REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
© 2018 PUCV, Chile DOI: 10.4067/S0718-09342018000200153 51(97) 153-174

Asignación de niveles de aprendizaje a las colocaciones del Diccionario de Colocaciones del español*

Assigning proficiency levels to the collocations of the 'Diccionario de Colocaciones del español'

Marcos García Salido
UNIVERSIDADE DA CORUÑA
GRUPO LENGUA Y SOCIEDAD DE LA INFORMACIÓN
ESPAÑA
marcos.garcias@udc.es

Margarita Alonso Ramos
UNIVERSIDADE DA CORUÑA
GRUPO LENGUA Y SOCIEDAD DE LA INFORMACIÓN
ESPAÑA
lxalonso@udc.es

Recibido: 03-V-2016 / **Aceptado:** 21-VII-2017

Resumen

Este artículo propone un método para nivelar las colocaciones del *Diccionario de Colocaciones del Español* de acuerdo con los niveles propuestos en el *MCER*. Como criterio nivelador se ha usado la frecuencia que las colocaciones del diccionario presentan en un corpus. Mediante el análisis de una muestra de colocaciones contenidas en el diccionario y en el *Plan Curricular del Instituto Cervantes* se comprueba una correlación negativa entre la nivelación propuesta para esas colocaciones en el *Plan Curricular* y la frecuencia de corpus. Tal correlación justifica la pertinencia de la frecuencia como criterio nivelador. Finalmente, se lleva a cabo un examen cualitativo de los resultados para comprobar la validez del método propuesto.

Palabras Clave: Colocaciones, diccionario, niveles de aprendizaje, MCER, frecuencia.

Abstract

This article puts forward a method to level the collocations of the *Diccionario de Colocaciones del Español* according to the levels proposed in the MCER. The frequency that the collocations contained in the dictionary exhibit in the corpus has been used as the levelling criteria. By means of the analysis of a sample of collocations included in both the dictionary and the *Plan Curricular del Instituto Cervantes,* it is proven a negative correlation between the levelling proposed for those collocations in the *Plan Curricular* and the corpus frequency. Such correlation justifies the relevance of the frequency as a levelling criterion. Finally, a qualitative exam of the results is carried out to prove the validity of the proposed method.

Key Words: Collocations, dictionary, learning levels, MCER, frequency.

INTRODUCCIÓN

El presente trabajo tiene como principal objetivo proponer un sistema para asignar los distintos niveles de aprendizaje establecidos en el *Marco Común Europeo de Referencia para las Lenguas* (Consejo de Europa, 2002, en adelante *MCER*) a las colocaciones presentes en el *Diccionario de Colocaciones del Español* (en adelante *DiCE;* Alonso Ramos, 2004a; Alonso Ramos, 2015, 2016 y las referencias allí incluidas para una descripción del diccionario), uno de los tres diccionarios combinatorios que existen para el español (véase Corpas Pastor, 2016, para una panorámica al respecto). El *DiCE* es un diccionario de correlatos léxicos que presta especial atención a las colocaciones, tal como las concibe la Lexicología explicativa y combinatoria (Mel'čuk, Clas & Polguère, 1995). Para esta corriente teórica, una 'colocación' es un sintagma compuesto por dos unidades léxicas, una de las cuales la 'base'–se elige por su significado, mientras que la elección de la otra –el 'colocativo'– viene determinada por la identidad léxica de la primera (Mel'čuk, 2012). Así, para predicar del nombre 'miedo' el significado 'causar', el propio nombre impone el verbo 'dar' y rechaza, por ejemplo, 'hacer', al contrario que 'ilusión' ('dar/*hacer miedo'; '??dar/hacer ilusión'). Por su propio contenido, pues, el diccionario es ya un instrumento valioso para el aprendiz de español, ya que puede serle de ayuda en la producción de colocaciones, aspecto más problemático para el hablante no nativo que la comprensión de dichas combinaciones (Nesselhauf, 2004). Además, el *DiCE* ha ido incorporando una serie de recursos dirigidos a incrementar la utilidad para sus posibles usuarios, en general, y de los aprendices de español en particular (Vincze & Alonso Ramos, 2013a), tales como la posibilidad de ordenar los colocativos de mayor a menor frecuencia (Alonso Ramos, 2012; Vincze & Alonso Ramos, 2013b), la presencia de marcas de uso (Vázquez Veiga, 2014), un módulo de actividades sobre colocaciones, etc.

Incorporar indicaciones de nivel al *DiCE* responde a la intención de ampliar su utilidad como herramienta de aprendizaje del español como lengua extranjera o lengua segunda (ELE/L2), tanto desde el punto de vista de los estudiantes como de los docentes. Existen diversos recursos que incluyen o proyectan incluir repertorios

léxicos –colocaciones incluidas– organizados de acuerdo con los criterios de nivelación del *MCER,* ya sea en formato de diccionarios o de listas de vocabulario (Capel, 2010, 2015; para el inglés y Spina, 2016, para el italiano). Hasta donde sabemos, el repertorio más similar del que dispone el español son las listas de exponentes[1] incluidas en diferentes secciones del *Plan curricular* (Instituto Cervantes, 1997-2016; en adelante *PCIC*). Este tipo de recursos tiene diversas aplicaciones, que van desde facilitar directrices o puntos de referencia en la elaboración de instrumentos didácticos (planes docentes, textos), hasta determinar patrones de orden de adquisición o criterios de evaluación (Capel, 2010). Dependiendo del propósito, las referencias para asignar niveles serán distintas. Así, el *English Vocabulary Profile* (Capel, 2010) añade datos procedentes de corpus de aprendices a la información de frecuencia obtenida a partir de corpus del inglés o listas de vocabulario para determinar cuál es el léxico que los hablantes no nativos son efectivamente capaces de producir en un determinado nivel. El *DICI-A,* por su parte, toma como punto de referencia un corpus de hablantes nativos (Spina, 2016) y para determinar el nivel de las colocaciones que incluye emplea un conjunto de parámetros: la frecuencia y la dispersión de la colocación en el corpus, su función (expresiones con significado descriptivo frente a marcas de organización textual y elementos pragmáticos) y el tema al que la colocación en cuestión se asocia.

En el caso particular del español, Ferrando Aramo (2012) lamentaba la ausencia de materiales que pudieran servir de guía para la secuenciación de colocaciones a lo largo de los distintos niveles de aprendizaje. Con la inclusión de indicaciones relativas al nivel de aprendizaje para cada colocación, el *DiCE* puede ser una herramienta interesante para subsanar esta y otras lagunas. Asimismo, el sistema para atribuir un determinado nivel de aprendizaje propuesto aquí podría aplicarse tanto a nuevos conjuntos de colocaciones que se vayan a incorporar al *DiCE,* como a otras listas de colocaciones que pretendan ordenarse de acuerdo con la nivelación del *MCER* (con las adaptaciones oportunas).

El trabajo se organiza del modo siguiente. Tras esta introducción, el siguiente apartado da cuenta del marco teórico que informa el *DiCE*. A continuación se expone el método usado para asignar los niveles del *MCER* a las colocaciones del diccionario tomando como referencia las colocaciones del *DiCE* que incluye el *PCIC*, de una parte, y su frecuencia en un extenso corpus de español peninsular, de otra. Tras ello se presentan los resultados y se discute la adecuación del método propuesto. El artículo se cierra con un apartado de conclusiones.

