Orientaciones para la nivelación de secuencias formulaicas en ELE según el criterio de frecuencia

NARCISO CONTRERAS IZQUIERDO[1] & FERMÍN MARTOS ELICHE[2]
Universidad de Jaén[1], Universidad de Granada[2]

Resumen
Con este trabajo pretendemos orientar al profesorado de ELE en la introducción de secuencias formulaicas (SF) en el aula de acuerdo con el criterio de frecuencia. Se trata de ofrecer datos objetivos para su nivelación desde los niveles A1 a C2. Hemos seleccionado 3260 SF del *Plan Curricular del Instituto Cervantes* (PCIC) y hemos analizado su frecuencia en *Google*, como motor de búsqueda general, y en tres corpus lingüísticos especializados en el español: dos de la RAE (CREA y CORPES XXI) y el corpus de Mark Davies (*Corpus del Español*). Nuestro objetivo es establecer un índice de frecuencia para su nivelación, comprobando igualmente si existe un grado de coincidencia entre la frecuencia y el nivel establecido en el PCIC. Finalmente, completamos dicho índice de frecuencia con el análisis de las 100 SF más frecuentes atendiendo a otros criterios, rentables comunicativamente, de tipo lexicométrico, lingüístico y nociofuncional.
Palabras clave: competencia léxica; secuencias formulaicas; corpus lingüísticos; enseñanza de español LE/L2; niveles de enseñanza/aprendizaje; frecuencia léxica

Abstract
With this work we intend to guide SFL teachers in the introduction of formula sequences (FS) in the classroom according to the frequency criteria. It is about offering objective data for leveling from levels A1 to C2. We have selected 3260 FS from the *Plan Curricular del Instituto Cervantes* (PCIC) and have analyzed its frequency in Google, as a general search engine, and in three specialized linguistic corpus in Spanish: two of the RAE (CREA and CORPES XXI) and the corpus by Mark Davies (*Corpus del Español*). Our goal is to establish a frequency index for its leveling, also checking if there is a degree of coincidence between the frequency and the level established in the PCIC. Finally, we complete the frequency index with the analysis of the 100 most frequent SFs according to other criteria, profitable communicatively, of a lexicometric, linguistic and notional-functional.
Keywords: lexical competence, formulaic sequences, linguistic corpus, teaching Spanish LE / L2, teaching / learning levels, lexical frequency.

1 Introducción

Si se considera la competencia léxica como elemento central del desarrollo de la competencia comunicativa en la enseñanza y aprendizaje de lenguas extranjeras, las denominadas *secuencias formulaicas* (en adelante SF) ocupan un lugar destacado. Se trata de unidades léxicas pluriverbales o "combinaciones de palabras que se almacenan y recuperan de la memoria como un todo en el momento de su uso" (Wray 2002: 9), y que los hablantes nativos emplean en su discurso de forma natural como recurso expresivo en situaciones comunicativas determinadas.

El dominio de estas SF mejora la fluidez y la comprensión de los aprendientes de lenguas extranjeras en la interacción comunicativa, tal y como propone Lewis

(1993,1997) en los conocidos postulados del enfoque léxico. Tan importante es su control que las SF pueden llegar a constituir más de la mitad del discurso nativo (Erman & Warren 2000).

El documento que recoge las SF en el caso del español (ELE/L2) es el *Plan Curricular del Instituto Cervantes* (PCIC), concretamente en el apartado de "Nociones específicas". Se ofrece allí una compilación de SF en series abiertas y se establece el nivel (desde A1 a C2) en que estas unidades deben aparecer sin establecer criterio alguno de su nivelación. Solo se arguye respecto a la selección de estas unidades la apreciación intuitiva basada en la experiencia docente. En este sentido, Penadés (1999) ya reconocía que no existen estudios que nos permitan precisar y determinar con precisión qué SF enseñar en los distintos niveles de aprendizaje.

En trabajos anteriores (Martos & Contreras 2018; Contreras & Martos 2020), comprobamos que dicha apreciación intuitiva es bastante precisa para la adscripción de las SF a los niveles en el PCIC y presentamos la base de una posible nivelación de esas SF a partir del análisis de frecuencia en diferentes corpus, por lo que ahora nos proponemos como objetivo ampliar dicha propuesta, estableciendo puntos de referencia en la frecuencia a partir de los cuales el profesor puede establecer en qué nivel enseñar una u otra SF.

En el ámbito de los estudios fraseológicos, el empleo de frecuencia mide la rentabilidad el léxico (Alvar 2005, García Salido 2017, García & Alonso 2018, Nation 2001, Sinclair & Renouf, 1985), por lo que en este estudio nos planteamos un criterio puramente objetivo (la media de uso a partir de la frecuencia), sabiendo que existen criterios más subjetivos que podrían tomarse en cuenta como la dispersión en los corpus de consulta, la rentabilidad y productividad, la opacidad o transparencia, los factores culturales, los intereses de los alumnos, las propias producciones de los estudiantes, o incluso las intuiciones de los hablantes nativos (Benigno, Kraiff, Grossmann & Velez 2016, García 2017, García & Alonso 2018, Gómez 2004, McGee 2008, Nation 2001).

Conscientes de todo ello, pero teniendo en cuenta también los objetivos y las dimensiones de esta propuesta, complementamos los datos de la frecuencia léxica absoluta atendiendo a otros criterios rentables comunicativamente como son los datos lexicométricos relativos a la frecuencia normalizada, la dispersión y la densidad léxica. Igualmente, atenderemos a aspectos lingüísticos como la estructura y la opacidad semántica de las SF, así como la variación diatópica y la modalidad (oral/escrita). Finalmente, en consonancia con el enfoque nociofuncional adoptado por el PCIC, tendremos en cuenta un aspecto fundamental para la nivelación de estos exponentes léxicos como son los ámbitos temáticos en los que se emplean, relacionados con los descriptores de las escalas ilustrativas del MCER y las nociones específicas asignadas en el PCIC.

2 Marco teórico

2.1 Competencia léxica y secuencias formulaicas

La competencia léxica, definida en el MCER (2002: 126) como el conocimiento del vocabulario de una lengua y la capacidad para utilizarlo, está formada por diversos tipos de unidades: palabras simples y compuestas, fórmulas rutinarias o expresiones institucionalizadas, colocaciones, compuestos sintagmáticos, locuciones idiomáticas o modismos (Gómez 2004). Dicha diversidad en la composición de esta competencia léxica ya aparece recogida tanto en el MCER como en el PCIC, puesto que manejan un concepto amplio de unidad léxica, enfoque que "se sitúa en una línea de investigación que parte de la base de que un hablante cuenta, además de con unidades léxicas simples o palabras, con un número amplio de bloques semiconstruidos que puede combinar al hablar" (PCIC, III, 2006: 385)[1].

No es por tanto de extrañar que el nivel léxico haya centrado la atención en los últimos años de la lingüística y de la adquisición de lenguas, y que actualmente se considere el léxico como componente central y transversal en el discurso (Battaner & Arias 2019), de ahí la necesidad de desarrollar nuevos avances teóricos y metodológicos para su enseñanza y aprendizaje.

Entre estas unidades que forman la competencia léxica, destacan las pluriverbales, para las que existen diversas y variadas denominaciones, definiciones, clasificaciones y caracterizaciones, aunque debido a la naturaleza de este trabajo, no profundizaremos en cuestiones teóricas, ofreciendo una visión general de estas unidades.

