Cum corde et in nova grammatica

DEPARTAMENTO DE LINGUA ESPAÑOLA

Cum corde et in nova grammatica

Estudios ofrecidos a Guillermo Rojo

EDICIÓN A CARGO DE
Tomás Jiménez Juliá
Belén López Meirama
Victoria Vázquez Rozas
Alexandre Veiga

2012
UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

Cum corde et in nova grammatica : estudios ofrecidos a Guillermo Rojo / edición a cargo de Tomás Jiménez Juliá, Belén López Meirama, Victoria Vázquez Rozas, Alexandre Veiga. - Santiago de Compostela : Universidade de Santiago de Compostela, Servizo de Publicacións e Intercambio Científico, 2012
927 p. ; 24 cm. – (Homenaxes (Universidade de Santiago de Compostela))
D.L. C 1286-2012. – ISBN: 978-84-9887-914-8
1. Filoloxía 2.Español (Lingua) I. Rojo Sánchez, Guillermo, 1947- II. Jiménez Juliá, Tomás, ed. lit. III. López Meirama, Belén, ed. lit. IV. Vázquez Rozas, Victoria, 1960- , ed. lit. V. Veiga, Alexandre, 1958- , ed. lit. V. Universidade de Santiago de Compostela. Servizo de Publicacións e Intercambio Científico, ed.

80
806.0

Edita
Servizo de Publicacións
e Intercambio científico
Campus Vida
15782 Santiago de Compostela
usc.es/publicacions

Imprime
Imprenta Universitaria
Campus Vida
15782 Santiago de Compostela

Dep. Legal: C 1286-2012
ISBN 978-84-9887-914-8

Índice

Explorando la frecuencia léxica para el *Diccionario de colocaciones del español*

Margarita ALONSO RAMOS
Universidade da Coruña

1. INTRODUCCIÓN

En este trabajo[1] nos proponemos explorar la funcionalidad del concepto de *frecuencia léxica* en un diccionario de colocaciones, más en particular en el *Diccionario de colocaciones del español* (*DiCE*, Alonso Ramos 2004, 2010, Vincze *et al.* 2011). Como es sabido, el fenómeno léxico de las colocaciones no recibe una única definición. Simplificando, hay dos grandes corrientes teóricas: una, que viene desde Firth (1957), se apoya especialmente en la estadística para considerar una combinación dada como colocación[2], y otra, promovida especialmente por el enfoque lexicográfico de Hausmann y Mel'čuk[3], se centra en la coocurrencia léxica restringida entre dos palabras por la que una de ellas exige la presencia de otra para expresar un sentido dado. Obviamente, la frecuencia es el factor clave para determinar qué es una colocación dentro de la primera corriente, pero no es así dentro de la corriente lexicográfica, en donde los criterios son lingüísticos. Ahora bien, frente a la "objetividad" de los datos estadísticos, los criterios lingüísticos no son los mismos para todos los autores. El debate entre qué es o no colocación gira en torno a concepciones distintas de qué es una combinación libre o restringida (Bosque 2001, Muñiz 2004). El *DiCE* se enmarca dentro de la corriente lexicográfica en donde la frecuencia no desempeña ningún papel en la determinación de qué es o no colocación. Así, desde nuestro enfoque, tanto *gran cariño*, que es frecuente y trivial, como *cariño entrañable*, no muy frecuente y más idiosincrásica, son consideradas colocaciones, como veremos abajo.

Aquí no queremos contribuir más al debate de qué es o no colocación. Lo que pretendemos es valorar la utilidad de la frecuencia en la selección del material léxico para un diccionario. El *DiCE* ha sido construido inductivamente, a partir principalmente del *Corpus de referencia del español actual* (*CREA*), pero en ese proceso no se ha valorado suficientemente hasta qué punto una determinada colocación es muestra de la "voluntad de estilo" de un autor o es representativa de la lengua y usada normalmente por cualquier hablante nativo. Así, por ejemplo, actualmente ofrecemos 33 adjetivos que sirven para intensificar el nombre *odio*. Ahí entran desde adjetivos como *acérrimo*, que parecen ejercer ese papel de "palabra justa", hasta palabras como *fuerte* o *profundo*, que funcionan casi como el intensificador por defecto, y pa-

[1] Quisiera agradecer a Leo Wanner y a Orsolya Vincze las discusiones sobre frecuencias léxicas y más especialmente a Nancy Vázquez y a Aquilino Sánchez por haber aceptado leer la primera versión de este texto y haber aportado sus comentarios. Este trabajo se enmarca en el proyecto de investigación financiado por el MINECO (FFI2011-30219-C02-01).

[2] Desde la lingüística británica, se considera colocación cualquier combinación de unidades léxicas que tienen una alta probabilidad de coexistencia en el corpus: "collocation is the occurrence of two or more words within a short space of each other in a context" (Sinclair 1991: 170).

[3] Entre otros, Mel'čuk (1998, 2006) y Hausmann (1979, 1989).

sando por *radical* o *satánico*, que quizás puedan ser más muestras del estilo de un determinado autor. Nos gustaría poder ordenar, de algún otro modo además del alfabético, estos 33 adjetivos y poder basarnos en algún criterio para su selección. Aquí es donde entra el concepto de frecuencia.

En lo que sigue, nos disponemos a valorar cómo se ha utilizado la frecuencia para seleccionar el material léxico que se debe incluir tanto en los diccionarios como en el material didáctico destinado a los aprendices de lenguas. Antes presentaremos someramente la noción de colocación tal y como la concebimos desde el *DiCE*, así como una breve exposición de cómo se codifica la información en este diccionario. A continuación, nos detendremos en reflexionar sobre la equiparación entre palabras frecuentes y palabras útiles. Puesto que la nomenclatura del *DiCE* está constituida en su gran parte por nombres de sentimiento, examinaremos en diferentes listados de frecuencias y diccionarios de colocaciones cómo se han seleccionado estos nombres y algunos de sus colocativos. Finalmente, exploraremos diferentes vías para poder asignar información de frecuencias a los datos incluidos en el *DiCE*.

2. BREVE PRESENTACIÓN DEL *DiCE*

El *DiCE* ha sido concebido como una base de datos, que se puede consultar en la web. Se centra en recoger para una unidad léxica dada todas las elecciones sintagmáticas que dependen de elecciones léxicas hechas previamente. Así, por ejemplo, bajo la entrada de *alegría*, incluimos las distintas maneras de expresar 'causar una alegría', entre las que está el frecuente verbo *dar* (como en *dar una alegría* [*a alguien*]) y también el no tan frecuente *despertar*. Lo que caracteriza específicamente a las colocaciones es la coocurrencia léxica restringida entre los dos constituyentes de la colocación. La coocurrencia de dos unidades léxicas L_1 y L_2 es léxicamente restringida si, para expresar un significado 'L_2' aplicándose a la unidad léxica L_1, la elección de L_2, que expresa el significado 'L_2', está léxicamente determinada por L_1. La combinación de L_1 y L_2 formará una colocación, en donde L_1 es la *base* y L_2, el *colocativo*, en los términos de Hausmann (1979); en el caso mencionado, *alegría* será la base y los verbos son los colocativos.

Cada colocación recibe una descripción semántica y sintáctica y es atestiguada con varios ejemplos, la mayoría extraídos del *CREA*. Para describir las colocaciones, usamos las *funciones léxicas* (FL, Mel'čuk *et al.* 1995). Una FL codifica la relación entre dos unidades léxicas de las cuales una de ellas (la *base* de la colocación) controla la elección léxica de la otra (el *colocativo*). Por ejemplo, la FL Magn codifica la relación entre *aburrimiento* y los siguientes adjetivos: *profundo*, *mortífero*, *solemne*, *inmenso*, etc. Todos son seleccionados para expresar, en combinación con *aburrimiento*, aproximadamente el mismo significado, 'intenso', aunque no todos son igual de frecuentes ni del mismo registro.