1. Marco teórico: La Lexicología explicativa y combinatoria

Tal como se explica en Alonso Ramos (2015), el *DiCE* es un diccionario que sigue los presupuestos teóricos defendidos por la Lexicología explicativa y combinatoria

(Mel'čuk et al., 1995). Aunque presta especial atención a las colocaciones del español (relaciones sintagmáticas), propiamente se trata de un diccionario de 'correlatos léxicos', esto es, de relaciones tanto sintagmáticas como paradigmáticas en las que una determinada unidad léxica condiciona la elección de otra y que Mel'čuk (1996) formalizó en su día mediante la herramienta de las 'funciones léxicas'. Las funciones léxicas expresan relaciones relativamente sistemáticas dentro del léxico de las lenguas naturales y, cuando toman como argumento una unidad léxica (denominada 'palabra llave'), dan como resultado otra cuya elección depende de la primera. El funcionamiento de esta herramienta se puede comprobar en el ejemplo siguiente:

(1) **Able$_2$**(admiración) = admirable, digno de admiración

La función léxica devuelve adjetivos que expresan el significado 'ser objeto de' para una determinada palabra llave. En este caso en concreto, uno de los resultados ilustra un correlato léxico de tipo paradigmático ('admirable') y el otro uno sintagmático ('digno de admiración').

Según nuestro marco teórico, 'digno de admiración' es una colocación en tanto que la elección de uno de sus componentes se realiza en función de la identidad léxica del otro. Esto se hace especialmente evidente en la expresión de un significado como 'hacer, efectuar' en compañía de nombres como 'paseo' o 'excursión'. La expresión de dicho significado está restringida por la identidad léxica del nombre en cuestión y, en el primer caso, es 'dar' ('dar/*hacer un paseo'), mientras que en el segundo es 'hacer' ('*dar/hacer una excursión'), a pesar de las similitudes que los dos nombres puedan presentar si se atiende exclusivamente a su semántica ('tipo de desplazamiento, etc.'). Dentro de una colocación, denominaremos 'base' a la unidad que condiciona la elección de la otra y 'colocativo' a la unidad condicionada.

Cabe señalar además que, dentro del marco teórico adoptado, se consideran colocaciones únicamente las combinaciones de tipo composicional, es decir, cuyo significado se puede segmentar en partes correspondientes a cada uno de sus miembros (p. ej.: 'dar' = 'hacer, efectuar'; 'paseo' = 'tipo de desplazamiento, etc.'). Ahora bien, el colocativo puede ser una unidad léxica que solo aparece en el contexto de la colocación: p. ej., 'de leche' con el sentido de 'provisional, previo a la dentición definitiva', solo se encuentra en el contexto de la colocación 'diente de leche'. A pesar de la posible falta de transparencia del sintagma 'de leche', la combinación no deja de ser composicional: 'diente'= 'pieza ósea de la boca' + 'de leche'='provisional, previo a la dentición definitiva'.

Asimismo, como se desprende de lo anterior, las colocaciones se conciben como relaciones binarias asimétricas en las que la base guía la elección del colocativo (por razones de espacio, remitimos a Mel'čuk et al. (1995) para más detalles sobre la Lexicología explicativa y combinatoria).

La estructura del *DiCE* se deriva de los presupuestos teóricos expuestos. Los correlatos léxicos aparecen en la entrada de la unidad léxica que funciona como palabra llave (o base, en el caso de las colocaciones). Es, pues, un diccionario orientado eminentemente a la producción, toda vez que los resultados de las búsquedas devuelven, o bien colocativos de la unidad léxica buscada, o bien otro tipo de correlatos léxicos. Mediante el modo de 'Consulta general', agrupadas bajo cada lema el usuario encuentra las distintas unidades léxicas que le corresponden acompañadas de una breve glosa, un ejemplo de uso y enlaces a sus colocaciones y correlatos léxicos clasificados según su esquema sintáctico. Además se brinda información de la frecuencia de cada unidad léxica. Los colocativos (y demás correlatos léxicos) de una unidad léxica se presentan, nuevamente, con glosas breves (paráfrasis en lengua natural de la función léxica correspondiente) y ejemplos de uso.

Si bien, como se ha dicho, el *DiCE* está fundamentalmente orientado a la producción, su formato electrónico permite diversas maneras de acceso a la información que contiene y, mediante las opciones de búsqueda avanzada, se pueden hacer consultas partiendo de un determinado colocativo, del significado que este predica de su base (información que proporcionan las funciones léxicas) o de toda la colocación para obtener su glosa.

2. Propuesta metodológica para asignar niveles de aprendizaje a las colocaciones

La presente propuesta parte de un repertorio colocacional ya nivelado –el que se incluye en varias secciones del *PCIC*– e intenta adaptar este tipo de nivelación a un repertorio de colocaciones más amplio –el contenido en el *DiCE*–. Para trasladar los niveles de un repertorio a otro nos basaremos en la frecuencia de corpus, teniendo en cuenta que la mayoría de los criterios esgrimidos en el *MCER* y el *PCIC* tienen algún tipo de relación o repercuten de algún modo en esta dimensión, tal como se justifica más abajo.

2.1. Obtención de datos

Los datos que se han utilizado en este estudio provienen, por una parte, del *PCIC* y, por otra, del propio *DiCE*. El *PCIC* es un documento que se presenta como un 'repertorio de material' para los profesionales de la enseñanza de ELE (*PCIC*: Introducción general). Parte de un enfoque nocio-funcional y, en ciertos aspectos, supone una concreción con respecto a las directrices del *MCER*. El vocabulario, que es el centro de interés del presente estudio, aparece relacionado fundamentalmente en las secciones dedicadas al componente nocional (en el propio documento se indica que las 'nociones' son "contenidos que tienen que ver, en sentido amplio, con el significado"; *PCIC*: Nociones generales) y se organiza en diferentes niveles de aprendizaje. En este sentido, cabe apuntar que este documento se ha tomado ya como

referencia en algún otro estudio de nivelación de colocaciones para ELE (Laya Gómez, 2014; Rojo Mejuto, 2015).

Del *PCIC* nos hemos basado fundamentalmente en la sección 'Nociones específicas', por ser la que más coincidencias podría presentar con el *DiCE*. La sección citada contiene un repertorio léxico que abarca desde unidades léxicas consideradas de forma aislada (p. ej. 'pelo', 'ojo', 'nariz' en el nivel A1) a secuencias pluriverbales de distintos tipos (colocaciones como 'cometer un error', locuciones como 'hacer buenas migas', proverbios como 'hecha la ley, hecha la trampa', etc.). Del vocabulario de los apartados mencionados extrajimos las colocaciones que figurasen también en el *DiCE*. La muestra (n=89) contiene, por una parte, información relativa al nivel de aprendizaje (tomada del *PCIC*) e información relativa a la frecuencia de uso de cada colocación (contenida ya en el *DiCE*).