Una de las denominaciones más aceptadas es la de "unidad fraseológica", que según la definición ya clásica de Corpas (1996: 20), son "unidades léxicas formadas por más de dos palabras gráficas en su límite inferior, cuyo límite superior se sitúa en el nivel de la oración compuesta. Dichas unidades se caracterizan por su alta frecuencia de uso, y de coaparición de sus elementos integrantes; por su institucionalización, entendida en términos de fijación y especialización semántica; por su idiomaticidad y variación potenciales; así como por el grado en el cual se dan todos estos aspectos en los distintos tipos". Igualmente, en el ámbito de la enseñanza de lenguas se ha extendido el uso de "secuencias formulaicas" (SF), que para Wray (2002: 9) son una combinación de palabras, continua o discontinua, que es o parece ser prefabricada, esto es, que se almacena y recupera de la memoria

[1] Así, para Ferrando (2012: 362), uno de los aspectos más destacables del PCIC es la definición de los contenidos léxico-semánticos desde una perspectiva nocional, que permite dar cuenta de la dimensión combinatoria del léxico, por lo que se incluyen, además de lexías simples, toda una serie de unidades léxicas pluriverbales. No obstante, autores como Velázquez (2018: 22) critican el tratamiento superficial y asistemático de estas unidades en el PCIC, ya que se basa en el principio cuantitativo-acumulativo en detrimento del criterio de la frecuencia de uso y de la funcionalidad comunicativa.

como un todo en el momento de su uso, en lugar de ser un enunciado construido palabra por palabra a partir del conocimiento que se tiene de la gramática[2].

Por lo que se refiere a su caracterización, sus propiedades básicas serían el carácter poliléxico, la idiomaticidad, la fijación y la independencia, si bien estas propiedades pueden estar sometidas a gradación y no darse en todos los casos (Alvarado 2010, Moreno 2012, Ruiz 1997 2000, Velázquez 2018, Corpas 1996, Zuloaga 1980).

De igual modo, se han propuesto diversas clasificaciones para estas unidades, pero quizás la más general y extendida es la de Corpas (1996), quien, partiendo de una concepción más amplia de la fraseología, las divide en colocaciones, locuciones y enunciados fraseológicos[3].

Igualmente, es importante destacar que son secuencias de palabras que los nativos sienten como la manera preferida y natural de expresar una idea particular o propósito, aunque haya otras formas expresivas también posibles (Sánchez 2011 2017). De este modo, el dominio de estas unidades, que pueden llegar a constituir la mitad del discurso nativo (Erman y Warren 2000)[4], mejora la competencia comunicativa de los aprendientes, pues aporta fluidez, precisión y mayor expresividad al discurso, tanto en la comprensión como en la producción (Martín 2012: 61 y ss., Szyndler 2015: 203, Moreno 2012: 36 y ss; Pérez 2017: 25, Timofeeva 2013: 326 y ss.). Igualmente, otro aspecto reseñable es que incluyen información de tipo sociocultural, reflejando aspectos de la idiosincrasia, la historia, los hábitos y costumbres, así como la manera de pensar y conceptualizar el mundo por parte de las comunidades de habla que las emplean (Szyndler 2015: 203-4, Moreno 2012: 40, Timofeeva 2013).

Por tanto, el lenguaje formulaico muestra el léxico como una representación directa del modo en el que el lenguaje funciona dentro de la comunidad de habla del individuo (Moreno 2012: 40), por lo que si el alumno no conoce o no sabe

[2] Sinclair (1991: 110 y ss.) distingue entre el "principio de la elección libre" (*open choice principle*) y el "principio de la locución" (*principle of idiom*): el primero es la capacidad de seleccionar palabras individuales, lo que no es suficiente para producir discurso por parte de un hablante, por lo que, para complementar la capacidad comunicativa, tiene que completarse con recursos ya preestablecidos y almacenados en la mente, esto es, el "principio de la locución". En la lingüística actual se reconocen ambos modelos, pero no existe acuerdo sobre cuál de ellos prevalece en la producción y el procesamiento lingüísticos: mientras que para Sinclair (2004) y Wray (2002) el principio de locución predomina sobre el de elección libre, Chomsky da más importancia a los procesos analíticos. Además, existen modelos duales en los que los procesos analíticos y holísticos actúan de forma complementaria, como la propuesta de Coseriu (1981: 297-302), que habla de "técnica libre del discurso" y "discurso repetido". Finalmente, estudios más actuales revelan que dichos modelos se asocian a las circunstancias de uso: el empleo del modelo basado en las reglas requiere condiciones en las que se pueda planear el discurso, mientras que el basado en la memoria favorece la fluidez en situaciones de comunicación a tiempo real (Moreno Teva 2012: 40-41; Pérez 2017: 32).

[3] Otras clasificaciones clásicas son las Casares (1950), Zuluaga (1980), Ruiz Gurillo (1997, 2000), Erman & Warren (2000), aunque estimados que la de Corpas (1996) es muy útil desde el punto de vista didáctico, ya que es más amplia y su clasificación en tres esferas facilita la práctica de las UF.

[4] Moreno (2012: 20 y 55) presenta algunos estudios sobre las estimaciones de la extensión del lenguaje formulaico en el discurso nativo y no nativo en los que existen diferencias importantes.

emplear bien las SF su discurso no se asemejará, en ningún caso, al de un hablante nativo, sino que sonará forzado, ensayado, superficial o, "no idiomático" (como lo denomina Lennon 1998: 12), pues no se ajusta a lo esperado. Así pues, el uso inadecuado de las UF puede romper convenciones de cortesía, llevar a malentendidos y perjudicar el proceso de comunicación. Como indica Velázquez (2018: 32): "Recuérdese que, en última instancia, el usuario de la lengua ha de ser capaz de desenvolverse en diversos contextos comunicativos, para lo cual requiere –además de las competencias generales– no sólo competencias lingüísticas, sino también socioculturales y pragmáticas"[5].

Parece, por tanto, evidente, la necesidad de integrar estas unidades pluriverbales en la enseñanza de lenguas, aunque la cuestión de cuándo introducir las SF suscita diversas opiniones, que pueden agruparse entre los que consideran que deben presentarse en los niveles intermedio y avanzado y los que defienden que deben integrarse en todos los niveles. Frente a esto, sí que existe un amplio consenso en las dificultades para su aprendizaje debido a sus características formales, semánticas y pragmáticas (Timofeeva 2013: 324-5, Olímpio 2006: 6, Wray 2002: 468).

No obstante, y a pesar de su evidente interés en el desarrollo de la competencia léxica de los alumnos, faltan investigaciones sobre estas unidades que promuevan un aprendizaje efectivo, comenzando por estudios que nos permitan determinar qué SF enseñar en los diferentes niveles de aprendizaje. En este sentido, en los últimos años se viene hablando de "fraseodidáctica" (Corpas 2003, Crida & Alessandro 2019, Ettinger 2008, González Rey 2006, López 2011, Szyndler 2015), cuyo objetivo es la descripción de los fenómenos relacionados con las expresiones fraseológicas, así como su enseñanza en lengua materna y en lengua extranjera, atendiendo tanto a la comprensión como a la expresión de fraseologismos (López 2011: 533).

En esta línea, una orientación metodológica adecuada para la enseñanza y aprendizaje de estas fórmulas es el enfoque léxico, desarrollado por Michael Lewis (1993, 1997)[6]. Para dicho autor, el léxico es el eje central del proceso de desarrollo de la competencia comunicativa ("la lengua consiste en léxico gramaticalizado, no en gramática lexicalizada", Lewis 1993: 51), y para ello propone partir del aprendizaje de bloques prefabricados de palabras (*chunks*), segmentos léxicos empleados con mucha frecuencia por los hablantes nativos en sus interacciones.

A pesar de la importancia de las SF, muchos currículos, manuales y profesores no priorizan ni organizan su enseñanza, y prefieren tratarlas sin ninguna sistematización cuando aparezcan según el tema tratado. Es por esto por lo que Lewis (1993) sugiere que haya una mayor atención en la presentación del léxico en las clases de lengua extranjera y, por supuesto, un trabajo explícito proporcionando siempre un contexto (situación extralingüística) y un cotexto (entorno lingüístico).