Como hemos dicho, todas las colocaciones están apoyadas en ejemplos extraídos del corpus. Ahora bien, aunque el *DiCE* se compila inductivamente a partir del corpus, las FLs constituyen una plantilla que guía la búsqueda de colocaciones en el corpus. El corpus, por tanto, es filtrado desde el principio con búsquedas específicas. Así, por ejemplo, a la hora de buscar las colocaciones de un nombre como *opinión*, el análisis semántico junto con la plantilla de las FLs lleva al lexicógrafo a buscar colocativos específicos. No buscará, por ejemplo, un valor de la FL Magn porque el significado de ese nombre no es compatible con la intensificación, pero sí buscará, por ejemplo, adjetivos que expresan cuántas personas coinciden en

la opinión (*mayoritaria*, *generalizada*, *compartida*, *personal*) o adjetivos que caracterizan si el contenido de la opinión es positivo o negativo (*buena*, *mala*, *contraria*, *favorable*).

3. SOBRE LA RELEVANCIA DEL CONCEPTO DE FRECUENCIA LÉXICA

El concepto de frecuencia léxica ha estado siempre vinculado a diccionarios y al mundo de enseñanza de lenguas. Los estudios de frecuencias léxicas se originaron en relación con el aprendizaje del vocabulario y hoy siguen estando vinculados; de hecho, el diccionario de frecuencias de español publicado por Davies (2006) se subtitula "core vocabulary for learners". Los diccionarios de frecuencias son concebidos con una finalidad especialmente didáctica: "comprobar la frecuencia de las palabras para poder usar en obras pedagógicas las palabras más frecuentes" (Haensch & Omeñaca 2004: 154). Además de los diccionarios de frecuencia propiamente dichos, la información de frecuencia de las palabras puede ser útil para la Lexicografía, puesto que los recuentos pueden servir de criterio para incluir o excluir material en su macro o en su microestructura y, asimismo, la frecuencia de las acepciones puede servir para ordenarlas dentro de la entrada. La relevancia de la información de frecuencia en el ámbito del aprendizaje de lenguas ha sido señalada reiteradamente, aludiendo principalmente a conceptos como *utilidad*, *cobertura* y *rentabilidad comunicativa*. Expongamos primero las virtudes de la frecuencia léxica para pasar más tarde a presentar algunos de sus defectos.

Ante la enormidad del léxico de una lengua y el dilema de un lexicógrafo (qué incluir), de un profesor de lengua (qué enseñar) o de un autor de material didáctico (qué incluir para que un profesor enseñe o para que un alumno aprenda), la frecuencia léxica viene con el halo de objetividad y de base científica, que sirve para contrarrestar la subjetividad que puede caracterizar a los diccionarios y para racionalizar el aprendizaje del vocabulario. Así, por ejemplo, Alvar Ezquerra (2004) insiste en que se deben enseñar antes las palabras más frecuentes a los estudiantes de L2 porque aparecen en todos los textos y son de alto rendimiento. Asimismo, McEnery y Rayson, prologando el diccionario de Davies (2006), argumentan que parece razonable priorizar el aprendizaje de palabras que tienen más probabilidades de ser oídas o usadas a menudo. La rentabilidad comunicativa es el argumento mencionado por otros autores. Por ejemplo, Almela *et al.* (2005) señalan que es mucho más rentable aprender una palabra muy usada y altamente polisémica que otra poco usada. Para medir la rentabilidad, se suele acceder al concepto de cobertura. Según Alvar Ezquerra (2003: 100), con las cinco mil palabras más frecuentes se cubre el 90% de un texto. Asimismo, para el inglés, Nation (2001: 11-13) propone incidir en las dos mil o tres mil palabras más frecuentes porque son estas las que permiten al aprendiz desenvolverse en las situaciones cotidianas. También señala que las cuatro mil o cinco mil palabras más frecuentes dan cuenta de hasta un 95% de un texto escrito[4]. Para medir la utilidad o importancia de la palabra, además de la frecuencia, se suele tener en cuenta la *dispersión*: una palabra que aparece en diferentes tipos de obras y de registros debe ser considerada más útil que otra que, aunque sea más frecuente, solo aparezca en un registro. En Davies (2006), el concepto de dispersión tiene gran importancia en la fórmula

[4] Según Izquierdo (2004: 337), los resultados sobre cobertura textual parecen coincidir independientemente de la lengua y del corpus.

para ordenar las palabas; por ejemplo, *desconfianza* aparece en la posición 3170 con 247 ocurrencias, frente a *pánico* que ocupa la posición 3227, con 297 ocurrencias, debido a que la primera tiene más dispersión que la segunda en el corpus explotado por Davies. Tanto la frecuencia como la dispersión son los criterios utilizados para asignar un índice de uso o de utilidad. Así, por ejemplo, Koprowski (2005) evalúa la utilidad de las expresiones fraseológicas utilizadas en tres manuales de inglés como L2 midiendo su frecuencia y su dispersión en el corpus *COBUILD*.

Si la frecuencia y la dispersión de una palabra son índices de su utilidad, cabe ahora preguntarse en qué consiste la utilidad de una palabra. Como señala Bogaards (1994: 111), se ofrecen criterios objetivos para seleccionar palabras útiles, pero la utilidad de una palabra es una noción vaga. Ni las palabras ni ninguna otra cosa son útiles de una manera general e indeterminada, sino que tendrán mayor o menor utilidad dependiendo de su adecuación al objetivo. En palabras de Bogaards (1994): "est utile ce qui sert à atteindre un but qu'on s'est fixé". Así, por ejemplo, para un médico profesional español que quiera leer publicaciones en inglés, le serán útiles palabras de su campo como *anaesthesia* o *bleed*, que no figuran entre las cinco mil palabras más frecuentes[5]. Igualmente, el que aprende la L2 en el país donde la lengua se habla normalmente tiene especial interés en las palabras que le sirven para su actividad cotidiana en donde entran palabras poco frecuentes como *empadronamiento* o *tarjeta sanitaria*[6]. Es necesario hacer también una importante distinción entre el vocabulario receptivo y el productivo (Bogaards 1994: 115). Dado que para comprender lo que se dice o se escribe, el aprendiz depende de las palabras que los demás escogen, en la selección del vocabulario para la competencia receptiva, el factor frecuencia puede tener un papel importante. Sin embargo, en producción, el aprendiz escoge los elementos que tiene a su disposición que no son necesariamente los más frecuentes.

Como vemos, la utilidad no está determinada siempre por la frecuencia. La ausencia de palabras de uso común en los listados de frecuencias fue lo que llevó a Gougenheim y a otros investigadores franceses (Gougenheim *et al.* 1964) a proponer los estudios de *disponibilidad léxica*, que han tenido mucho seguimiento en España[7]. Palabras que refieren a realidades cotidianas como *ducha* y *cepillar*(*se*) empleadas habitualmente por cualquier hablante nativo no son frecuentes, pero son disponibles si se activa un centro de interés como 'aseo personal', por ejemplo. Hay, por tanto, palabras muy útiles pero poco frecuentes y viceversa: muchas de las palabas más frecuentes son muy poco precisas y con escaso valor informativo, lo que necesariamente les resta rentabilidad. Bogaards (1994: 116-119), en un divertido experi-

[5] Siempre es posible establecer las frecuencias dentro de los campos específicos, tal y como se estudia en Lenguas para fines específicos, pero con todo, es dudoso que el criterio más importante para un médico sea aprender los términos de su especialidad por orden de frecuencias.

[6] Estas palabras proceden de la sección *Papeleo* del *Manual de español para inmigrantes Nivel A1* (consultable en www.obrasocialcajamadrid.es/Ficheros/CMA/ficheros/OSEduca_EPIManual.PDF). Obviamente conocer el vocabulario asociado al papeleo es algo útil para un inmigrante. Creemos que la proliferación en los últimos años de cursos de ELE para inmigrantes es una clara muestra de que no todas las palabras son útiles para cualquier aprendiz, puesto que cada uno tiene sus prioridades comunicativas.