La información de frecuencia proviene de la sección correspondiente al español peninsular del corpus *esTenTen11,* que contiene más de 2000 millones de ocurrencias (Kilgarriff & Renau, 2013). La elección de este corpus obedece principalmente a las posibilidades que en cuanto a la recuperación de colocaciones ofrece su interfaz de consulta, ausentes en otros corpus de referencia del español. Así, cada colocación se ha buscado aprovechando la anotación morfológica del corpus para encontrar ciertas configuraciones sintácticas concretas. En un ejemplo como la colocación formada por los lemas 'miedo' y 'tener', este enfoque permite incluir tanto casos donde los constituyentes de la colocación no forman una secuencia continua (es decir, casos donde entre verbo y nombre aparecen modificadores u otros complementos), como los de posible anteposición (secuencias del tipo 'el miedo que... tenía'). Para ello se utilizaron en las consultas al corpus una serie de reglas que combinan el etiquetado morfológico con la lematización de las unidades buscadas (Vincze & Alonso Ramos, 2013b). La interfaz permite además automatizar este proceso en gran medida mediante el uso de APIs. De otro modo, la obtención de información de frecuencia para todas las colocaciones incluidas en el *DiCE* hubiese sido una tarea imposible.

En contrapartida a sus posibilidades de consulta, el corpus elegido tiene ciertas limitaciones. Quizá la principal para el propósito de la investigación que aquí se presenta sea que no está divido en secciones representativas de distintos géneros discursivos, lo cual impide medir la dispersión de una determinada colocación según este parámetro. Esto puede constituir un problema, en la medida en que la dispersión de una determinada forma permite asegurar que dicha forma no es frecuente exclusivamente en un género textual (Juilland & Chang-Rodríguez, 1964; Alvar Ezquerra, 2004). Esta distribución desigual a través de distintos géneros textuales afecta también a las colocaciones (Corpas Pastor, 2015). Una alternativa sería medir la dispersión de las colocaciones teniendo en cuenta, no secciones temáticas o genéricas, sino la propia división en documentos del corpus (con coeficientes como DP, propuesto por Gries (2008) o TF-IDF (Spärck, 1972)). No obstante, debido al gran

volumen de documentos del propio corpus y a que su consulta solo es posible mediante una interfaz, la implementación de esta solución no es trivial y tendrá que esperar a investigaciones futuras. Con todo, parece razonable asumir que en un corpus de gran tamaño como el utilizado, la sobrerrepresentación de formas específicas de ciertos géneros se diluya en gran medida.

Además, el corpus incluye solamente documentos extraídos de la *Web*, por lo que no están propiamente representadas las variedades orales del español. Aun así, entre los incluidos, están presentes textos con un grado de planificación mínima (entrevistas, foros, ciertos blogs, etc.), más próximos por ello a textos orales.

Cabe notar asimismo que para el estudio descrito se ha usado exclusivamente la parte del corpus correspondiente al español de España. Esto se debe a un intento por mantener cierta coherencia con el actual estado del *DiCE*, toda vez que desde las fases iniciales de su elaboración se contó con las intuiciones lingüísticas de lexicógrafos hablantes de esta variedad y los ejemplos estaban tomados de corpus en los que esta variedad era también predominante.

2.2. La frecuencia y su papel en la gradación del vocabulario

La aplicación de estudios de frecuencia a la enseñanza del vocabulario de una L2/LE tiene una historia relativamente larga. El primer listado de léxico frecuente en español del que tenemos noticia es el de Keniston (1920). En el ámbito francófono, la idea de que en la enseñanza de una lengua ha de primarse el léxico más frecuente (a lo que posteriormente se añadió el criterio de la disponibilidad léxica) tuvo gran pujanza a mediados del siglo XX y culminó en el proyecto del *français fondamental* (Christ & Christ, 1951; Gougenheim, Michéa, Rivenc & Sauvageot, 1964). Este interés por la frecuencia léxica –y por el léxico, en general– decreció considerablemente con la irrupción del paradigma generativista. El desplazamiento de este último por el enfoque nocio-funcional, que sigue condicionando la planificación de la enseñanza de lenguas de manera notable (está presente, por ejemplo, tanto en el *MCER* como en el *PCIC*), no redundó, sin embargo, en una concreción en cuanto a los contenidos léxicos y su gradación. Así, a principios de la década de 1980, Hindmarsh advertía de que tener en cuenta nociones y funciones comunicativas en la enseñanza de una lengua no implica necesariamente una concreción de los contenidos lingüísticos pertinentes en las sucesivas fases de su aprendizaje y defendía que tal concreción es útil y, hasta cierto punto, necesaria (Hindmarsh, 1980). En este sentido es ilustrativo el caso del *PCIC*, donde, las nociones y funciones contribuyen solo parcialmente a la concreción de contenidos y vuelve a invocarse el criterio de la frecuencia léxica para la gradación del vocabulario.

Probablemente, el desarrollo de la lingüística de corpus haya sido mucho más decisivo en el renovado interés por el vocabulario que se observa en las últimas

décadas del siglo XX (Sinclair & Renouf, 1985; Lewis, 1993) que la reacción a diversas corrientes estructuralistas (incluido el generativismo) que supuso el enfoque nocio-funcional. Este renovado interés por el vocabulario viene acompañado casi siempre por la vuelta a los estudios de frecuencia léxica, en tanto que este parámetro supone la posibilidad de medir de forma relativamente objetiva la rentabilidad del léxico (Sinclair & Renouf, 1985; Nation, 2001) o de las combinaciones léxicas (Martínez, 2013). La lingüística hispánica no es una excepción al respecto. Por ejemplo, Alvar Ezquerra (2004) hace una encendida defensa del uso de diccionarios de frecuencia o listas extraídas de corpus para determinar cuál es el vocabulario que con más urgencia necesita un aprendiz de español, hasta el punto de proponer que sea la frecuencia lo que determine a qué formas de un paradigma ha de prestarse atención preferente en la enseñanza (Alvar Ezquerra, 2004). Esta defensa de la frecuencia como determinante de la rentabilidad de los elementos lingüísticos hasta sus últimas consecuencias es ciertamente llamativa, ya que otros autores ven un inconveniente en subordinar la presentación de un paradigma a la frecuencia de sus distintos elementos, por la posible incoherencia del resultado.

Además de la frecuencia, se citan otros criterios con respecto a la selección del léxico. El primero de ellos es la dispersión en el corpus que se consulte (Juilland & Chang-Rodríguez, 1964; Alvar Ezquerra, 2004): una dispersión alta es señal de que la forma que la presenta forma parte del léxico común y no es frecuente solo en un tipo de discurso. Como indicamos más arriba, el corpus utilizado no nos permitía medir este parámetro. No obstante, confiamos en que la gran extensión del corpus compense esta deficiencia en cierta medida: es decir, con una muestra tan grande, es de esperar que los sesgos introducidos por textos de un determinado tipo tengan un impacto menor.

Junto con la frecuencia, Gómez Molina (2004) cita como criterios útiles para la gradación del léxico en la enseñanza de una lengua su productividad y los intereses de los aprendices. Según el autor, la productividad agruparía la cobertura (*coverage*) de Nation (2001) y el rango, que, en Lingüística de corpus, no es sino una medida de dispersión. En cuanto a la cobertura, cabe notar que es una dimensión íntimamente ligada con la frecuencia, en tanto que son las formas más frecuentes, por el propio hecho de serlo, las que cubren un mayor porcentaje de las palabras corrientes de un texto (Zipf, 1935; Alvar Ezquerra, 2004). Por lo que respecta a los intereses de los aprendices, cabe notar que seguramente es un criterio muy eficaz para planificar los contenidos de un curso con un grupo de alumnos concreto, pero no es aplicable a una obra de consulta dirigida a un público general, como un diccionario.