[5] Alvarado (2010: 47 y ss.) analiza algunos de estos hechos pragmáticos como el significado, la modalidad discursiva, la evidencialidad, los actos de habla, la cortesía o la ironía, confirmando la importancia de los aspectos sociales, discursivos y expresivos para la explicación de estas unidades.
[6] Para conocer más sobre enfoque léxico y su aplicación en ELE, véase Pérez (2017).

De ese modo, los aprendientes no solo deben conocer el significado de las palabras, sino también el uso y las restricciones de cada unidad léxica.

Como vemos, estas secuencias formulaicas representarían los segmentos léxicos (*chunks*) mencionados por Lewis, y su tratamiento en el aula sería muy productivo, puesto que si los hablantes los perciben y almacenan en su lexicón de forma holística, como una unidad léxica, y asociados a contextos comunicativos determinados, su aprendizaje repercutiría positivamente en la competencia comunicativa de los aprendientes: al percibir, entender, recuperar y producir estas unidades como un bloque mejoraría su fluidez y nivel de comprensión en la interacción comunicativa.

No obstante, aunque desde la década de los años 90 del siglo pasado este tema empieza a ganar importancia y ha aumentado bastante la atención que se le dedica, todavía faltan investigaciones que traten aspectos teóricos y prácticos de la fraseología en el aula de ELE, y así "se transfiere al profesor la tarea de investigador, dado que, para llevar a cabo ciertas propuestas didácticas, le corresponde determinar las unidades fraseológicas más frecuentes o rentables, y definir ciertos parámetros discursivos y pragmáticos de estas unidades, tareas que se encuentran en estado embrionario en la fraseología del español" (Olímpio 2006: 141-2)[7].

2.2 El tratamiento del léxico y las SF en el MCER y el PCIC

Partiendo del hecho de la dificultad de precisar qué unidades léxicas debe aprender un alumno de ELE, debemos intentar determinar los criterios que, tomando como base la frecuencia, sean más rentables para seleccionar y ordenar los ítems léxicos dentro de los diseños curriculares. En este sentido, Baralo (2007: 286) precisa que dichos criterios están interrelacionados con las habilidades comunicativas, con las funciones, con los temas y las situaciones que se quieran enseñar.

Para la selección del léxico que debería adquirir un hablante de español/LE en cada nivel, debemos partir del MCER y del PCIC, tomando la descripción de los niveles de referencia (A1-C2) como base que contribuya a determinar la competencia léxica que debería adquirir un aprendiente de ELE. Por tanto, la descripción de dichos niveles debe servir como instrumento práctico que rija la selección de las unidades léxicas que deben enseñarse y aprenderse en cada etapa.

A pesar de que, como señala Baralo (2005: 31), el tratamiento de la competencia léxica en el MCER es muy sencillo y bastante incompleto, sí que puede servir como base y marco para el tratamiento del léxico en el aula de lenguas extranjeras. Así, en el apartado 6.4.7.2. del MCER se destaca que los parámetros que debemos tener en cuenta para, entre otras, la planificación del aprendizaje y enseñanza de lenguas

[7] Olímpio (2006: 129 y ss.) recoge diversos aspectos, tanto desde el punto de vista teórico como didáctico, en el tratamiento de estas unidades léxicas en la enseñanza de ELE, sobre los que sería necesaria una investigación más amplia: su adquisición, su producción, la frecuencia de uso, las estrategias involucradas en su aprendizaje, el influjo de la L1 sobre la L2, etc. Además de esta obra, Aguilar (2013), González Rey (2012) y Serradilla (2014) realizan una importante recopilación bibliográfica sobre este tema.

son la riqueza (es decir, el número de palabras y de expresiones hechas) y el alcance (los ámbitos, temas, etc.) del vocabulario que deberá controlar el alumno.

En las siguientes tablas podemos observar con más detalle las escalas ilustrativas para la gradación del conocimiento léxico por lo que se refiere a dichos parámetros de riqueza y dominio del vocabulario (MCER, 2002: 109):

	RIQUEZA DE VOCABULARIO
C2	Tiene un buen dominio de un repertorio léxico muy amplio, que incluye expresiones idiomáticas y coloquiales; muestra que es capaz de apreciar los niveles connotativos del significado.
C1	Tiene un buen dominio de un amplio repertorio léxico que le permite superar con soltura sus deficiencias mediante circunloquios; apenas se le nota que busca expresiones o que utiliza estrategias de evitación. Buen dominio de expresiones idiomáticas y coloquiales.
B2	Dispone de un amplio vocabulario sobre asuntos relativos a su especialidad y sobre temas más generales. Varía la formulación para evitar la frecuente repetición, pero las deficiencias léxicas todavía pueden provocar vacilación y circunloquios.
B1	Tiene suficiente vocabulario para expresarse con algún circunloquio sobre la mayoría de los temas pertinentes para su vida diaria, como, por ejemplo, familia, aficiones e intereses, trabajo, viajes y hechos de actualidad.
A2	Tiene suficiente vocabulario para desenvolverse en actividades habituales y en transacciones cotidianas que comprenden situaciones y temas conocidos.
	Tiene suficiente vocabulario para expresar necesidades comunicativas básicas. Tiene suficiente vocabulario para satisfacer necesidades sencillas de supervivencia.
A1	Tiene un repertorio básico de palabras y frases aisladas relativas a situaciones concretas.

	DOMINIO DEL VOCABULARIO
C2	Utiliza con consistencia un vocabulario correcto y apropiado.
C1	Pequeños y esporádicos deslices, pero sin errores importantes de vocabulario.
B2	Su precisión léxica es generalmente alta, aunque tenga alguna confusión o cometa alguna incorrección al seleccionar las palabras, sin que ello obstaculice la comunicación.
B1	Manifiesta un buen dominio del vocabulario elemental, pero todavía comete errores importantes cuando expresa pensamientos más complejos, o cuando aborda temas y situaciones poco frecuentes.
A2	Domina un limitado repertorio relativo a necesidades concretas y cotidianas.
A1	No hay descriptor disponible.

Como podemos apreciar, y tal y como puntualiza Baralo (2005: 35), esta descripción de las escalas pone de manifiesto la relación entre el léxico y las situaciones de comunicación, pues son estas las que establecen los niveles. De este modo, al A1 le corresponde un repertorio léxico propio de situaciones concretas; al A2, el de entornos cotidianos y habituales; al B1, el léxico propio de la ampliación de la vida diaria, con temas relativos al trabajo, las aficiones e intereses, los viajes y los hechos de actualidad; el B2 incluye los asuntos relacionados con su especialidad; el C1, más riqueza y precisión, así como el uso de expresiones idiomáticas en registros formales y coloquiales; el C2, con su variedad y calidad, incluyendo los aspectos connotativos del significado, es semejante al del hablante nativo culto.

Por tanto, como precisa Pérez Serrano (2017: 216), de estas escalas (MCER), se puede deducir que el grado de coloquialidad, de idiomaticidad, el carácter regional

y la especialización van asociados a una mayor riqueza de vocabulario y, por tanto, a un nivel de superior, mientras que la brevedad, habitualidad, sencillez y cotidianidad se asocian a los niveles más iniciales. Esto concuerda con la opinión de Muñoz-Basols (2015), quien propone que, a la hora de enseñar el lenguaje idiomático, habría que tener en cuenta tres aspectos: la frecuencia de uso, la noción de registro y la variación diatópica.