[7] Para más información, puede consultarse la página web sobre el *Proyecto panhispánico de léxico disponible*, dirigido por López Morales (www.dispolex.com).

mento, borró de un texto periodístico todas las palabras plenas, con mayor contenido semántico, y dejó solo las que entran en las 500 más frecuentes, y demostró que el texto era incomprensible. Solo cuando subió hasta las cinco mil más frecuentes, el texto era adivinable, pero todavía muy pobre informativamente. Probó también a la inversa: tras eliminar de un texto las palabras que entran en la lista de las mil más frecuentes, mostró que lo esencial del mensaje se había transmitido. Por lo tanto, la equivalencia entre frecuencia y utilidad debe ser muy matizada. También Izquierdo (2004: 360) subraya que la frecuencia no puede ser el único criterio en el que determinar la selección léxica del material destinado a aprendices, puesto que estos necesitan conocer palabras con alto contenido informativo para poder comunicarse.

Si los estudios sobre disponibilidad sirvieron para poner de relevancia las palabras útiles, aunque poco frecuentes, solo Galisson (1979) ha subrayado la importancia del eje sintagmático para medir la disponibilidad: "on reconnaît souvent l'étranger à ce que, tout en faisant un choix correcte du vocable qui rend compte de l'essentiel de ce qu'il veut signifier, il n'emploie pas avec celui-ci les cooccurrents attendus" (Galisson 1979: 58). Así, propuso encuestas de disponibilidad sintagmática en donde se preguntaba por los verbos y adjetivos que vienen a la mente en coocurrencia con determinados nombres. Efectivamente, para aprender palabras como *ducha* o *cepillarse*, que mencionábamos arriba, y poder usarlas es necesario saber cómo se combinan: *ducha* coocurre con *darse*, *pegarse* y *tomar*; *cepillarse* es sinónimo de *lavar* cuando va con *dientes* pero no cuando va con *pelo* y la frecuencia en el corpus de *cepillarse* con *pelo* es diferente de la de *cepillarse* con *dientes*[8]. La frecuencia de las palabras aisladas no es directamente utilizable ni por el lexicógrafo ni por el autor de material didáctico que se interese por las colocaciones. Tomemos el ejemplo de verbos muy frecuentes como *dar*, *tomar*, *hacer* o *tener*, que funcionan a menudo como verbos de apoyo. De nada sirve incluir estos cuatro verbos en un manual de ELE (o seleccionarlos para un vocabulario de ELE) sin el nombre que los selecciona léxicamente porque un aprendiz tendrá que aprender que la combinatoria de estos verbos está controlada por el nombre de la base: los paseos se *dan*, los viajes se *hacen*, al igual que las proposiciones, pero no como las decisiones que se *toman*. En otras palabras, incluir esos cuatro verbos como prioritarios en el aprendizaje del vocabulario sin los nombres asociados no parece adecuado porque muchas de las palabras frecuentes lo son porque son escogidas frecuentemente como colocativos de bases que son menos frecuentes. Así, se puede dar el caso de dejar fuera de la lista de lo priorizable las bases que seleccionan los colocativos frecuentes[9]. Sánchez (2000: 24) incide en la misma idea de que los aprendices comunicarán más fácilmente si conocen no solo las palabras más frecuentes sino las que las acompañan, que pueden ser no tan frecuentes. Menciona el caso de *teléfono*, que está entre las mil más frecuentes, pero no algunas de las que la acompañan usualmente. Podríamos dar un paso más: incluso aunque las palabras que la acompañen sean frecuentes, como es el caso de *coger*, el significado de este verbo con *teléfono* nada tiene que ver al que tiene con *manzana*, pongamos por caso.

[8] En el *CREA* encontramos 63 apariciones con *dientes*, frente a 43 con *pelo*.

[9] En el ámbito del inglés como L2, varios autores llamaron la atención sobre el hecho de que los llamados *delexical verbs* no pueden ser enseñados sin referencia a las colocaciones en que entran. *Vid*, por ejemplo, Sinclair & Renouf (1996: 152).

Y este es otro de los problemas planteados por la mayoría de las listas de frecuencias[10]: cuentan formas o lemas, no unidades léxicas. Como ha señalado Bogaards (2008: 1234), entre otros, existe una fuerte correlación entre frecuencia y polisemia: cuantos más sentidos tenga una palabra, más ocurrencias tiene. El hecho de que *interés* aparezca con una frecuencia alta comparada a otros nombres de sentimiento o de estados mentales no quiere decir que sea el nombre de sentimiento más frecuente, sino la palabra más frecuente que incluye entre una de sus acepciones la referida a la 'inclinación del ánimo'. La falta de discriminación de la unidad léxica puede dar una impresión no totalmente ajustada a la realidad. Así, por ejemplo, en Davies (2006), *carta* recibe una posición dada, sin distinción entre el sentido 'letter' y el de 'playing card', a pesar de que es probable que la frecuencia de los dos sentidos no sea la misma. Igualmente se le asigna la misma frecuencia y orden en la lista a *cólera* 'enfermedad' y al 'sentimiento'. Sin embargo, el aprendiz necesita aprender formas vinculadas a sentidos y el lexicógrafo necesita saber qué acepciones son más frecuentes que otras. Esa es la meta formulada por Rojo (2009) cuando señala que los diccionarios basados en corpus, no solo deberían dar la frecuencia de las palabras sino de las acepciones. El problema es que esto solo es posible con un corpus desambiguado semánticamente, tarea ardua donde las haya[11]. Un corpus sin desambiguar puede ayudar al hablante nativo pero no tanto al aprendiz de la lengua. A pesar de que en los últimos años se ha promovido el uso de las concordancias para que los aprendices puedan consultar directamente los corpus, la dificultad, en ocasiones, es demasiado alta para que un aprendiz de nivel intermedio pueda distinguir los diferentes sentidos de una palabra en las concordancias (Aston *et al.* 2004). Aunque hay cierta tendencia a creer, sobre todo desde la óptica sinclairiana, que los corpus hablan solos[12], pensamos que es necesario un filtrado y una descripción que solo puede venir bien de un diccionario, bien de un corpus anotado (Alonso Ramos 2009).

Las frecuencias de las unidades léxicas son cruciales también para los tests de vocabulario de los aprendices. Por ejemplo, Casso (2010: 82) se basa en la lista de las cinco mil palabras más frecuentes del corpus *Cumbre* para extraer de ahí un test de vocabulario receptivo en donde el aprendiz debe vincular glosas o definiciones mínimas con una lista de palabras, en la línea de los test propuestos por Nation (2001). Dado que la lista no refleja la frecuencia de las acepciones, al construir las definiciones para el test, el autor se ve obligado a actuar por intuición y elegir el sentido de la palabra que cree más frecuente. Así por ejemplo para el adjetivo *ligero* propone la definición 'no pesa'. Sin embargo, si consultamos el *CREA* buscando solo la forma "ligero" y con un primer filtrado, en la primera página de resultados

[10] Una excepción es West (1953), que indica cuántas veces se emplea cada uno de los sentidos de las palabras seleccionadas.

[11] Existe todo un ámbito de investigación, conocido en inglés como *Word sense disambiguation*, que trata de la desambiguación automática de los sentidos de las palabras. *Vid.* Agirre & Edmonds (2006).