Además de las limitaciones apuntadas en cuanto al corpus concreto que se ha usado para el estudio, ciertos autores han señalado las limitaciones que presenta cualquier corpus a la hora de establecer la frecuencia de un determinado elemento léxico. En primer lugar, se ha observado que la información contenida en un corpus

no coincide exactamente con el input de ningún hablante concreto (Hoey, 2005; Schmitt, 2010). En segundo lugar, se ha llamado la atención sobre la existencia de discrepancias entre corpus distintos en cuanto a la frecuencia que presenta una determinada forma (McGee, 2008)[2]. Una alternativa propuesta por McGee (2008) sería combinar la información de frecuencia léxica extraída de un corpus con las intuiciones de los hablantes. En esta línea parece moverse la reciente propuesta de Benigno, Kraiff, Grossmann y Velez (2016). Este estudio retoma la noción de léxico fundamental y pretende establecer un método para encontrar un repertorio de colocaciones fundamentales en francés. Los autores presentan a hablantes nativos diversas listas de combinaciones léxicas obtenidas por medio de diversas medidas de asociación y filtradas estableciendo un valor mínimo de dispersión. A los informantes se les pregunta qué combinaciones tienen para ellos un carácter 'fundamental'. Es interesante apuntar que la medida de asociación que muestra un mayor grado de correspondencia con las intuiciones de los informantes es la frecuencia de coaparición en un corpus.

En resumen, aun considerando la posibilidad de complementarlo con otro tipo de información en el futuro (frecuencia percibida, áreas temáticas propias de los distintos niveles de aprendizaje, etc.), la frecuencia léxica parece un criterio de partida lo suficientemente sólido para establecer una gradación de las colocaciones del *DiCE* en distintos niveles de aprendizaje.

Al aplicar el criterio de la frecuencia, cabe plantearse, además, si, al tratar con colocaciones, tal como las define el marco teórico adoptado, esto es, unidades léxicas en una relación asimétrica, debería considerarse este parámetro para todo el conjunto o dar de algún modo más relevancia a la frecuencia de la base. El segundo tipo de razonamiento se defendía en Alonso Ramos (2012) y se aplicó a una clasificación inicial de las colocaciones del *DiCE* en franjas de frecuencia. Una revisión inicial de muestras tomadas del *DiCE* desaconsejó, sin embargo, esta manera de proceder en la presente investigación. Así, partiendo de una frecuencia ponderada, calculada tal como se propone en Vincze y Alonso Ramos (2013b), dando más peso a la frecuencia de la base, obtendríamos que todas las colocaciones de 'miedo' incluidas en el diccionario recibirían un nivel B1 o inferior al aplicar el sistema que se propone más abajo. El conjunto incluye casos como 'miedo cerval' o 'cagarse de miedo', ambas pertenecientes a registros marcados (literario uno, coloquial el otro). Este resultado no se adapta a las directrices de *MCER* y *PCIC*, de acuerdo con las cuales, las expresiones marcadas en cuanto al registro se irían presentando a partir del nivel B2. Partiendo de la frecuencia sin ponderación alguna, las colocaciones de 'miedo' se reparten de modo más uniforme a través del de abanico de niveles posibles, desde 'tener miedo' (A1/A2), hasta 'miedo cerval' (C2), lo cual parece más deseable (*vid. infra*).

2.3. Criterios de nivelación usados en el PCIC y su aplicación al DiCE

Una vez consideradas la utilidad y limitaciones de la frecuencia y, en concreto, la frecuencia obtenida a partir de corpus, para la gradación del léxico en la enseñanza de lenguas, repasaremos los criterios utilizados en el *PCIC*, documento del que partimos para establecer la gradación del *DiCE*. Los criterios que maneja el *PCIC* son los siguientes (cf. introducciones a los capítulos 5 y 9 del citado documento):

i) registro, estrato, etc.
ii) rentabilidad comunicativa
iii) frecuencia léxica, establecida a partir de la introspección de los autores
iv) ser una "frase hecha" o "expresión idiomática"

Es razonable pensar que i) y ii) repercuten en iii). Así, cabe suponer que las expresiones no marcadas en cuanto al registro aparecerán en cualquier tipo de texto y serán más frecuentes que las marcadas (Schmitt, 2010). Igualmente, salvo casos de discrepancia[3], parece que, en general, una expresión comunicativamente rentable lo es en virtud de su aplicabilidad a una gran variedad de contextos y situaciones. Teniendo en cuenta que las expresiones comunicativamente más rentables y neutras son las asociadas a niveles más bajos, sería esperable una correlación negativa entre frecuencia léxica y nivel.

Cabe notar, con todo, que si bien el tercer criterio utilizado en la nivelación es propiamente el de la frecuencia léxica, en el *PCIC* esta se ha establecido a partir de la introspección de los autores[4] y, por tanto, en teoría, no tendría por qué coincidir con la frecuencia extraída a partir de un corpus concreto. Existen, sin embargo, una serie de estudios que han venido realizándose desde la década de 1960 y que sugieren que las intuiciones sobre la frecuencia léxica de los hablantes nativos coinciden hasta cierto punto con los datos de frecuencia extraídos de corpus (McGee, 2006; Siyanova & Schmitt, 2008; y sobre colocaciones en particular, Siyanova & Spina, 2015)[5], si bien existe una variabilidad considerable en cuanto al grado de intuición de cada individuo (Schmitt, 2010).

Teniendo en cuenta lo anterior, sería de esperar que el nivel atribuido a las colocaciones que figuran en el *PCIC* muestre algún grado de correlación con la frecuencia con las que estas aparecen en un corpus del español. Para verificar esta expectativa hemos comparado la frecuencia de nuestra muestra de colocaciones en el corpus *esTenTen11* de español peninsular y los niveles que se les atribuyen en el *PCIC*. La muestra incluye solamente colocaciones de cuatro de los niveles del *MCER* (B1, B2, C1 y C2), ya que no se han encontrado colocaciones del *DiCE* en el nivel A1 y hemos estimado que el número de colocaciones encontradas con nivel A2 era demasiado exiguo (solo cuatro) como para ser tenido en cuenta. La estadística descriptiva de la muestra manejada se presenta en la Tabla 1.

Tabla 1. Datos relativos a la frecuencia de las colocaciones del *DiCE* que se registran en el *PCIC*.

nivel	n	media	desv. est.	mediana	min.	1er cuart.	3er cuart	máx.
B1	10	10915,4	10186,48	7672	1877	2259	17740	31664
B2	18	6134,72	10755,01	1439	102	746	5957	39909
C1	11	2368,82	4027,42	116	10	48,5	2361	10538
C2	46	1922,48	11156,77	31	0	9	167	75770

Para determinar si existe una correlación entre el nivel de las colocaciones de nuestra muestra y su frecuencia en el corpus manejado se ha calculado el coeficiente τ_b de Kendall. Este coeficiente mide la correlación entre dos rangos: un valor de 1 indicaría una correlación positiva perfecta, un valor de 0, falta de correlación alguna y un valor de -1 una correlación negativa perfecta, esto es, que a valores altos en una dimensión se corresponden valores bajos en la otra. Además, este coeficiente de correlación tiene en cuenta posibles empates, que se dan necesariamente aquí, puesto que hay cuatro valores posibles (B1, B2, C1 y C2) para 85 casos. Lo esperable sería una correlación negativa entre nivel y frecuencia: es decir, que a niveles más altos correspondiesen frecuencias más bajas y viceversa. Esta expectativa se confirma con una correlación negativa moderada (τ_b=–0,57) estadísticamente significativa ($p<0,0001$). Parece, pues, justificado asumir que de acuerdo con los criterios del *PCIC*, en general, a medida que descienda la frecuencia de una colocación, más alto será su nivel.