No obstante, como venimos repitiendo y como igualmente defienden diversos autores, la idiomaticidad y la coloquialidad no deberían ser criterios de selección privativos de las unidades léxicas que forman parte de los currículos avanzados, "pues, creemos que la habilidad para reconocer el carácter figurativo que presentan numerosas expresiones idiomáticas, colocaciones y palabras, no debería dejarse para los niveles más altos de dominio lingüístico sino que debe trabajarse desde los niveles iniciales" (Pérez Serrano, 2017: 221). De este modo, y como sugiere Muñoz-Basols (2015: 449), el tratamiento del lenguaje idiomático en diseños curriculares como el MCER y el PCIC, relegándolo a niveles superiores, ha podido contribuir a que su enseñanza no se haya integrado en todos los niveles de aprendizaje. Frente a esto, y teniendo siempre en cuenta las dificultades de su enseñanza y aprendizaje por sus características inherentes, es un componente del léxico que, siguiendo unas pautas determinadas, puede y debe incorporarse en el diseño curricular de cualquier nivel.

Otra orientación esencial que recoge el MCER, y que fundamenta igualmente la orientación de nuestro análisis, son las opciones metodológicas para la selección léxica (6.4.7.3.), y que se articulan en cuatro grandes principios: a. elegir palabras y frases clave: en áreas temáticas necesarias para la consecución de tareas comunicativas adecuadas a las necesidades de los alumnos; […] b. seguir unos principios léxico-estadísticos que seleccionen las palabras más frecuentes en recuentos generales y amplios o las palabras que se utilizan para áreas temáticas delimitadas; c. elegir textos (auténticos) hablados y escritos y aprender o enseñar todas las palabras que contienen; d. no realizar una planificación previa del desarrollo del vocabulario, pero permitir que se desarrolle orgánicamente en respuesta a la demanda del alumno cuando éste se encuentre realizando tareas comunicativas.

Vemos, por tanto, que estos principios se articulan en torno a la rentabilidad comunicativa, la frecuencia, la autenticidad y las necesidades de los alumnos, todo ello en línea con el enfoque nociofuncional (Wilkins, 1972) que rige el PCIC, organizando el aprendizaje en términos de los propósitos sociales –y no lingüísticos– de los aprendientes. Este mismo autor (Wilkins, 1972) considera que las funciones y las nociones son unidades de análisis que se basan en el significado, frente a las unidades de análisis de los programas estructurales, que son de carácter lingüístico. Además, organizar la instrucción lingüística en términos semánticos posibilita que el aprendiente acceda a un rango mucho más amplio de elementos lingüísticos que lo que permitía el enfoque estructural.

Por lo que se refiere a las SF, como hemos mencionado, el PCIC incluye exponentes fijos (expresiones idiomáticas y frases hechas) desde los niveles

iniciales, aunque es en los superiores donde reciben mayor atención (PCIC, Introducción, I: 208)[8]. Así, en dicha obra se admite que no se hace referencia explícitamente a estas unidades en los niveles iniciales (A1 y A2) debido a que la falta de conocimientos puede impedir a los aprendientes principiantes analizar los fraseologismos (PCIC, III, 2006: 173). Esto puede presentar ciertos inconvenientes, tanto desde el punto de vista del estudiante, pues las SF están presentes en el proceso de aprendizaje desde el primer momento, sin que el alumno deba ser consciente de su complejidad formal o su pertenencia a la fraseología[9], como desde el punto de vista del docente, ya que el PCIC debería ofrecer indicaciones claras acerca de qué unidades pluriverbales es necesario adquirir en los niveles del MCER (A1-C2) para que los profesores pudiesen organizar estos contenidos (López 2011: 534-5).

Por otro lado, hay que tener en cuenta que, tal y como se indica en el PCIC, sus inventarios deben considerarse abiertos y no exhaustivos "en cuanto que siempre podrían añadirse nuevos elementos, tanto en sentido horizontal (completando las series), como en sentido vertical (con nuevos elementos en cada nivel)" (2006: 393-394). Por tanto, es labor de los programadores, profesores y creadores de materiales adaptar estos inventarios a las necesidades de los estudiantes y a la situación particular de enseñanza, así como añadir nuevos elementos que los completen. A pesar de esta falta de exhaustividad, estos catálogos constituyen una ayuda inestimable para determinar qué llevar al aula en los diferentes niveles, pues, tal como se indica en su introducción, es "una base de orientación general que podrá servir de referencia a quien lo utilice a la hora de seleccionar y distribuir por niveles las unidades léxicas que precise para sus propios fines" (PCIC, 2006: 333).

Así pues, compartimos con la mayoría de estudiosos sobre el tema la opinión de que es fundamental que estas SF se trabajen desde los niveles iniciales, tanto por la rentabilidad comunicativa que hemos comentado anteriormente, como para el desarrollo de las redes léxicas (Ferrando 2012: 360, Penadés 1999, Gómez 2004, Higueras 2004, Leontaridi, Ruiz & Peramos 2011: 188, Muñoz-Basols 2015; Saracho 2016, Velázquez 2018: 32), aunque su estudio deberá intensificarse en los niveles avanzados con el objetivo de aumentar la fluidez y precisión lingüística del aprendiente.

De este modo, pretendemos que nuestro estudio pueda servir de guía y orientación a los profesionales en el ámbito de ELE en esa labor de adaptación y ampliación del inventario propuesto en el PCIC en función de las necesidades de los estudiantes y de las situaciones particulares de enseñanza. Para ello, y como comentaremos a continuación, la lingüística de corpus (LC) se ha convertido en un valioso instrumento para dicha nivelación.

[8] Velázquez (2018: 22) constata que el 99% de las unidades fraseológicas recogidas en el PCIC se inscriben en los niveles C1 (27%) y C2 (72%). Para conocer más sobre combinatoria léxica en el PCIC, remitimos a Pérez (2015: 229).

[9] Además, como señala Timofeeva (2013: 331), curiosamente, y debido a la capacidad holística que se tiene en la lengua materna, los alumnos, incluso en los niveles iniciales, suelen detectar con éxito estas SF en un texto adecuado a su grado de competencia, aun sin saber qué significan estas.

2.3 SF, lingüística de corpus y frecuencia léxica

Como acabamos de comentar, pensamos que el uso de la lingüística de corpus (LC) en ELE puede aportar importantes beneficios en este campo pues, en primer lugar, pueden constituir herramientas didácticas eficaces para favorecer el desarrollo de la competencia comunicativa de nuestros alumnos al mostrar el uso lingüístico contextualizado. Además, en segundo lugar, puede ofrecer datos estadísticos sobre la frecuencia de uso de las unidades léxicas muy útiles para la nivelación de las SF.

En términos generales, podríamos apuntar que la LC, dentro de la lingüística aplicada, se centra en la creación, la explotación, el análisis de textos orales y escritos de un corpus electrónico, así como también sus aplicaciones a otros campos[10]. En palabras de Sinclair (2004: 16): "A corpus is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research". No obstante, es necesario aclarar que un corpus no constituye la totalidad de la lengua, sino que recoge una muestra representativa de la misma (Cruz 2017: 35).

Aunque en otros ámbitos como el anglosajón este campo evidencia un amplio desarrollo desde hace tiempo, el interés creciente por los corpus lingüísticos se manifiesta ahora como una novedad en ELE, debido principalmente a que son un recurso relativamente reciente en nuestro idioma, y también a que los profesores no suelen poseer formación en lingüística de corpus (Villayandre & Llanos 2017: 58).

A pesar de este retraso, encontramos obras como la de Cruz (2017), que explora adecuadamente este surgimiento desde la óptica ELE, presentando conceptos básicos de la lingüística de corpus, así como diversas aplicaciones a la enseñanza del español, todo ello acompañado por actividades para poner en práctica lo expuesto en la obra. Igualmente, varios autores, incluso ya con perspectiva histórica, y desde la lingüística teórica, analizan los corpus de español a los que podemos acceder como profesionales (Bolaños 2015, Briz & Albelda 2009, Cruz 2014 2017, Pérez 2017, Rojo 2016).