[12] Según Moon (2008), en el momento de redactar el *Collins Cobuild* se consideró incluir las secuencias de ejemplos antes de las definiciones con el objetivo de mostrar las pruebas antes de la explicación y así permitir a los usuarios a localizar el significado heurísticamente. Más tarde, Sinclair planeó un diccionario de colocaciones, estructurado simplemente alrededor de las concordancias, que nunca fue completado. Esta idea hubiera sido el prototipo de diccionario en donde las palabras no son más que los puntos de acceso al corpus en donde los textos muestran el significado.

no devuelve ni un solo ejemplo en donde signifique 'que pesa poco' pero sí, muchos ejemplos del adjetivo en colocación con los nombres siguientes: *retroceso*, *aumento*, *recorte*, *descanso*, etc. en donde el adjetivo sirve para atenuar la intensidad de lo designado por el nombre. No queremos decir que este sea necesariamente el primer sentido que debamos enseñar al aprendiz y quizás haya razones pedagógicas para enseñar prioritariamente los sentidos más físicos. Lo que nos gustaría destacar es que otra vez por vía de la objetividad de las frecuencias, se nos cuela la subjetividad, al asignar la mayor frecuencia a un sentido dado, basándose en la intuición.

Además de enriquecer el corpus con la desambiguación semántica para mejorar el valor de las frecuencias léxicas, también sería necesaria una identificación de locuciones o expresiones fraseológicas antes de hacer los recuentos. La razón de que tanto en Davies (2006) como en Almela *et al.* (2005), *vez* esté entre los tres nombres más frecuentes es que esta forma es muy recurrente en locuciones como *a la vez*, *a veces*, *de una vez*, *de vez en cuando*, *tal vez*, etc. Sería necesario realmente saber la frecuencia de todas estas locuciones independientemente del nombre *vez*. Lo mismo ocurre en todas las listas de frecuencia consultadas con *embargo* (Almela *et al.* 2005: 28)[13]. Al ver ese nombre en posiciones tan altas, quien no conozca la locución *sin embargo*, podría pensar que a los españoles los embargos nos quitan el sueño (y a pesar de la crisis, no es para tanto).

Se desprende que merece la pena el esfuerzo en anotar sentidos y locuciones en los corpus porque de corpus enriquecidos se derivan resultados más fidedignos. Con todo, no basta con un corpus anotado: todavía más importante es que el corpus sea equilibrado. De no ser así, de nuevo, la objetividad de los números se ve reducida por la subjetividad en la selección de muestras que constituyen el corpus. Con todo, por muy equilibrado que sea el corpus, no podemos dejar de ver los dados del azar en el hecho de que una palabra entre en las cinco mil más frecuentes o se quede fuera; o más en particular, en lo que nos concierne, el hecho de que una colocación aparezca o no en un corpus, por muy representativo que este sea. En la sección siguiente, veremos si el azar desempeña o no un papel en el orden que tienen los nombres de sentimiento y sus colocaciones, en diferentes listas de frecuencias y en diferentes corpus.

4. SELECCIÓN DE NOMBRES DE SENTIMIENTO Y SUS COLOCATIVOS

Dado que nuestro objetivo es seleccionar y ordenar los datos contenidos en el *DiCE*, creemos oportuno revisar lo que otros diccionarios y listas de frecuencias han seleccionado y qué orden han dado a los nombres de nuestra nomenclatura, es decir las bases de las colocaciones, y a los colocativos. Para explorar las bases, hemos elegido las siguientes tres listas de frecuencia, principalmente por seleccionar las cinco mil palabras más frecuentes y estar basadas en corpus del español actual: Davies (2006), el corpus *Cumbre* (Almela *et al.* 2005) y la lista de las cinco mil formas más frecuentes del *CREA* accesible en la web[14]. Para examinar la

[13] Una excepción es Ávila (1999), que ha prestado especial atención a los marcadores discursivos como *o sea*, *por ejemplo*, *a lo mejor*, etc.

[14] Desde la página web de la RAE se puede acceder a distintos listados de frecuencias: http://corpus.rae.es/lfrecuencias.html.

selección de colocativos, optamos por los diccionarios de colocaciones. Hemos escogido dos diccionarios para el inglés, el *OCD* (McIntosh 2009) y el *McMillan* (Rundell 2010), y para el español, el *Práctico* (Bosque 2006). Nos limitaremos a hacer algunas calas en la descripción de los adjetivos colocativos para mostrar las diferentes selecciones que hace cada diccionario.

4.1. Bases en las listas de frecuencia y en vocabularios de aprendizaje

De los 211 nombres que constituyen la actual nomenclatura del *DiCE*, los datos son los siguientes: (1) en Davies (2006) figuran 96, (2) en el *Cumbre* aparecen 71, (3) en la lista de la RAE, solo 48. La primera observación concierne a la lista del RAE, puesto que trata formas y no lemas, por lo que los datos hay que interpretarlos con precaución. Así, no es posible saber si la posición de *deseo*, *gana*, *ganas* y algún otro, se refiere al nombre o al verbo. Dejando aparte *pesar*, *disgusto* y *dicha*, que aparecen en los primeros puestos de la RAE debido a la homonimia con el verbo y también a la locución *a pesar de*, los diez primeros nombres de la lista de la RAE entran entre los doce primeros de las otras dos listas, aunque no exactamente ordenados igual.

Nombres del *DiCE*	**Puesto[15] en Davies**	**Puesto en *Cumbre***	**Puesto en *CREA***
amor	2	2	1
atención	3	3	3
confianza	12	10	7
deseo	6	7	6
dolor	8	8	5
esperanza	9	5	10
interés	1	1	2
miedo	4	4	4
pena	5	6	8
respeto	11	11	9

Tabla 1. Los 10 nombres del *DiCE* más frecuentes

De los doce primeros nombres en Davies, los dos que no figuran en la tabla son *gusto* (posición 7) y *sentimiento* (10). Con respecto al *Cumbre*, la única ausencia es también *sentimiento* en la posición 9. El nombre *gusto* no aparece en el *Cumbre* entre las 5 mil más frecuentes y en el *CREA* aparece en el puesto 1107. En cuanto a *sentimiento*, en el *CREA* aparece la forma plural en el puesto 2067 y en singular en el 2106. Salvo estos desajustes, podemos decir que en los primeros puestos las listas coinciden bastante.

Intuitivamente, pueden sorprender algunas ausencias como *antipatía*, *enemistad* o *enfado*, que no figuran en ninguna de las listas. Ninguno de estos nombres parece especialmente marcado en un nivel de registro. Así, por ejemplo, parece que encaja con la intuición el que aparezca *asco* en Davies en el puesto 4375, pero no *repugnancia*, dado que esta palabra

[15] El puesto en las listas de frecuencia es con respecto a los nombres que configuran la nomenclatura del *DiCE*; es decir no es que *amor* figure en el puesto 2 de las 5 mil más frecuentes, sino que de los nombres del *DiCE*, *amor* ocupa el segundo más frecuente en la lista de Davies.

no es usada en lengua oral y Davies ha favorecido las palabras que aparecen en los subcorpus orales, pero con el mismo criterio no se entiende por qué *chasco* queda fuera, mientras que *decepción* entra. Una razón puede deberse a que quizás *chasco* tiene menos distribución que *decepción*. En definitiva, los datos sobre la frecuencia no tienen que coincidir necesariamente con la intuición del hablante nativo[16]. Lo que importa es que tenemos unos datos que han sido recogidos en el caso especialmente de Davies con criterios explícitos.

Comparemos ahora la selección de nombres de sentimiento que hacen tres autores que se dirigen a aprendices de ELE: la selección de Izquierdo (2004: 447), orientada como vocabulario disponible; el *Léxico fundamental* (LF) de Sánchez Lobato & Aguirre (1992), sin información de cómo se ha hecho la selección ni a quién va dirigido; y, por último, el *Vocabulario medio B1* (VM) de Baralo *et al.* (2009), dirigido a estudiantes de nivel B1. Como se puede observar en la siguiente tabla, las coincidencias entre las tres selecciones de nombres de sentimiento y estados de ánimo no son muchas; solo cuatro nombres coinciden en las tres: *alegría*, *amistad*, *humor* y *miedo*. Hay otras coincidencias entre dos. Así, tanto Izquierdo como LF seleccionan *amor*, *cariño*, *felicidad*, *odio* y *tristeza,* mientras que LF y VM coinciden en proponer *pena*.