Una vez establecido que la frecuencia de corpus tiene cierta correlación con la nivelación de las colocaciones registradas en el *PCIC*, queda por precisar el procedimiento para asignar nivel de las colocaciones del *DiCE*. Más en concreto, decidir a partir de qué frecuencia a una colocación dada se le asignará un nivel u otro. Para tal fin, hemos decidido usar como punto de corte el valor coincidente con el primer cuartil de cada nivel, antes que medidas más extremas como la frecuencia máxima[6] o la frecuencia mínima. Este valor actúa como límite inferior. Por ejemplo, el valor del primer cuartil del nivel B2 es la frecuencia de 'perder las ganas' (746 ocurrencias). Así, todas las colocaciones con una frecuencia inferior a esta e igual o superior al primer cuartil del siguiente nivel (C1) se han clasificado como C1. Puesto que, debido a su escasez en la muestra, las colocaciones de nivel A2 no se han tenido en cuenta, para fijar el límite inferior del nivel B1 se ha usado el valor coincidente con el tercer cuartil del propio B1.

3. Resultado de la aplicación al *DiCE*, coincidencia con el *PCIC* y adecuación

En la presente sección se aborda el análisis de los resultados obtenidos. En primer lugar, mediante un análisis cuantitativo se da cuenta de la coincidencia entre la nivelación de la muestra extraída del *PCIC* y la que resulta de aplicar el método

propuesto. A continuación se analiza de manera cualitativa el resultado de la nivelación en el nivel A1-A2 de forma más o menos exhaustiva. Ya que el número de colocaciones incluidas en los niveles superiores no permite un análisis con el mismo grado de detalle, se analizan a continuación los resultados para colocaciones que incluyen dos bases especialmente productivas ('miedo' y 'amor'), que ofrecen un panorama ilustrativo del funcionamiento del método propuesto a través de los distintos niveles del *MCER*. Por último, se considera qué solución sería más conveniente en aquellos casos en los que el criterio de la frecuencia léxica parece chocar con la supuesta idiomaticidad de las combinaciones consideradas.

3.1. Coincidencia entre la nivelación del DiCE y el PCIC

Revisemos en primer lugar la correspondencia que se da entre los niveles que asigna el *PCIC* a las colocaciones de la muestra y los que resultan de aplicar únicamente el criterio de frecuencia de corpus. Para verificar si la decisión de tomar como referencia el valor del primer cuartil de cada nivel es adecuada, compararemos las nivelaciones que resultarían de aplicar la frecuencia mínima de cada nivel o la frecuencia del primer cuartil como punto de corte. Los datos se exponen en la Tabla 2.

Tabla 2. Coincidencia entre nivelación del PCIC y nivelación basada en frecuencia.

	punto de corte = 1er cuartil		**punto de corte = frecuencia mínima**	
	nº	**%**	**nº**	**%**
mismo nivel	45	52,27	36	40,1
PCIC+1	10	10,23	3	3,41
PCIC–1	18	20,45	32	36,37
PCIC+2	12	13,64	15	17,05
PCIC–2	1	1,14	0	0
PCIC–3	3	3,4	3	3,4
Total	89		89	

Cabe notar, en primer lugar, que una nivelación basada en la frecuencia de corpus coincide de forma notable con la que se presenta en el *PCIC*. Así, tanto si tomamos como punto de corte la frecuencia mínima o el valor del primer cuartil, el porcentaje de las colocaciones que o bien coinciden con el nivel que se les adjudica en el *PCIC* o bien aparecen en un nivel inmediatamente adyacente ronda el 80% de la muestra en ambos casos. El número de coincidencias entre la nivelación propuesta es llamativo en cualquier caso, ya que las colocaciones que forzosamente conservan su nivel del *PCIC* son cuatro: aquellas cuyas frecuencias coinciden con los puntos de corte utilizados. La nivelación que resulta de tomar el primer cuartil como punto de corte, además, presenta un porcentaje mayor de coincidencias al mismo nivel (algo más de la mitad) que la resultante de tomar como referencia la frecuencia mínima. Parece, pues, que optar por el primero ha sido una decisión acertada.

Se apuntaba a propósito de la muestra manejada que las colocaciones presentes en el *DiCE* incluidas en niveles bajos del *PCIC* son escasas. Teniendo en cuenta los objetivos propuestos en este documento para los niveles A1 y A2 y la composición actual del *DiCE* (colocaciones de nombres de sentimiento), este hecho puede resultar esperable. Así, en la descripción general de los mencionados niveles queda de manifiesto que los recursos comunicativos que se esperan de los aprendices estarán dirigidos fundamentalmente a cubrir necesidades materiales básicas, como saber su ubicación, desenvolverse en transacciones comerciales, etc., funciones comunicativas que no pasan necesariamente por la expresión de los propios sentimientos (véase http://cvc.cervantes.es/Ensenanza/biblioteca_ele/plan_curricular/niveles/01_objetivos_relacion_a1-a2.htm).

Ahora bien, el simple hecho de que la nivelación propuesta esté basada en la frecuencia de corpus en sí mismo favorece una distribución desigual de las colocaciones entre los diferentes niveles. Hace tiempo que se ha observado en cuanto a la distribución del léxico –por ejemplo, Zipf (1935) – que existe un conjunto reducido de formas de palabra que presentan una frecuencia elevada en cualquier texto, mientras que el conjunto de formas que aparecen de forma esporádica es mucho más amplio. Las colocaciones no parecen ser una excepción a esta tendencia, y teniendo en cuenta que la asignación de niveles se ha hecho de forma inversamente proporcional a la frecuencia registrada para cada colocación, es esperable encontrar pequeños grupos de colocaciones muy frecuentes en niveles bajos y vastos repertorios de colocaciones infrecuentes en los más altos. Este es precisamente el resultado de examinar la distribución de las colocaciones del *DiCE* agrupadas según los distintos niveles que hemos establecido. La mayoría de las más de once mil colocaciones niveladas mediante el sistema propuesto aparecen en el nivel C2 (9.574 colocaciones). El resto de colocaciones se distribuyen como sigue: C1, 951 colocaciones; B2, 637 colocaciones; B1, 393 colocaciones; y finalmente un nivel único para A1-A2 con 31 colocaciones.