Desde este punto de vista de la enseñanza y aprendizaje, los corpus constituyen excelentes recursos, tanto para los docentes como para los discentes, que propician la práctica de aspectos pragmáticos, discursivos y culturales, puesto que contribuyen a la contextualización de la lengua que se trabaja en clase, potenciando la interacción comunicativa y la negociación del significado en el aula. Del mismo modo, posibilitan que el alumno esté expuesto a muestras lingüísticas de hablantes nativos de diversas características sociales y dialectales. En cuanto a las SF,

[10] Como señala Bolaños (2015: 34-5), la lingüística de corpus se inscribe en el mismo paradigma de las disciplinas lingüísticas que, con el giro de la lingüística en los años sesenta del siglo pasado, se dedican a estudiar las diferentes manifestaciones del uso del lenguaje en contextos reales de interacción comunicativa con un énfasis en lo comunicativo y pragmático y con un enfoque lingüístico funcional, frente a los estudios teóricos del sistema de la tradición lingüística anterior. Para conocer con más detalle la relación entre la LC y el pensamiento lingüístico, recomendamos Cruz (2017: 29 y ss.). Igualmente, para las aplicaciones de la LC en el caso del español para la elaboración de gramáticas, diccionarios, así como en disciplinas y fenómenos como la dialectología, la disponibilidad léxica, la traducción, la sociolingüística, etc., véase Cruz (2017: 77 y ss.).

propician el desarrollo de la competencia léxica mediante la detección de patrones de palabras (*chunks*) que suelen aparecer juntas (Albelda 2011, Bolaños 2015, Buyse 2017, Cruz 2017, Villayandre & Llanos 2017: 53-55)[11].

En segundo lugar, los corpus son igualmente una valiosa herramienta de información estadística sobre frecuencia de uso de las unidades léxicas, algo que podría ayudar al profesor a establecer prioridades en la enseñanza y seleccionar y graduar, así, los contenidos que son más habituales en la lengua, posibilitando una elección más objetiva y menos intuitiva. En el caso del español, el criterio de la frecuencia para la selección de exponentes ya es asumido en el PCIC, aunque parte de la apreciación intuitiva basada en la experiencia docente, y lo completa con la rentabilidad comunicativa, es decir, que se trate de exponentes que sirvan para realizar las funciones comunicativas propias de cada nivel (PCIC, 2006: Nociones específicas, Introducción).

La aplicación de estudios de frecuencia como fuente de información lingüística en general y a la enseñanza del vocabulario de una L2/LE en particular, tiene una historia relativamente larga (Cruz 2017: 25, Ferrando 2012, García 2017, García & Alonso 2018: 159-61, Nation 2001: 329) y, precisamente el desarrollo de la LC ha contribuido al renovado interés por el vocabulario que se observa en las últimas décadas, interés que viene acompañado por la vuelta a los estudios de frecuencia léxica. Por ejemplo, en el ámbito hispánico, Alvar (2005) propone que sea la frecuencia lo que determine cuál es el vocabulario que con más urgencia necesita un aprendiz de español, aunque este mismo autor llame la atención sobre el hecho de que "la información sobre la frecuencia de uso no puede ser un valor rígido al que debamos sujetarnos, pues hay otros factores que influyen en el aprendizaje de las palabras" (Alvar 2005: 19).

Es por esto por lo que, aunque la frecuencia ha sido un criterio esencial para la selección del vocabulario fundamental que debe enseñarse, por sí sola puede no ser suficiente, pues como ya observaba Nation (1990), por un lado, algunas palabras útiles presentan niveles muy bajos de frecuencia y, por otro, las palabras más frecuentes suelen ser las de contenido gramatical[12]. Por tanto, parece necesario emplear más parámetros para la selección, como el de la disponibilidad léxica (Bustos 2001), si bien los estudios en este ámbito se centran en la palabra, por lo que debería desarrollarse un modelo de encuesta que tuviera en cuenta también las unidades léxicas pluriverbales para determinar qué SF enseñar en cada nivel (Ferrando 2012: 361).

Además de la frecuencia, en los estudios sobre este tema se recogen otros criterios con respecto a la selección del léxico para determinar cuál es el vocabulario propio de los diferentes niveles distinguidos por el MCER, entre otros: la dispersión en el corpus que se consulte (Alvar 2005), su productividad y los intereses de los aprendices (Gómez 2004, Pérez 2017: 56), las intuiciones de los hablantes (McGee

[11] Cruz (2017: 35-37) resume la utilidad de los corpus.

[12] Por ejemplo, véase el inicio del listado de las 1000 palabras más frecuentes que ofrece el corpus CREA (http://corpus.rae.es/frec/1000_formas.TXT).

2008, Benigno Kraiff Grossmann & Velez 2016), o las producciones de los propios aprendices (García 2017: 32). Debido a las restricciones de espacio de este trabajo, estos otros criterios de selección no son tratados aquí, y merecerían otros trabajos aparte.

3 Metodología

Sobre la base de trabajos anteriores (Martos & Contreras 2018; Contreras & Martos 2020), partimos de la hipótesis de que tanto Internet (a través de *Google*, su buscador más eficiente) como los corpus lingüísticos pueden emplearse como herramienta para el cálculo de la frecuencia de uso de las SF en el discurso nativo, y que dicha frecuencia podría, objetivamente, contribuir a que el profesor pueda establecer el nivel en el que deben enseñarse estas unidades léxicas en el aula de ELE.

Para indagar en dicha hipótesis, se ha analizado en diferentes corpus la frecuencia de 3260 SF extraídas del PCIC, comprobando el grado de correlación de su frecuencia con los niveles establecidos en el MCER, desde A1 a C2. Igualmente, realizamos una propuesta que establezca un índice de frecuencia de referencia para la nivelación de las SF, y complementamos esta propuesta con el análisis de las 100 SF más frecuentes de nuestro inventario atendiendo a otros criterios rentables comunicativamente como son los datos lexicométricos relativos a la frecuencia normalizada, la dispersión y la densidad léxica. Igualmente, atendemos a aspectos lingüísticos como la estructura y la opacidad semántica de las SF, así como la variación diatópica y la modalidad (oral/escrita). Finalmente, analizamos los ámbitos temáticos en los que se emplean estas expresiones en relación con los descriptores de las escalas ilustrativas del MCER y las nociones específicas asignadas en el PCIC.

Para justificar la elección de los corpus debemos hacer una consideración. En primer lugar, se ha recurrido al motor de búsqueda de *Google* para analizar la frecuencia de estas expresiones en la Red. Lindstromberg & Boers (2008) y otros como Buyse (2017), Cruz (2017), González Fernández (2017) o Sinclair (2004) comentan las ventajas y desventajas del empleo de Internet como corpus, destacando entre los inconvenientes su continuo crecimiento, con lo que la información cambia constantemente, o que el buscador no haya sido diseñado desde una perspectiva lingüística, por lo que no está lematizado. No obstante, todos los autores reconocen que parece indiscutible que la Red ofrece un repertorio de textos y datos infinitamente mayor que cualquier otro corpus existente, por lo que su consideración como un gran repertorio textual está ya asumida por la mayoría de los investigadores.