[16] Para Nation (2001: 27), sin embargo, la frecuencia es uno de los conocimientos implicados en saber una palabra.

Nombres de sentimiento (N_{sent})	**Izquierdo**	**LF**	**VM**	**Puesto en N_{sent} Davies**	**Puesto en todo Davies**
aburrimiento			x	**86**	4488
afecto		x		35	2132
alegría	x	x	x	24	1429
amistad	x	x	x	18	1232
amor	x	x		**2**	423
angustia		x		32	1957
cariño	x	x		40	2246
celos	x			49[17]	2924
depresión			x	64	3419
desesperación		x		47	2780
disgusto		x		68	3660
diversión			x	62	3345
dolor		x		8	705
emoción		x		27	1577
enemistad		x		No	No
enamoramiento			x	No	No
enfado			x	No	No
envidia	x			56	3272
esperanza	x			9	805
felicidad	x	x		31	1892
humor	x	x	x	30	1889
ilusión		x		23	1372
lástima			x	60	3294
miedo	x	x	x	4	450
nostalgia	x			58	3290
odio	x	x		36	2177
pena		x	x	5	575
preocupación			x	13	996
respeto	X			11	894
satisfacción			x	21	1299
sentimiento		x		10	2646
simpatía		x		41	2327
sufrimiento		x		46	2658
temor		x		15	1154
ternura		x		70	3797
tristeza	x	x		39	2233
vergüenza	x			38	2206
Total: 37	Total: 15	Total: 23	Total: 13		

Tabla 2. Nombres de sentimiento en Vocabularios para ELE

De los cuatro nombres en que coinciden los tres vocabularios, solo *miedo* figura entre las mil palabras más frecuentes. El nombre *amor* sería la palabra más alta en un listado de

[17] Recordemos que Davies no distingue entre *celo* y *celos*.

frecuencias y *aburrimiento*, la menos frecuente. Como se puede ver en la tabla 3, solo siete nombres entran en las mil palabras más frecuentes, pero curiosamente, tres nombres están fuera de las cinco mil más frecuentes: *enemistad*, *enfado* y *enamoramiento*. Con todo, aunque ninguno de los tres autores se apoye explícitamente en la frecuencia, hay que señalar que si de 37 solo 4 quedan fuera de cinco mil más frecuentes, parece haber cierta correlación entre importancia para el aprendizaje y frecuencia.

Entre las 1000	**Entre las 2000**	**Entre las 3000**	**Entre las 4000**	**Entre las 5000**
amor	alegría	afecto	depresión	aburrimiento
dolor	amistad	cariño	diversión	
esperanza	angustia	celos	disgusto	
miedo	emoción	desesperación	envidia	
pena	felicidad	odio	lástima	
preocupación	humor	sentimiento	nostalgia	
respeto	ilusión	simpatía	ternura	
	satisfacción	sufrimiento		
	temor	tristeza		
		vergüenza		

Tabla 3. Distribución de 37 Nsent entre las 5000 palabras más frecuentes

4.2. Colocativos en diccionarios de colocaciones y corpus

Para poder encontrar algún criterio que nos ayude a incluir o a excluir un determinado colocativo, hemos querido examinar la selección de colocativos hecha por dos diccionarios de colocaciones del inglés. Hemos comparado los colocativos adjetivos del nombre *fear* en el *OCD* y en el *McMillan*[18]. Puesto que cada diccionario se ha basado en diferentes corpus, puede ser instructivo comparar las ocurrencias de estas colocaciones en un único corpus. Elegimos el corpus Internet-En[19] (150 millones de palabras) porque tiene un corpus paralelo en español, con el que poder comparar más tarde.

[18] *McMillan* distingue dos sentidos para *fear*. Algunos adjetivos son compartidos por los dos, como es el caso de *genuine*, lo que señalamos en la siguiente tabla con "x, 2". Si solo se combina con la UL *fear* 2, añadimos el número 2 a la x.

[19] Este corpus así como el Internet-ES son accesibles desde la interfaz Intellitext desarrollada por la Universidad de Leeds y accesible en la web: http://corpus.leeds.ac.uk/it/.

Collocates(fear)	**OCD**	**McMillan**	**Intellitext**[20]
big	x		8
biggest 2		x	46
childhood	x		11
constant	x	x	43
deep	x	x	19
deepest 2		x	22
deep-seated	x		**1**
general	x		10
genuine	x	x, 2	14
great	x	x	51
greatest		x 2	57
groundless		x 2	**1**
growing	x		17
intense	x	x	36
irrational	x	x 2	53
justified		x 2	4
legitimate	x	x 2	12
mortal	x	x	14
nagging	x	x	3
overwhelming	x		4
paralysing	x		9
paranoid		x 2	5
primal	x		8
public	x		24
Pure	x		**2**
Real	x	x, 2	50
terrible	x	x	9
understandable		x 2	6
unfounded	x	x 2	9
unjustified 2		x 2	3
unreasonable	x		12
utter	x		**1**
well-founded	x	x 2	10
widespread	x		18
worst	x	x 2	**79**
	total: 27	total: 21	

Tabla 4. Adjetivos colocativos de *fear*

[20] Para las consultas en Intellitext nos hemos limitado a buscar el adjetivo antepuesto al nombre con una distancia de dos palabras con la siguiente fórmula: [word="worst"] [word!="\W"]{0,2} [lemma="fear"]. Hemos filtrado manualmente los resultados.

Como vemos, hay diferencias entre ambos diccionarios[21]: *McMillan* desecha *big* para *fear* 1, pero elige el superlativo *biggest* para 2. También rechaza *childhood*, *deep-seated*, *general*, *growing*, *primal*, *pure*, *public*, *unreasonable*, *utter* y *widespread*. Por su parte, el *OCD* solo distingue una forma superlativa y desecha *biggest*, *deepest*, *greatest*. Tampoco incluye la serie *justified*, *legitimate*, *understandable*, pero sí *well-founded*. En cuanto a sus antónimos, rechaza *groundless*, *irrational***,** *paranoid* y *unjustified* y elige *unfounded*. Estas ausencias pueden basarse en diferentes razones: (1) diferentes criterios de si una combinación dada es o no es colocación, (2) diferente frecuencia en el corpus de referencia en el que se hayan basado, (3) error o lapsus. El corpus en el que se basa el McMillan es de mil seiscientos millones de palabras, mientras que el corpus de Oxford era en la primavera de 2010 de 2000 millones de palabras[22]. Ambos diccionarios escogen colocaciones que aparecen 1 o dos veces en este corpus. Los desajustes entre los dos diccionarios son mayores cuanto menor es el número de ocurrencias, pero tampoco esta correlación es exacta: por ejemplo, *growing* y *widespread* que tienen, respectivamente, 17 y 18 ocurrencias, solo son escogidos por el *OCD*. E igualmente, el *McMillan* elige *paranoid* con solo 5 ocurrencias, mientras que rechaza *unreasonable* con 12. Quizás los datos en sus corpus respectivos son diferentes de lo que muestra nuestro corpus de referencia, pero lo que queremos indicar es que no podemos tener la "certeza estática" de que tal colocación es más frecuente y por eso se ha escogido en un diccionario, sino que es más frecuente en un corpus dado y, por esa y otras razones, se ha escogido.

Examinemos ahora los adjetivos que se combinan con *miedo* comparando el *Práctico* con el *DiCE*. Asignamos también el número de ocurrencias en el corpus Internet-ES de Intellitext para poder comparar los dos diccionarios. El *Práctico* está basado en un corpus periodístico de 250 millones de palabas. Los datos del *DiCE* proceden, en su mayor parte, del *CREA*, pero limitándose a España.

[21] También Bogaards (2000) llama la atención sobre los desajustes entre las palabras consideradas frecuentes en cinco diccionarios de aprendices de inglés.

[22] Tomo la información de la página de diccionarios Oxford: http://oxforddictionaries.com/words/about-the-oxford-english-corpus [febrero 2012].