3.2. Colocaciones de nivel A1-A2

Como se veía más arriba, las colocaciones marcadas con el nivel A1-A2 forman el grupo menos numeroso una vez aplicado el sistema propuesto. El conjunto se reduce aun más en la versión pública del diccionario, pues en casos de discrepancia entre los resultados de nuestra propuesta y el *PCIC,* el *DiCE* conserva el nivel asignado en el *PCIC*. Repasemos alguna de estas discrepancias. 'Tener gana(s)', por ejemplo, según la frecuencia con que se documenta en el corpus debería aparecer en el nivel inicial. En el *PCIC* aparece, sin embargo, en un nivel más alto (B1). Por su frecuencia también le correspondería el nivel A1-A2 a 'llamar la atención', que aparece en el *PCIC* en un nivel mucho más alto (C1). Teniendo en cuenta, por un lado, la elevada frecuencia de uso de esta colocación y, por otra, el hecho de que no parece que sea propia de un

registro marcado, cabría pensar que la decisión del *PCIC* responde a uno de las criterios de nivelación esgrimidos tanto en el *PCIC* como en el *MCER*: asociar el dominio de lo que denomina 'usos' o 'expresiones idiomáticas' con los niveles C1 y C2 (*MCER*: 119). Desde el enfoque adoptado en el *DiCE*, la decisión de considerar 'llamar la atención' como una colocación y, por tanto, incluirla en el diccionario, está totalmente justificada, ya que se trata de un sintagma plenamente composicional (es posible segmentar el significado de la expresión entre sus dos constituyentes: 'llamar', 'atraer, causar que X se dirija a Y', y 'atención', 'acción de dedicar los sentidos, la actividad mental, etc. a algo'). Ahora bien, este sentido de 'llamar' no parece especialmente productivo fuera del contexto de 'atención', con la excepción quizá del derivado 'llamativo'. Este hecho puede haber llevado a los autores del *PCIC* a considerar 'llamar la atención' un tipo de expresión idiomática –a pesar de su composicionalidad– y a reservarla intencionadamente para un nivel alto. Un caso hasta cierto punto similar es el de 'prestar atención', también muy frecuente en el corpus –según el sistema propuesto recibiría el nivel A1-A2– que en el *PCIC* aparece en el nivel B2.

Si en los casos de 'llamar la atención' y 'prestar atención' podría argumentarse tanto una marca de nivel alto –debido a la relativa falta de transparencia de 'llamar' en ese contexto–, como una marca de nivel relativamente bajo –por su alta frecuencia y, por tanto, por su rentabilidad comunicativa–, la inclusión de otras colocaciones del *DiCE* en A1-A2 podría resultar más cuestionable. Así, las colocaciones 'tipo de interés' o 'interés general' probablemente deben su alta frecuencia a la fuerte presencia de textos periodísticos en el corpus consultado. En tales casos es razonable cuestionar que frecuencia y rentabilidad comunicativa vayan de la mano. Intuitivamente, se diría que las colocaciones en cuestión están restringidas a textos relativamente especializados en economía y política. Para confirmar la validez de esta intuición, hemos acudido al CORPES XXI (Real Academia Española, s.f.), que permite delimitar las búsquedas por la temática del documento donde se atestiguan estas combinaciones. Como era de esperar, estas colocaciones se concentran en los textos políticos y económicos, con casi 3 ('interés general') y 9 ('tipo de interés') casos por millón. En el resto de áreas temáticas no alcanzan nunca una ocurrencia por millón. Este resultado sugiere además la conveniencia de tener en cuenta la dispersión al lado de la frecuencia como criterio nivelador y aconseja incorporarla en futuras revisiones del *DiCE*.

De lo anterior cabe concluir que, a pesar de que el conjunto de colocaciones incluidas en el nivel A1-A2 es pequeño, podría decirse que la nivelación basada en la frecuencia peca por exceso más que por defecto. Es decir, probablemente a una parte de las colocaciones que, según su frecuencia, les correspondería nuestro nivel A1-A2 deberían aparecer en niveles más altos según los criterios del *PCIC*.

3.3. La gradación de las colocaciones de las bases 'miedo' y 'amor'

Mientras que el conjunto de colocaciones incluidas en el nivel más bajo, gracias a su reducido tamaño, permite una revisión cualitativa relativamente exhaustiva, con el resto de niveles esto no sucede. Por esta razón, el enfoque adoptado será distinto: limitaremos la revisión de las colocaciones correspondientes a dos bases que presentan ocurrencias en prácticamente todos los niveles distinguidos, 'miedo' y 'amor'.

Según el sistema de nivelación propuesto, 'miedo' presentaría dos colocaciones en el nivel A1-A2: 'tener' y 'dar miedo'. La coincidencia con el *PCIC* es parcial: 'tener miedo' recibe en este documento un A2, pero 'dar miedo' aparece en el nivel B1 en compañía de las colocaciones 'dar pena' y 'dar lástima'. Seguirían en un nivel más alto (B1) sinónimos menos frecuentes de los colocativos ya presentados en el nivel A1-A2: los verbos de apoyo 'pasar' y 'sentir' y el verbo causativo 'meter'. En cuanto a los verbos de apoyo[7], la progresión parece adecuada. Si se trata de priorizar el registro neutro, intuitivamente 'tener miedo' parece menos marcado que 'pasar' o 'sentir miedo' –la última colocación, quizá, restringida a discursos de cierta formalidad–. Más discutible podría ser la opción de marcar 'meter miedo' con un B1 si nos guiamos por las recomendaciones en cuanto al registro incluidas en el *PCIC* y el *MCER*. En primer lugar, 'meter' como colocativo causativo parece estar restringido fundamentalmente a la lengua oral, como revela un estudio de corpus llevado a cabo por Alba-Salas (2009). En segundo lugar, el hecho de que 'meter' como colocativo causativo no sea especialmente productivo (está prácticamente limitado a los nombres 'miedo', 'prisa', y quizá 'ganas') puede hacer que tales colocaciones se vean como ejemplos de 'frases hechas'. En ambos casos, los documentos citados recomiendan de un nivel B2 en adelante.

Conforme va subiendo el nivel, se presentan colocativos de 'miedo' que parecen cada vez más propios o bien de discurso literario –o en todo caso, de un grado de formalidad y planificación considerable–, o bien restringidos a registros muy coloquiales. Así, mientras 'perder el miedo' aparece en B1, colocativos como 'vencer' o 'superar', con un significado similar, pero intuitivamente restringidos a registros más formales aparecen en un nivel más alto (B2). En C1 aparece una serie de verbos causativos ('causar', 'sembrar', 'infundir miedo') que se esperarían en textos con cierta vocación de estilo, así como adjetivos intensificadores propios de hablantes con un considerable dominio del léxico ('miedo intenso', 'atroz'). Por último, en el nivel C2 encontramos colocativos como 'cerval', que en español contemporáneo aparece casi exclusivamente en el contexto de 'miedo' y, probablemente, en textos literarios, o 'cagarse de', que sería esperable casi exclusivamente en registros muy coloquiales.

En el caso de 'amor' la situación es relativamente similar. Las colocaciones que se presentan en los niveles más bajos presentan significados relativamente genéricos y

podrían considerarse de un registro neutro (B1: 'sentir amor', 'dar amor', 'gran amor'…). A medida que se va subiendo aparecen significados más específicos (B2: 'amor a primera vista', 'amor platónico'…). Por último, en los niveles superiores (C1, C2) aparecen colocaciones que exigen el dominio de un repertorio léxico considerable y que parecen propias de textos con un alto grado de planificación o de estilo muy cuidado ('profesar amor', 'amor acendrado', etc.).