Hemos seleccionado por su fiabilidad los dos corpus de la RAE: el *Corpus de Referencia del Español Actual* (CREA)[13], compuesto por un amplio repertorio de textos de diversa procedencia, tanto escritos (libros, revistas, periódicos) como

[13] http://corpus.rae.es/creanet.html. Puede ampliarse la información sobre el corpus en el manual de consulta, accesible en http://corpus.rae.es/ayuda_c.htm

orales, procedentes en su mayoría de la radio y la televisión, desde 1975 hasta 2004, y que comprende algo más de 160 millones de formas y el *Corpus del Español del Siglo XXI* (CORPES XXI)[14] que, al igual que el CREA, está formado por textos escritos y orales procedentes de España, América, Filipinas y Guinea Ecuatorial, y que en su versión 0.91 (diciembre de 2018) cuenta con más de 285000 documentos que suman alrededor de 286 millones de formas. La otra herramienta es el *Corpus del Español* (CdE) de Mark Davies[15], que consta de dos partes: la primera es el corpus original y más pequeño, que permite buscar cambios históricos y variación de géneros (incluye más de cien millones de palabras procedentes de más de veinte mil textos del español de los siglos XIII al XX), mientras que la segunda es el corpus nuevo, mucho más amplio, y que permite buscar variaciones dialectales. Este segundo contiene casi 5500 millones de palabras de páginas web de veintiún países de habla hispana, y permite hacer búsquedas en textos en español muy recientes, por lo que será el que utilicemos para nuestro estudio.

Por lo que se refiere a las SF seleccionadas para nuestro estudio, las hemos extraído de los catálogos de Nociones Específicas del PCIC (2006), desde el nivel A1 al C2, aunque hemos unido las pertenecientes a los niveles A1 (49) y A2 (69) para que sean estadísticamente representativas. De este modo, nuestro catálogo está formado por 3260 SF (A1-A2 = 118; B1 = 373; B2 = 913; C1 = 952; C2 = 904). Para su selección, hemos adoptado un criterio amplio, incluyendo todos los tipos de SF (colocaciones, locuciones y enunciados fraseológicos) y atendiendo no tanto a sus propiedades semánticas de opacidad o no composicionalidad, como a la rentabilidad comunicativa para el aprendiente. Nuestra decisión se basa en el hecho de que, como señalan diversos autores (Martínez 2013: 187 y ss.), la no composicionalidad de estas unidades léxicas, a pesar de ser un criterio que debe considerarse para complementar otros, es subjetivo, puesto que la opacidad puede depender de factores individuales de los aprendientes, y no solo de las características de las propias SF, y por lo tanto su interpretación y comprensión puede ser diferente en cada aprendiente. Así, una expresión aparentemente transparente puede resultar opaca por diversas razones para ciertos estudiantes.

De acuerdo con estos criterios, hemos seleccionado SF de diversa tipología y características como "café cortado", "hijo natural", "pronóstico reservado", "reproducción asistida", "echar un trago", "entrado en carnes", "ir hecho un adefesio", "navegación de recreo", "pagar a tocateja," "a bombo y platillo", "estar como un niño con zapatos nuevos", "hecha la ley, hecha la trampa", "ojo por ojo y diente por diente".

La recolección de los datos se realizó entre julio y octubre de 2019, y debemos comentar algunos aspectos que inciden en los datos sobre la frecuencia de las SF analizadas. En primer lugar, hemos tenido en cuenta, siempre que ha sido posible, la variación morfológica, tanto de género y número en los sustantivos y adjetivos,

[14] http://www.rae.es/recursos/banco-de-datos/corpes-xxi. Para más información sobre la conformación y codificación del corpus véase https://bit.ly/33Lhx4M

[15] http://www.corpusdelespanol.org

como la flexión verbal. Para ello, en las consultas utilizamos las reglas de etiquetado morfológico y lematización que permiten los propios corpus.

En este sentido, en el caso concreto de *Google*, que no está lematizado, hemos contabilizado la variación de género y número realizando búsquedas individuales para singulares, plurales, masculinos, femeninos y sus combinaciones pero no ha sido posible, por la ausencia de lematización, atender la flexión verbal, por lo que en el caso de las SF cuya base es un verbo, hemos empleado el infinitivo como lema.

Igualmente, por las dimensiones de este estudio, y en ocasiones por la propias posibilidades que en cuanto a la recuperación de SF ofrecen los corpus seleccionados, no se han atendido configuraciones sintácticas concretas como, por ejemplo, en el caso en el que los componentes de la SF no forman una secuencia continua (esto es, cuando entre dichos componentes aparecen modificadores u otros complementos: "mar [un poco, bastante, muy...] picado", "pagar [siempre, a veces...] a escote", "relación [un tanto, a veces, bien, de lo más...] tormentosa"), así como en los posibles cambios de orden sintáctico de los constituyentes de la SF (por ejemplo, la anteposición en ciertas colocaciones: "actitud extraña/extraña actitud", "amigo íntimo/íntimo amigo", "barba poblada/poblada barba", "carácter fuerte/fuerte carácter", "fisonomía extraña/extraña fisonomía", "sensación amarga/amarga sensación", "trabajo ímprobo/ímprobo trabajo"). Por todo ello, deben tenerse en cuenta estas consideraciones para la interpretación de los datos ofrecidos en este trabajo.

Finalmente, para el tratamiento estadístico hemos utilizado el programa *IBM SPSS Statistics* (versión 24), el cual nos ha permitido realizar tanto la descripción estadística descriptiva de los datos recabados como la representación gráfica de los mismos.

4 Estudio y propuesta sobre la frecuencia de las SF

Como ya hemos comentado, la frecuencia es un índice que tradicionalmente se ha empleado para la selección del léxico en ELE, produciéndose una correlación negativa entre frecuencia léxica y nivel, es decir, que las unidades más frecuentes se aprenden en niveles más bajos y viceversa.

Por tanto, en nuestro estudio comprobamos en primer lugar si se produce dicha correlación negativa entre la frecuencia de las SF analizadas y el nivel asignado en el PCIC, tomando como base los datos extraídos de los corpus analizados. Tras esto, realizaremos un análisis descriptivo y comparativo de dicha frecuencia y su distribución en los corpus estudiados para, finalmente, presentar una propuesta de distribución de las SF en función de la frecuencia.

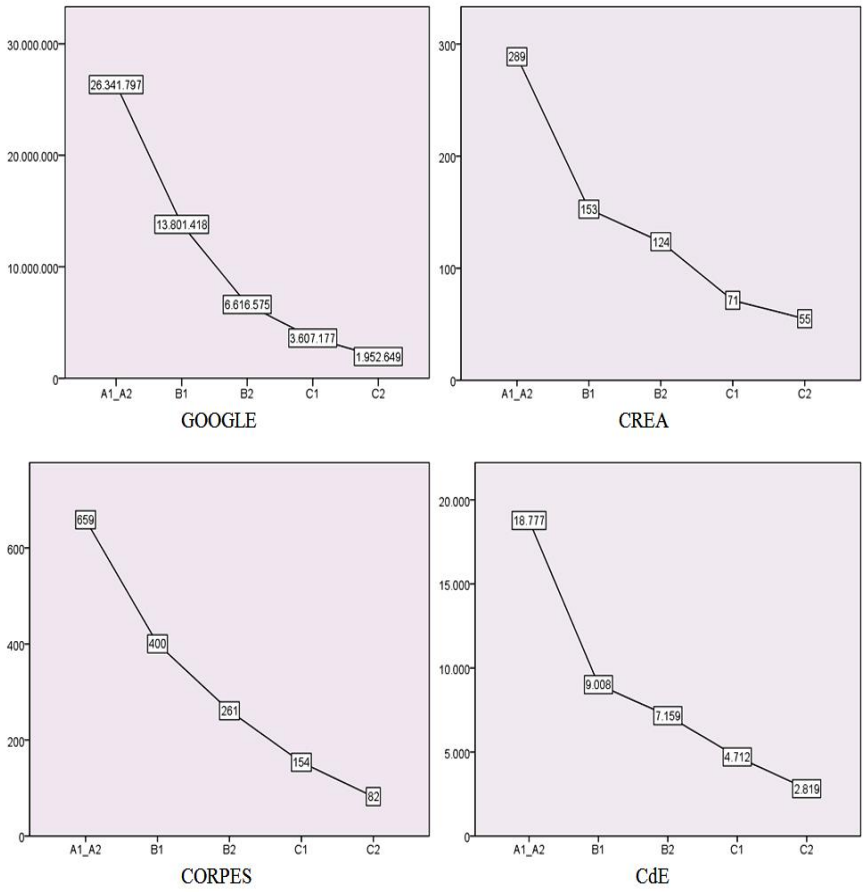
Los datos obtenidos en cuanto a la relación entre frecuencia y nivel se reflejan en la Tabla 1:

Tabla 1. Estadística descriptiva. Datos generales

	A1-A2 (n=118)	**B1 (n=373)**	**B2 (n=913)**	**C1 (n=952)**	**C2 (n=904)**
GOOGLE					
media	26341796,61	13801417,69	6616574,93	3607177,44	1952648,73
desv. est.	62994255,45	31141526,62	17767531,96	12762003,91	9383841,71
mediana	7388000	3244000	1260000	451500	137000
mín.	168000	12100	4	119	0
1er cuart.	2446500	761000	364250	85625	21900
3er cuart.	19400000	13739000	5990000	2568250	840950
máx.	498000000	316000000	276000000	295000000	217000000
CREA					
media	288,73	152,60	123,58	71,35	54,86
desv. est.	744,6	538,17	312,70	145,67	205,54
mediana	59	39	35	21	11
mín.	0	0	0	0	0
1er cuart.	18	13	9	5	2
3er cuart.	205,75	91	106	74,75	38,25
máx.	5649	7998	5427	2078	4274
CORPES					
media	658,65	400,41	260,98	153,57	82,22
desv. est.	1326,40	1141,89	791,54	353,91	315,32
mediana	174	83	63	40	15,5
mín.	0	0	0	0	0
1er cuart.	51,25	25	16	9	2
3er cuart.	600,75	278	210	149,5	63
máx.	7293	13687	17281	3776	6106
CdE					
media	18777,28	9007,52	7159,28	4711,68	2818,73
desv. est.	59224,42	32437,44	33007,67	20675,79	14253,77
mediana	2244,5	1177	1041	488	201
mín.	1	0	0	0	0
1er cuart.	517	288	195	92,5	31
3er cuart.	6506	4642	4333	2376,5	1017
máx.	476454	435498	801302	365368	247772

En cuanto a la relación entre la frecuencia y la nivelación, se percibe una evidente correlación negativa entre la frecuencia de las SF y el nivel asignado en el PCIC, es decir, que a mayor frecuencia, menor nivel y viceversa. Esto se evidencia en todos los datos estadísticos más robustos como la media, la mediana, y los cuartiles, y sin considerar mínimos y máximos, pues no son representativos. Tomando como referencia la frecuencia media por niveles, en la Figura 1 se aprecia ostensiblemente dicha correlación negativa en todos los corpus.

Figura 1. Media de frecuencias por nivel y por corpus



La correlación negativa entre el nivel asignado en el PCIC a las SF de nuestra muestra y su frecuencia en los corpus manejados queda confirmada estadísticamente por el coeficiente τb de Kendall, que establece la correlación entre dos rangos (en nuestro caso, el nivel y la frecuencia), puesto que se aprecia una correlación negativa en todos los corpus: *Google* (τb = -0,342), CREA (τb = -0,192), CORPES (τb = -0,243) y CdE (τb = -0,208). Aunque el valor de la correlación sea bajo, sí que es estadísticamente significativo ($p<0,0001$), y además, dicho grado de correlación podría deberse, como comentaremos con más detalle a continuación, al tipo de distribución de las frecuencias en cada grupo.

Por tanto, se confirma lo que ya apuntamos en trabajos anteriores (Martos & Contreras 2018, Contreras & Martos 2020): la nivelación realizada por el PCIC, a pesar de basarse en un criterio intuitivo, es bastante fiable, puesto que a medida que desciende la frecuencia de las SF, más alto es el nivel asignado.

4.1 Grado de coincidencia entre la frecuencia y el nivel asignado en el PCIC

Diversos autores (Pérez 2015: 229, Velázquez 2018: 22) analizan el tratamiento de las SF en el PCIC, que incluye exponentes fijos (expresiones idiomáticas y frases hechas) desde los niveles iniciales, aunque es en los niveles superiores donde

reciben mayor atención (2006, Introducción, I: 208), admitiendo que no hace referencia explícitamente a estas unidades en los niveles iniciales (A1 y A2) debido a que la falta de conocimientos puede impedir a los aprendientes de estos primeros niveles analizar los fraseologismos (PCIC, 2006: III, 173).

Desde nuestro punto de vista, estimamos que esta decisión puede presentar ciertos inconvenientes, en primer lugar desde el punto de vista del estudiante, pues las SF están presentes en el proceso de aprendizaje desde el primer momento, sin que el alumno deba ser consciente de su complejidad formal o su pertenencia a la fraseología. Además, como señala Timofeeva (2013: 331), debido a la capacidad holística que se tiene en la lengua materna, los alumnos, incluso en los niveles iniciales, suelen detectar con éxito estas SF en un texto adecuado a su grado de competencia, aun sin saber qué significan. En segundo lugar, desde el punto de vista del docente, el PCIC debería ofrecer indicaciones claras acerca de qué unidades pluriverbales es necesario adquirir en los niveles del MCER (A1-C2) para que los profesores puedan organizar estos contenidos (López 2011: 534-5).

En este sentido, hay que tener en cuenta que, tal y como se indica en el PCIC, sus inventarios deben considerarse abiertos y no exhaustivos "en cuanto que siempre podrían añadirse nuevos elementos, tanto en sentido horizontal (completando las series), como en sentido vertical (con nuevos elementos en cada nivel)" (PCIC, 2006: 393-394). Por tanto, es labor de los programadores, profesores y creadores de materiales adaptar estos inventarios a las necesidades de los estudiantes y a la situación particular de enseñanza, así como añadir nuevos elementos que los completen. A pesar de esta falta de exhaustividad, estos catálogos constituyen una ayuda inestimable para determinar qué llevar al aula en los diferentes niveles, pues, tal como se reconoce en su introducción, es "una base de orientación general que podrá servir de referencia a quien lo utilice a la hora de seleccionar y distribuir por niveles las unidades léxicas que precise para sus propios fines" (PCIC, 2006: 333).

Compartimos con la mayoría de investigadores sobre el tema la opinión de que es fundamental que estas SF se trabajen en el aula desde los niveles iniciales, tanto por la rentabilidad comunicativa que hemos comentado anteriormente, como para el desarrollo de las redes léxicas (Ferrando 2012, Gómez 2004, Higueras 2004, Leontaridi, Ruiz & Peramos 2011, Penadés 1999, Velázquez 2018), aunque su estudio deberá intensificarse en los niveles avanzados con el objetivo de aumentar la precisión lingüística del aprendiente.

Así, con el objetivo de profundizar en la posible relación entre la frecuencia de las SF y el nivel asignado en el PCIC, y por tanto seguir confirmando si este es un criterio válido para la nivelación de estas unidades léxicas, en primer lugar tomamos las 20 SF más comunes en los cuatro corpus para comprobar la coincidencia entre ellos.