Colocativos(miedo)	**Práctico**	**DiCE**	**Intellitext**[23]
abrumador		x	0
ancestral		x	6
animal		x	1
arraigado	x		0
atávico	x		2
atenazador	x		0
aterrador	x		2
atroz	x	x	6
cerval	x	x	4
clínico	x		0
escénico	x	x	28
espantoso	x	x	1
feroz		x	2
fuerte		x	5
fundado	x	x	9
general	x		4
generalizado		x	0
horrible		x	2
horroroso	x	x	1
imponente		x	0
incontenible		x	0
inevitable	x		1
infundado	x	x	3
injustificado	x	x	5
insoportable	x		2
instintivo	x		4
insuperable	x		29
intenso		x	33
irracional	x		**53**
irrefrenable	x		0
irreprimible	x		0
justificado	x	x	5
latente	x		6
palpable	x		1
profundo		x	19
pueril		x	3
sacral		x	0
soterrado	x		0
terrible	x		23
verdadero		x	11
visceral	x	x	3
	total: 28	total: 22	

Tabla 5. Adjetivos colocativos de *miedo*

[23] Para la búsqueda en español hemos tenido en cuenta la posición antepuesta y pospuesta del adjetivo y hemos permitido hasta 5 palabras intermedias para poder dar cuenta de ejemplos como *el miedo al rechazo es muy fuerte*. También hemos filtrado manualmente los resultados.

El número de adjetivos elegidos por los diccionarios españoles es aproximadamente el mismo de los diccionarios ingleses: los cuatro ofrecen alrededor de una veintena de adjetivos. Al igual que ocurría antes, en este par de diccionarios no existe una perfecta coincidencia. Como vemos, hasta llegar al adjetivo *atroz*, cada diccionario ofrece unos adjetivos distintos. Y otra vez, las razones de las ausencias y de las coincidencias pueden ser varias. Aquí llama la atención que once adjetivos no ofrezcan ni una sola ocurrencia en el corpus español de Intellitext, que es de 145,6 millones de palabras. Las ocurrencias 0 están repartidas entre ambos diccionarios: el *Práctico*, 6 (*arraigado*, *atenazador*, *clínico*, *irrefrenable*, *irreprimible*, *soterrado*) y el *DiCE*, 5 (*abrumador*, *generalizado*, *imponente*, *incontenible* y *sacral*). Sin embargo, si consultamos otro corpus, las ocurrencias 0 se reducen.

Colocativos(miedo)	**Práctico**	***DiCE***	***CREA***
abrumador		x	1
arraigado	x		2
atenazador	x		0
clínico	x		0
generalizado		x	5
imponente		x	0
incontenible		x	2
irrefrenable	x		0
irreprimible	x		0
sacral		x	1
soterrado	x		3

Tabla 6. Ocurrencias 0 en Intellitext contrastadas con el *CREA*

Comparando los datos con los diccionarios ingleses, sorprende también el menor número de ocurrencias de las colocaciones españolas, puesto que los dos corpus son bastante aproximados en cuanto al número de palabras. Si en inglés se llega a 79 ocurrencias de una colocación (*worst fear*), en español el número más alto es de 53 (*miedo irracional*). Solo pasan la decena de ocurrencias seis colocaciones: *verdadero*, *profundo*, *terrible*, *escénico*, *insuperable*, *intenso* e *irracional*. Y tampoco puede decirse que los diccionarios españoles recojan todos los adjetivos colocativos relevantes. Contrastando con los diccionarios ingleses, en ambos diccionarios se echa de menos el adjetivo *constante*, con 15 ocurrencias en el Intellitext. Se podrían añadir otros como *excesivo* o *enorme* que también coocurren con *miedo*, pero este es el problema: hasta dónde incluir. Si fijamos límites como desechar los adjetivos con menos de 5 ocurrencias, dejamos fuera *miedo infundado*, que es una colocación especialmente útil e idiomática. Fijando los límites por arriba, ocurre lo mismo: si elegimos los adjetivos que tengan más de 20 ocurrencias entraría, por ejemplo, *escénico*, pero no *atroz* ni *profundo* que por su sentido más genérico pueden ser más útiles que el más frecuente.

4.3. ¿Es útil la información de frecuencia para el *DiCE*?

Sí. A pesar de todos los problemas que hemos planteado, pensamos que es útil esa información. Aunque el azar juega sus dados, hay que decir que de los 115 nombres del *DiCE* que no entran en las cinco mil palabras más frecuentes de Davies, tampoco entran en las otras dos listas examinadas, lo que indica que las tres listas están basadas en corpus equilibrados y representativos. Sin embargo, el examen de las frecuencias de las colocaciones en el corpus

muestra más desajustes: *miedo irracional* es la más frecuente en Intellitext, con 53 apariciones, mientras que en el *CREA* solo tiene 15. Creemos que hay que vivir con esto: todo diccionario basado en un corpus va a llevar su sesgo, pero podemos contrarrestarlo con la introspección. La introspección, aunque considerada un tabú para alguna visión radical de la Lingüística de corpus, permite incluir *miedo infundado* a pesar de su baja frecuencia en el corpus utilizado. Es cierto que da pie a introducir lo subjetivo, pero también tenemos que vivir con eso: un diccionario es la obra de unas personas, con una visión dada de la lengua y de lo que se considera natural o no. Por lo tanto, siguiendo a Bosque (2004: CLVIII), se trata de combinar frecuencia y *naturalidad*, pero, a diferencia de *Redes*, sin mezclarlas; es decir, si nuestros datos proceden de un corpus como Intellitext, tendremos que decir que *miedo infundado* tiene una frecuencia baja, pero si, con todo, lo incluimos en el diccionario, la razón se deberá a que lo consideramos natural y útil para expresar un sentido dado, específico de ese nombre. Así, la información de frecuencia indica lo que es: una colocación aparece mucho o poco en un corpus dado. En una primera fase, puede servir para una limpieza y desechar colocativos y, en una segunda fase, como una indicación útil para el usuario del diccionario: así, de los diferentes intensificadores de *miedo*, es importante que el usuario tenga la información de que *profundo*, *insuperable* e *intenso* son los más frecuentes en nuestro corpus.

5. HACIA UNA METODOLOGÍA DE ASIGNACIÓN DE FRECUENCIAS

Nos disponemos finalmente a esbozar una metodología de asignación de frecuencias al material incluido en el *DiCE* que pueda ser operativo como una guía. Debemos distinguir entre la frecuencia asignada a las bases y a los colocativos.

Para la asignación de frecuencia de las bases, es necesaria la desambiguación de los nombres. Dado que no disponemos de un corpus desambiguado semánticamente, deberemos proceder por muestras del corpus en donde desambiguaremos manualmente y extrapolaremos la frecuencia a las distintas unidades léxicas. Aunque los datos todavía no están cerrados[24], a modo de ilustración, podemos mostrar aquí que *interés*, el nombre más frecuente del *DiCE*, según los datos de Davies (2006) y Almela *et al.* (2005), incluye entre sus acepciones las cinco bandas de frecuencias establecidas por los autores de *Cumbre* (Almela *et al.* 2005: 16) que va de la frecuencia baja (1) a la frecuencia alta (5).

[24] En un próximo artículo presentaremos en detalle la metodología desarrollada por el equipo para la desambiguación semántica de las bases del *DiCE*.

INTERÉS	Cuasisinónimos	EJEMPLOS	frecuencia
interés I.1	inclinación	*Tengo interés por ver el museo*	4
interés I.2	interesante	*Es un monumento de gran interés*	5
interés I.3	beneficio	*Todo lo que hacemos es en interés tuyo*	2
interés I.4	conveniencias	*Si no defiendes tus intereses nadie los defenderá por ti*	1
interés II.1a	rédito	*En cuentas a plazo fijo obtendrás un mayor interés*	4
interés II.1b	cantidad pagada	*Pedimos un préstamo al banco y pagamos unos intereses bastante altos*	1
interés II.2	bienes	*Ese negociante tiene intereses en el extranjero*	2

Tabla 7. Frecuencia de las diferentes unidades léxicas de *interés*

A partir de este primer trabajo, tendremos una lista de unidades léxicas ordenadas por frecuencias, en donde aparecerá *interés* I.2 como la más frecuente, seguida de *amor* I.1a, etc. Y como mostramos, cada unidad léxica tendrá asignada una banda de frecuencia.