3.4. La frecuencia de uso y la idiomaticidad

Como se ha visto, en ciertas ocasiones se producen discrepancias entre la nivelación que resulta de la frecuencia de corpus y la que se propone en el *PCIC* y en tales casos el *DiCE* conserva el criterio establecido por aquel. Podría, sin embargo, cuestionarse qué debe primar cuando los criterios de nivelación previsiblemente relacionados con la frecuencia (frecuencia percibida, registro neutro y rentabilidad) parecen ir en dirección contraria al de la 'idiomaticidad', como sospechamos sucede en el caso de 'llamar la atención', que a pesar de ser una combinación frecuente y aparentemente neutra en cuanto al registro no aparece hasta el nivel C1. El *MCER* y el *PCIC* recomiendan dejar 'expresiones idiomáticas', 'frases hechas', etc., para niveles altos, pero esta recomendación podría no ser adecuada, especialmente cuando la expresión que se nivela tiene una frecuencia alta. Así, en un reciente trabajo, Martínez (2013) recomienda dar prioridad en la enseñanza de una lengua extranjera a las expresiones pluriverbales a la vez frecuentes y poco transparentes.

Como se apunta más arriba, podría considerarse que ciertas colocaciones, a pesar de su carácter composicional, son relativamente opacas, en la medida en que el colocativo o bien ocurre casi únicamente en el contexto de la colocación en cuestión (p. ej. 'miedo cerval') o bien transmite un significado que fuera de esa colocación no se asocia normalmente a su significante (p. ej. 'llamar la atención'). Si, como creemos, este argumento se usa en el *PCIC* para determinar si una expresión es idiomática (locución, frase hecha, etc.), dejar todas las expresiones de este tipo para niveles avanzados puede ser contraproducente. Si atendemos a su rentabilidad comunicativa, es evidente que expresiones del tipo de 'miedo cerval', por un lado, y 'llamar la atención', por otro, no son equivalentes: la segunda es mucho más rentable, a juzgar por su frecuencia de uso.

CONCLUSIONES

En este trabajo se ha propuesto un sistema para organizar las colocaciones del *DiCE* de acuerdo con los niveles del *MCER*. El sistema se basa, por un lado, en una muestra de colocaciones niveladas por profesionales de la enseñanza de español y, por otro, en la frecuencia con la que se registran las colocaciones en un corpus. Teniendo en cuenta que los criterios empleados por los autores de la nivelación de la muestra tienen previsiblemente relación con la frecuencia léxica, partíamos de la expectativa de cierta correlación entre estas dos dimensiones. Dicha correlación se ha comprobado

en la muestra manejada y se ha observado, además, que, usando únicamente la frecuencia de uso como criterio nivelador, la coincidencia con la nivelación de partida es considerable: aproximadamente la mitad de colocaciones reciben el mismo nivel y un 80% o bien coinciden con el nivel original o quedan en uno adyacente.

La revisión cualitativa de las colocaciones del *DiCE* niveladas mediante el sistema propuesto revela que los resultados son en su mayoría coherentes con las directrices defendidas en *MCER* y *PCIC*: se presenta el léxico más rentable y perteneciente a un 'registro neutro' en niveles más bajos, mientras que colocaciones no tan rentables o pertenecientes a registros marcados aparecen en niveles más avanzados. Además, a la luz de ciertas discrepancias entre la nivelación propuesta y la originaria de la muestra, se ha cuestionado la pertinencia de retrasar la presentación de colocaciones frecuentes pero relativamente opacas a niveles altos.

La propuesta presentada aquí constituye una aproximación inicial que permite asignar nivel de forma prácticamente automática a un conjunto muy amplio de colocaciones. Con todo, el método es susceptible de mejoras que incorporen criterios adicionales al de la frecuencia léxica (Spina, 2016), la dispersión entre ellos. Cabe apuntar también que la revisión de cada entrada del diccionario por parte de profesionales puede refinar el resultado considerablemente.

REFERENCIAS BIBLIOGRÁFICAS

Alba-Salas, J. (2009). Las estructuras tipo meter miedo en la diacronía: Más detalles sobre la evolución histórica de las colocaciones causativas. En A. Enrique-Arias (Ed.), *Diacronía de las lenguas iberorrománicas* (pp. 343-363). Madrid/Frankfurt: Iberoamericana-Vervuert.

Alonso Ramos, M. (2004a). *Diccionario de Colocaciones del Español* [en línea]. Disponible en: http://www.dicesp.com/

Alonso Ramos, M. (2004b). *Las construcciones con verbo de apoyo*. Madrid: Visor Libros.

Alonso Ramos, M. (2012). Explorando la frecuencia léxica para el Diccionario de colocaciones del español. En T. Jiménez Juliá, B. López Meirama, V. Vázquez Rozas & A. Veiga (Eds.), *Cum corde et in nova grammatica. Estudios ofrecidos a Guillermo Rojo* (pp. 19-40). Universidade de Santiago de Compostela.

Alonso Ramos, M. (2015). El Diccionario de Colocaciones del Español: Una puesta al día. *Estudios de Lexicografía*, 5, 103-122.

Alonso Ramos, M. (2016). Learning resources for Spanish collocations: From a dictionary towards a writing assistant. En B. Sanromán Vilas (Ed.), *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching*, volume C of Mémoires de la Société Néophilologique de Helsinki (pp. 65-95). Helsinki: Société Néophilologique de Helsinki.

Alvar Ezquerra, M. (2004). La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera. En M. A. Castillo, O. Cruz & J. P. Mora (Coords.), *Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: Deseo y realidad. Actas del XV Congreso de ASELE* (pp. 19-39). Sevilla: Universidad de Sevilla.

Benigno, V., Kraiff, O., Grossmann, F. & Velez. A. (2016). La notion de collocation fondamentale: Une étude de corpus. *Cahiers de Lexicologie*, *108*(1), 125-146.

Capel, A. (2010). A1-B1 Vocabulary: Insights and issues arising from the English Profile Wordlists Project. *English Profile Journal,* 1, 1-11.

Capel, A. (2015). Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal,* 3, 1-12.

Christ, H. & Christ, I. (1951). Les débats sur le Français fondamental et leur influence sur l'enseignement du français en Allemagne. *Documents pour l'histoire du français lange étrangère ou seconde*, *26*(1), 153-172.

Corpas Pastor, G. (2015). Register-specific Collocational constructions in English and Spanish: A usage-based approach. *Journal of Social Sciences*, *11*(3), 139-151.

Corpas Pastor, G. (2016). Collocations dictionaries for English and Spanish: The state of the art. En A. Orlandi & L. Giacomini (Eds.), *Defining collocations for lexicographic purposes: From linguistic theory to lexicographic practice* (pp. 173-208). Frankfurt: Peter Lang.

Consejo de Europa. (2002). *Marco Común Europeo de Referencia para las Lenguas*. Madrid: Instituto Cervantes/Ministerio de Educación/Anaya.

Ferrando Aramo, V. (2012). *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE*. Tesis doctoral, Universitat Rovira i Virgili, Tarragona, España.

Gómez Molina, J. R. (2004). Los contenidos léxico-semánticos. En J. Sánchez Lobato & I. Santos Gargallo (Dirs.), *Vademécum para la formación de profesores* (pp. 789-811). Madrid: SGEL.

Gougenheim, G., Michéa, R., Rivenc, P. & Sauvageot, A. (1964*). L'élaboration du français fundamental*. París: Didier.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403-437.

Hindmarsh, R. (1980). *The Cambridge English lexicon*. Cambridge: Cambridge University Press.

