibido el 26-11-2006

RæL 5 (2006): 87-106
ISSN 1885-9089