Pasemos ahora a esbozar la metodología de asignación de frecuencia a los colocativos. La frecuencia de la base no interviene en que tenga más o menos colocativos[25]. Así, si tomamos una base muy frecuente como *interés*, no se observa que tenga más colocaciones que cualquier otro nombre del *DiCE*. Y, en la lengua en general, los nombres más frecuentes como *año*, *día*, *país*, *gobierno*, *cosa* no son especialmente ricos en colocaciones. En la tabla que sigue indicamos si el colocativo aparece o no entre las cinco mil de Davies (2006), pero creemos que esta información no es realmente relevante para seleccionar una colocación o no, puesto que la frecuencia de, por ejemplo, *fuerte* aislado no tiene por qué ser la misma que con el nombre *miedo*.

La estimación de las probabilidades de que una base dada aparezca con un colocativo es la medida de "colocacionalidad" que utiliza Kilgarriff (2006) para elegir cuáles son las palabras que deben entrar como lemas en un diccionario, pero nosotros no nos preguntamos por las bases sino por los colocativos. Otros autores como Shin & Nation (2005) proponen como criterio de selección exigir que cada colocación tenga que ocurrir al menos 30 veces en un corpus de diez millones de palabras. Sin embargo, esta información no es útil para nuestros objetivos que son seleccionar y ordenar las colocaciones del *DiCE*[26]. Como se muestra en la siguiente tabla, ni una sola colocación *miedo* más adjetivo pasa un umbral mínimo. Según las bandas de frecuencia establecidas en *Cumbre*, hasta 3 apariciones por millón se marcaría como frecuencia baja pero cuando lo aplicamos a las colocaciones, la frecuencia sería extremadamente baja porque no llegan en ningún caso ni a una sola aparición por millón[27]. Como ya hemos dicho, si se trata de dar en un diccionario los adjetivos colocativos

[25] De la misma opinión es Kilgarriff (2006): "common words do not intrinsically have any stronger likelihood to be highly collocational than rare words".

[26] Su metodología no nos sirve por varias razones: entre otras, incluyen como colocaciones expresiones como *you know* y parten de elegir colocaciones cuyo colocativo (en sus términos *pivot*) sea frecuente sin tener en cuenta la frecuencia en relación con la base.

[27] Hay que decir que el tamaño del corpus es diferente. En el caso del *Cumbre* es de 20 millones, así que la correlación por el criterio de Shin & Nation sería de 60 apariciones, asumiendo que la distribución es homogénea, lo que tampoco es evidente.

de *miedo*, no tiene sentido dejar fuera un colocativo como *infundado* solo porque en el corpus utilizado la colocación *miedo infundado* sea poco frecuente. Y todavía esta frecuencia relativa por millón nos plantea otro problema ya que iguala por número de apariciones de la colocación sin tener en cuenta la frecuencia de la base. Así, si tenemos una colocación como *abatimiento profundo*, que ocurre en Intellitext al igual que *miedo infundado* tres veces, se les asignaría la misma frecuencia. Sin embargo, solo la base *miedo* está entre las cinco mil más frecuentes y parece obvio que un aprendiz puede tener más necesidad de usar y de entender la colocación *miedo infundado* que *abatimiento profundo*.

Colocativos(*miedo*)	**Intellitext**	**Frecuencia relativa / millón**	**Davies**
abrumador	0		
arraigado	0		
atenazador	0		
clínico	0		
generalizado	0		x
imponente	0		
incontenible	0		
irrefrenable	0		
irreprimible	0		
sacral	0		
soterrado	0		
animal	1	0,006	
espantoso	1	0,006	x
horroroso	1	0,006	
inevitable	1	0,006	x
palpable	1	0,006	x
atávico	2	0,013	
aterrador	2	0,013	
feroz	2	0,013	x
horrible	2	0,013	x
insoportable	2	0,013	x
infundado	3	0,020	
pueril	3	0,020	
visceral	3	0,020	
cerval	4	0,027	
general	4	0,027	x
instintivo	4	0,027	
fuerte	5	0,034	x
injustificado	5	0,034	
justificado	5	0,034	x
ancestral	6	0,041	
atroz	6	0,041	x
latente	6	0,041	x
fundado	9	0,061	
verdadero	11	0,075	x
profundo	19	0,13	x
terrible	23	0,15	x
escénico	28	0,19	
insuperable	29	0.19	
intenso	33	0,22	x
irracional	53	0,36	x

Tabla 8. Colocativos ordenados por frecuencias

Creemos, por tanto, que es necesario otro procedimiento de ordenación y de selección de las colocaciones. Pensamos que al usuario de un diccionario o al autor de material didáctico que quiera extraer datos, no le interesa tanto saber qué colocativo adjetivo es más o menos frecuente sino elegir el colocativo más frecuente que exprese un significado dado. Debemos, entonces, ordenar por grupos semánticos.

Glosa	**Colocativos(miedo)**	**Corpus**
‘intenso’	abrumador	1
	espantoso	1
	horroroso	1
	horrible	2
	visceral	3
	cerval	4
	fuerte	5
	atroz	6
	profundo	19
	insuperable	29
	intenso	33
‘razonable’	justificado	5
	fundado	9
‘no razonable’	infundado	3
	injustificado	5

Tabla 9. Colocativos ordenados por grupos de sentido

De esta manera, el usuario que quiera escoger un adjetivo que exprese el sentido ‘intenso’, puede elegir entre el más frecuente en el corpus de referencia (en este caso, el adjetivo *intenso*) y el menos frecuente (*abrumador*). A la hora de escoger un adjetivo de un grupo semántico, no interesa su frecuencia absoluta en el corpus sino saber de entre los que aproximadamente expresan el mismo significado cuál es el más y cuál el menos frecuente. Con todo, a la hora de seleccionar colocaciones para material didáctico, por ejemplo, puede interesar tener todas las colocaciones ordenadas y poder saber, por ejemplo, cómo se sitúa *abatimiento profundo* frente a *miedo infundado*. Se trata entonces de normalizar la frecuencia de la colocación y para ello, debemos tener en cuenta la frecuencia de la base. Así, cuanto más frecuente sea la base (recordemos, la unidad léxica, no la palabra polisémica), el puesto en la lista de las bases será más alto; por ejemplo, puesto que *miedo* 1 tiene un puesto mucho más alto que *abatimiento* 1, una colocación formada por *miedo* aunque tenga el mismo número de apariciones que otra formada por *abatimiento* estará ordenada en una posición más alta. De esta manera, si hay que hacer un corte, es preferible que caiga la que tenga menos probabilidades de ser usada y oída, que en este caso es la formada por *abatimiento.* Enfocándonos hacia el material didáctico, un aprendiz de español querrá saber antes un adjetivo que se combina con una palabra que conoce que otro que vaya con una palabra desconocida. Es cierto que desde otro punto de vista puede ser relevante llamar la atención sobre la fuerte coocurrencia entre un nombre y un adjetivo dado; por ejemplo, puede interesar señalar que de, pongamos por caso, cuarenta apariciones de *miedo*, solo en tres va con *infundado*, mientras que, de diez apariciones de *abatimiento*, en tres va con *profundo*. Sin embargo, desde la perspectiva del diccionario, no queremos destacar que un

nombre dado aparezca frecuentemente con un colocativo dado sino cuáles son las colocaciones en las que aparece. A la hora de ordenar el material para ser explotado con fines didácticos, serán necesarios otros factores como el registro o nivel de lengua que orienten a un aprendiz a escoger entre *miedo cerval* y *miedo horroroso*, a pesar de que el primer adjetivo de lengua literaria aparece más frecuentemente que el segundo de lengua hablada en un corpus dado. Por lo tanto, obviamente, la frecuencia no lo es todo, solo es un paso para poder ordenar el material incluido en el *DiCE*. Un paso de hormiga, no de gigante, mientras que no tengamos corpus de referencia desambiguados morfológica y semánticamente. Así que habrá que esperar a contarlo en el homenaje a Guillermo de los 70. No desesperes.