Hoey, M. (2005). *Lexical priming. A new theory of words and language*. Londres/Nueva York: Routledge.

Instituto Cervantes. (1997-2016). *Plan Curricular del Insituto Cervantes* [en línea]. Disponible en: http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/default.htm

Juilland, A. & Chang-Rodríguez, E. (1964). *Frecuency dictionary of Spanish words*. La Haya: Mouton.

Keniston, H. (1920). Common words in Spanish. *Hispania,* 3, 85-96.

Kilgarriff, A. & Renau, I. (2013). esTenTen, a Vast Web Corpus of Peninsular and American Spanish. *Procedia - Social and Behavioral Sciences*, 95, 12–19.

Laya Gómez, L. (2014). *Nivelación de las colocaciones en ELE*. Tesis de Magíster, Universidade da Coruña, A Coruña, España.

Lewis, M. (1993). *The lexical approach*. Londres: Language Teaching Publications.

Martínez, R. (2013). A framework for the inclusion of multi-word expressions in ELT. *ELT Journal*, 67, 184-198.

Mel'čuk, I. (1996). Lexical functions: A tool for the description of lexical relations in the lexicon. En L. Wanner (Ed.), *Lexical Functions in Lexicography and Natural Language Processing* (pp. 37-102). Amsterdam/Philadelphia: John Benjamins.

Mel'čuk, I. (2012). Phraseology in the language, in the dictionary and in the computer. *Yearbook of Phraseology*, 3, 31-56.

Mel'čuk, I., Clas, A. & Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Louvain-la-Neuve : Duculot.

McGee, I. D. (2006). *Lexical Intuitions and Collocation Patterns in Corpora*. Cardiff: Cardiff University.

McGee, I. D. (2008). Word frequency estimates revisited – A response to Alderson (2007). *Applied Linguistics, 29*(3), 509-514.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nesselhauf, N. (2004). *Collocations in a learner corpus*. Amsterdam/Philadelphia: John Benjamins.

Real Academia Española (s.f). *Banco de datos (CORPES XXI). Corpus del Español del Siglo XXI (CORPES)* [en línea]. Disponible en: <http://www.rae.es>

Rojo Mejuto, N. (2015). *Hacia una clasificación de las colocaciones léxicas por niveles de aprendizaje*. Tesis de magíster, Universidad de Salamanca, Salamanca, España.

Schmitt, N. (2010). *Researching vocabulary. A vocabulary research manual*. Londres: Palgrave MacMillan.

Sinclair, J. M. & Renouf, A. (1985). A lexical learning syllabus for language. En C. Ronald & M. McCarthy (Eds.), *Vocabulary and Language Teaching*, (pp. 140-160). Londres/Nueva York: Longman.

Siyanova, A. & Schmitt, N. (2008). L2 learner production and processing of Collocation : A multi-study perspective. *The Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes*, 3, 429-458.

Siyanova, A. & Spina, S. (2015). Investigation of native speaker and second language learner intuition of Collocation frequency. *Language Learning*, *65*(3), 533-562.

Spärck, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28, 11-21.

Spina, S. (2016). Learner corpus research and phraseology in Italian as a second language: The case of the *DICI-A*, a learner dictionary of Italian collocations. En B. Sanromán Vilas (Ed.), *Collocations Cross-Linguistically. Corpora, Dictionaries and Language Teaching* (pp. 219-244). Helsinki : Société Néophilologique de Helsinki.

Vázquez Veiga, N. (2014). Marcas de uso del Diccionario de Colocaciones del Español. *Zeitschrift für Romanische Philologie, 130*(3), 698-724.

Vincze, O. & Alonso Ramos, M. (2013a). Testing an electronic collocation dictionary interface: Diccionario de Colocaciones del Español. En I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets & M. Tuulik (Eds.), *Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 conference, 17-19 October 2013, Tallinn, Estonia* (pp. 328-337). Trojina/Eesti Keele Instituut, Ljubljana/Tallinn: Institute for Applied Slovene Studies.

Vincze, O. & Alonso Ramos, M. (2013b). Incorporating frequency information in a Collocation dictionary: Establishing a methodology. *Procedia -Social and Behavioral Sciences*, 95, 241-248.

Zipf, G. (1935). *The Psycho-Biology of Language*. Cambridge: The MIT Press.

NOTAS

[1] En el enfoque nocio-funcional (Instituto Cervantes, 1997-2016: Introducción) los 'exponentes' son expresiones lingüísticas asociadas a 'nociones' (significados) y 'funciones' (intenciones comunicativas).

[2] Aunque el propósito de McGee es cuestionar la fiabilidad de los corpus como fuentes de frecuencia léxica, los coeficientes de correlación que halla entre listas de frecuencia de corpus distintos (rho de Spearman del 0,82 en un caso, de casi el 0,90 en otro) son más que notables en cualquier caso (McGee, 2008).

[3] Alonso Ramos (2012) apunta casos de vocabulario asociado a contextos muy específicos (p. ej. tarjeta sanitaria'). Son expresiones probablemente infrecuentes, pero útiles y muy necesarias en ciertos casos.

[4] Al respecto se afirma: "[c]omo criterios de selección, siempre partiendo de la apreciación intuitiva basada en la experiencia docente, han primado la frecuencia y la rentabilidad comunicativa" [en línea]. Disponible en: (http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/09_nociones_especificas_introduccion.htm; cursiva nuestra)

[5] Siyanova y Schmitt (2008) obtienen índices de correlación (Spearmann) entre la ordenación hecha por hablantes nativos de inglés y la basada en frecuencia de corpus que van del 0,58, para un conjunto de 31 colocaciones, al 0,74 para un conjunto de 10 colocaciones de frecuencia alta y se muestran relativamente optimistas en este sentido: "[…] N[ative] S[speaker]s not only have good intuitions of what collocations are very frequent and very infrequent in language but can also distinguish finer shades of frequency" (Siyanova & Schmitt, 2008: 445). Siyanova y Spina (2015) emplean una metodología diferente. Según su análisis (esta vez tienen en cuenta la influencia de diversos factores en las respuestas de un grupo de hablantes), la frecuencia de cada una de las colocaciones del experimento no resulta un factor significativo en las estimaciones de sus informantes, pero sí la pertenencia de las colocaciones en cuestión a bandas de frecuencia alta, media, baja y muy baja. Las autoras aquí son más cautelosas sobre la intuición de los hablantes: "It seems that it is almost impossible to answer the question of whether or not language users have accurate intuitions about collocation frequency; it all depends on the frequency range in question […]" (Siyanova & Spina, 2015: 555).

[6] La frecuencia máxima ni siquiera se ha tenido en cuenta, pues como se aprecia en la Tabla 1, tomarla como límite superior de cada nivel hubiera supuesto incluir todas las colocaciones del *DiCE* en el nivel C2, ya que a la colocación con la frecuencia más alta de la muestra se le asigna ese nivel en el *PCIC*.

[7] Para la noción de 'verbo de apoyo', véase Alonso Ramos (2004b).

AGRADECIMIENTOS

***** Esta investigación ha contado con financiación de los proyectos FFI2011-30219-C02-01 (Ministerio de Ciencia e Innovación), FFI2016-78299-P (Ministerio de Economía y Competitividad) y la ayuda postdoctoral (POS-A/2013/191) de la Xunta de Galicia.