REFERENCIAS BIBLIOGRÁFICAS

AGIRRE, E. & P EDMONDS (eds.) (2006): *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer.

ALMELA, R., P. CANTOS, A. SÁNCHEZ, R. SARMIENTO & M. ALMELA (2005): *Frecuencias del español. Diccionario y estudios léxicos y morfológicos*. Madrid: Editorial Universitas.

ALONSO RAMOS, M. (2004): *Diccionario de colocaciones del español.* <http://www.dicesp.com>.

ALONSO RAMOS, M. (2009): "Hacia un nuevo recurso léxico: ¿fusión entre corpus y diccionario". En P. CANTOS & A. SÁNCHEZ (eds.): *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus*. Murcia: AELINCO, 1191-1207.

ALONSO RAMOS, M. (2010): "No importa si la llamas o no colocación, descríbela". En C. MELLADO *et al.* (eds.): *La fraseografía del S. XXI: Nuevas propuestas para el español y el alemán*. Berlin: Frank & Timme, 55-80.

ALVAR EZQUERRA, M. (2003): *La enseñanza del léxico y el uso del diccionario*. Madrid: Arco/Libros.

ALVAR EZQUERRA, M. (2004): "La frecuencia léxica y su utilidad en la enseñanza de español como lengua extranjera". En M. A. CASTILLO CARBALLO *et al*. (eds.): *Las gramáticas y los diccionarios en la enseñanza de español como segunda lengua: deseo y realidad, Actas del XV Congreso internacional de ASELE*. Sevilla: Universidad de Sevilla, 19-39.

ASTON. G., S. BERNARDINI & D. STEWART (eds.) (2004): *Corpora and Language Learners*. Amsterdam: John Benjamins.

ÁVILA, A. M. (1999): *Léxico de frecuencia del español hablado en la ciudad de Málaga*. Málaga: Universidad de Málaga.

BARALO, M., M. GENÍS & M. E. SANTANA (2009): *Vocabulario medio B1*. Madrid: Anaya.

BOGAARDS, P. (1994) : *Le vocabulaire dans l'apprentissage des langues étrangères*. Paris: Hatier-Didier.

BOGAARDS, P. (2008): "Frequency in Learners' Dictionaries". En E. BERNAL, E. & J. DECESARIS (eds.) *Proceedings of the XIII EURALEX International Congress.* Barcelona: IULA, Universitat Pompeu Fabra, 1231-1236.

BOSQUE, I. (2001). "Sobre el concepto de 'colocación' y sus límites". *Lingüística Española Actual* XXIII/1, 9-40.

BOSQUE, I. (dir.) (2004): *Redes. Diccionario combinatorio del español contemporáneo*. Madrid: SM.

BOSQUE, I. (dir.) (2006): *Diccionario combinatorio práctico del español contemporáneo*. Madrid: SM.

CASSO, J. (2010): *Análisis y revisión crítica de los materiales de evaluación de la competencia léxica. Elaboración de un test de vocabulario de nivel umbral*. Memoria de máster, Universidad Antonio de Nebrija.

CREA = REAL ACADEMIA ESPAÑOLA: Banco de datos [en línea]. *Corpus de referencia del español actual*. <http://www.rae.es>.

DAVIES, M. (2006): *A Frequency Dictionary of Spanish. Core Vocabulary for Learners.* New York: Routledge.

FIRTH, F. R. (1957): "Modes of Meaning". En *Papers in Linguistics* 1934-1951. London: Oxford University Press, 190-215.

GALISSON, R. (1979): *Lexicologie et enseignement des langues*. Paris: Hachette.

GOUGENHEIM, G., R. MICHEA, P. RIVENCL & A. SAUVAGEOT (1964): *L'élaboration du français fondamental (1er degré). Etude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.

HAENSCH, G. & C. OMEÑACA (2004): *Los diccionarios del español en el siglo XXI.* Salamanca: Ediciones Universidad Salamanca.

HAUSMANN, F. J. (1979): "Un dictionnaire des collocations est-il possible? ". *Travaux de Littérature et de Linguistique de l'Université de Strasbourg* 17/1, 187-195

HAUSMANN, F. J. (1989): "Le dictionnaire de collocations". En F. J. HAUSMANN *ET AL.* (eds.): *Wörterbücher – Dictionaries – Dictionnaires*, vol. 1. Berlin: de Gruyter, 1010-1019.

IZQUIERDO, M. C. (2004): *La selección de léxico en la enseñanza del español como lengua extranjera. Su aplicación al nivel elemental en estudiantes francófonos*. Tesis doctoral. Universitat de València.

KILGARRIFF, A. (2006): "Collocationality (and how to measure it)". En E. CORINO *et al.* (eds.): *Proceedings of the Twelfth EURALEX International Congress*. Torino: Accademia della Crusca, Università di Torino, Edizioni dell.Orso Alessandria, 997-1004.

KOPROWSKI, M. (2005): "Investigating the usefulness of lexical phrases in contemporary coursebooks". *ELT Journal* 59/4, 322-332.

MCINTOSH, C., B. FRANCIS & R. POOLE (2009): *Oxford collocations dictionary: for students of English*. Oxford: Oxford University Press.

MEL'ČUK, I. (1998): "Collocations and lexical functions". En A. P. COWIE (ed.), *Phraseology. Theory, Analysis and Applications*. Oxford: Clarendon Press, 23-53.

MEL'ČUK, I. (2006): "Colocaciones en el Diccionario". En M. ALONSO RAMOS (ed.): *Diccionario y fraseología*. Coruña: Universidade da Coruña, 11-43.

MEL'ČUK, I, A. CLAS & A. POLGUÈRE (1995): *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve: Duculot.

MOON, R. (2008): "Sinclair, Phraseology, and Lexicography". *International Journal of Lexicography* 21/3, 243-254.

MUÑIZ, E. (2004): *El concepto de colocación en español.* Tesis doctoral. Universidade de Santiago de Compostela.

NATION, I. S. P. (2001): *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

ROJO, G. (2009): "Sobre la construcción de diccionarios basados en corpus". *Tradumática* 7. <http://webs2002.uab.es/tradumatica/revista/num7/articles/02/02art.htm>.

RUNDELL, M. (2010): *Macmillan Collocations Dictionary.* Oxford: Macmillan.

SÁNCHEZ, A. (2000): "Language Teaching before and after 'Digitalized corpora'. Three main issues". *Cuadernos de Filología Inglesa* 9/1, 5-37.

SÁNCHEZ LOBATO, J. & B. AGUIRRE (1992): *Léxico fundamental del español*. SGEL: Madrid.

SHIN, D. & P. NATION (2008): "Beyond Single Words: the most frequent collocations in spoken English". *ELT Journal* 62/4, 339-348.

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

SINCLAIR, J. & A. RENOUF (1996): "A lexical syllabus for language learning". En R. CARTER & M. MCCARTHY (eds.): *Vocabulary and Language Teaching*. London: Longman, 140-160.

VINCZE, O., E. MOSQUEIRA & M. ALONSO RAMOS (2011): "An online collocation dictionary of Spanish". En I. BOGUSLAVSKY, & L. WANNER (eds.): *Proceedings of the 5th International Conference on Meaning-Text Theory*. Barcelona: Universitat Pompeu Fabra, 275-286.
WEST, M. (1953): *A General Service List of English Words*. London: Longman.