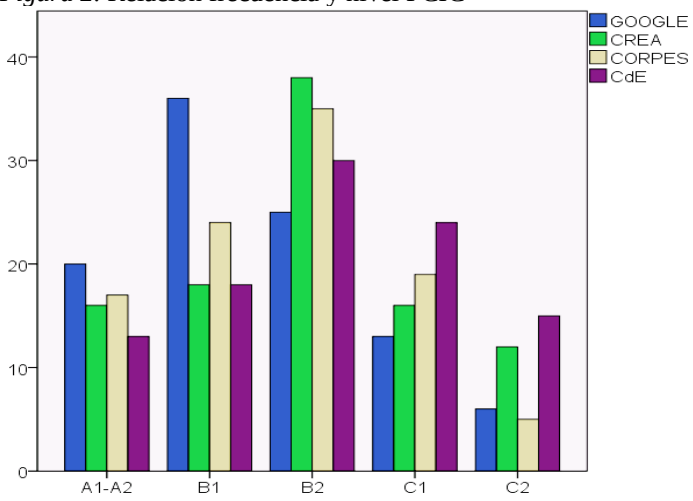
Entre estas 20 unidades más frecuentes, en total hay 57 SF distintas, y se confirma una amplia concurrencia en los cuatro corpus, pues 21 coinciden entre las 20 más frecuentes en al menos dos corpus. Aunque solo "medios de comunicación" aparece en los cuatro, ocho aparecen en tres de los corpus ("base de datos", "correo electrónico", "derechos humanos", "página web", "partido político", "primer

ministro", "proyecto de ley", "rueda de prensa") y 12 en al menos dos de ellos ("cambio climático", "centro comercial", "comunidad autónoma", "en directo", "en vivo", "fuerzas armadas", "guerra civil", "materia prima", "nuevas tecnologías", "poder judicial", "recursos naturales", "tipo de interés").

Asimismo, hemos comprobado que el resto de SF que no coinciden dentro de estas 20 más frecuentes sí aparecen entre las 100 primeras en los otros corpus, lo que vuelve a confirmar la amplia coincidencia en los cuatro corpus en cuanto a las SF con un grado de frecuencia más elevado.

En segundo lugar, y por lo que se refiere a la relación entre la frecuencia en los corpus y el nivel asignado por el PCIC, hemos analizado las 100 primeras SF más frecuentes en los corpus, entendiendo que dicho nivel de frecuencia debe corresponder a los niveles A1-A2 en el PCIC. Aparentemente, como se observa en la Figura 2, se aprecia una tendencia a asignar niveles superiores a estas expresiones con un alto grado de frecuencia, ya que en todos los corpus el porcentaje de ellas que se incluyen en el nivel A1-A2 del PCIC es bajo. Si subimos un nivel más (B1), este porcentaje se incrementa, pero en todos los casos, el porcentaje de SF en los niveles altos (B2-C2) es muy elevado (44% *Google*, 66% CREA, 59% CORPES, 69% CdE).

Figura 2. Relación frecuencia y nivel PCIC



Estas amplias diferencias entre los niveles establecidos en el PCIC y el grado de frecuencia pueden deberse, como ya señalan García & Alonso (2018: 165-6), a causas como que la SF sea propia de un registro marcado, con una alta frecuencia en textos especializados en diversos ámbitos temáticos, por lo que su dispersión, y por tanto su rentabilidad comunicativa es menor.

En nuestro caso, hemos podido confirmar esta hipótesis comparando, en el CORPES XXI, los datos de estas 100 SF relativos a la frecuencia normalizada

general con la del tema específico en la que mayor frecuencia presentan[16], comprobando así que las expresiones muy frecuentes propias de niveles avanzados presentan una frecuencia extrema en temas especializados como la política, la economía y la justicia, y algo menos en ámbitos temáticos como la ciencia, la tecnología, las artes, la cultura y los espectáculos.

De este modo, entre los primeros (política, economía y justicia), se encuentran "cámara de comercio" (B2: 5,26/17,99)[17], "campaña electoral" (B2: 10,31/43,50), "derechos humanos" (B2: 56,01/201,62), "comercio exterior" (B2: 5,79/24,61), "crisis financiera" (B2: 5,38/22,20), "elecciones generales" (B2: 5,31/24,12), "elecciones municipales" (B2: 3,66/16,54), "entidad financiera" (C1: 5,29/ 22,28), "estado de derecho" (C1: 7,63/30,21), "fuerzas armadas" (C2: 20,44/79,01), "fuerzas de seguridad" (C1: 7,13/26,72), "gobierno central" (B2: 9,27/33,75), "libre comercio" (C2: 10,43/46,95), "poder judicial" (B2: 14,08/60,77), "policía nacional" (C1: 13,19/41,79), "proyecto de ley" (C2: 16,91/55,19), "rueda de prensa" (B2: 17,24/39,45), "tipo de cambio" (C2: 7,16/34,75), "tipo de interés" (C2: 4,65/20,52).

Por lo que se refiere a las ciencias y tecnología, ejemplos de estas SF muy frecuentes en este ámbito son "nuevas tecnologías" (B2: 16,83/58,64), "energía eléctrica" (B2: 10,87/51,35), "disco duro" (B2: 5,41/37,30), "base de datos" (B2: 16,47/74,99), "energía solar" (B2: 5,13/42,82), "sistema operativo" (C1: 11,17/97,66), "teléfono celular" (C1: 6,10/20,18). Por último, expresiones con frecuencias absolutas muy altas en artes, cultura y espectáculos: "banda sonora" (B2: 3,74/27,82), "en directo" (C1: 11,23/45,27), "en vivo" (C1: 12,65/ 67,69), "obra maestra" (C1: 6,15/30,01), "primer plano" (B2: 5,40/14,92).

Así pues, entendemos que estas cifras confirman que la alta frecuencia de ciertas SF se debe a su pertenencia a un registro marcado temáticamente, por lo que tanto su dispersión general como su rentabilidad comunicativa son menores y, por tanto, su inclusión en niveles superiores, tal y como se realiza en el PCIC, parece acertada.

Otra posible causa de lo que comentamos es la opacidad semántica de algunas expresiones, y así, tal y como señalan tanto el PCIC como el MCER, se incluyen en los niveles altos las SF que, a pesar de ser frecuentes no son transparentes, pues su significado no es composicional, como podría ocurrir en casos como como "alto el fuego" (C2), "cámara de comercio" (B2), "dar a luz" (B2), "efectos secundarios" (B2), "encogerse de hombros" (C1), "estado de derecho" (C1), "golpe de estado" (B2), "guerra fría" (C1), "moción de censura" (C2), "tipo de cambio" (C2), "tipo de interés" (C2). En estos casos, como hemos comentado, sí que estimamos que, frente a la posible opacidad, que no deja de ser una característica subjetiva, pues depende no tanto de las propias SF sino de factores individuales, distancia entre la

[16] La frecuencia normalizada indica el número de apariciones de la expresión por millón de palabras del corpus, mientras que la frecuencia absoluta muestra el número de apariciones de la unidad en el total de los materiales que forman la base del corpus, por lo que estimamos que la conjunción de ambos índices puede contribuir a una mejor comparación de la frecuencia de las SF analizadas.

[17] Entre paréntesis, indicamos el nivel asignado en el PCIC, frecuencia normalizada general/frecuencia normalizada en el ámbito temático específico.

lengua materna y la LE, etc., y por lo tanto su interpretación y comprensión puede ser diferente en cada aprendiente, debería primar, por su posible rentabilidad comunicativa, la frecuencia de uso, por lo que tal vez estas SF deberían incluirse en niveles inferiores, siempre en función de las necesidades y contextos de enseñanza.

4.2 Propuesta de distribución de las SF atendiendo a la frecuencia

En un trabajo anterior (Contreras & Martos 2020) proponíamos un posible punto de corte para decidir a partir de qué frecuencia se podría asignar un nivel u otro a cada SF, y ahora ampliamos los datos y cálculos que completen y apoyen dicha propuesta.

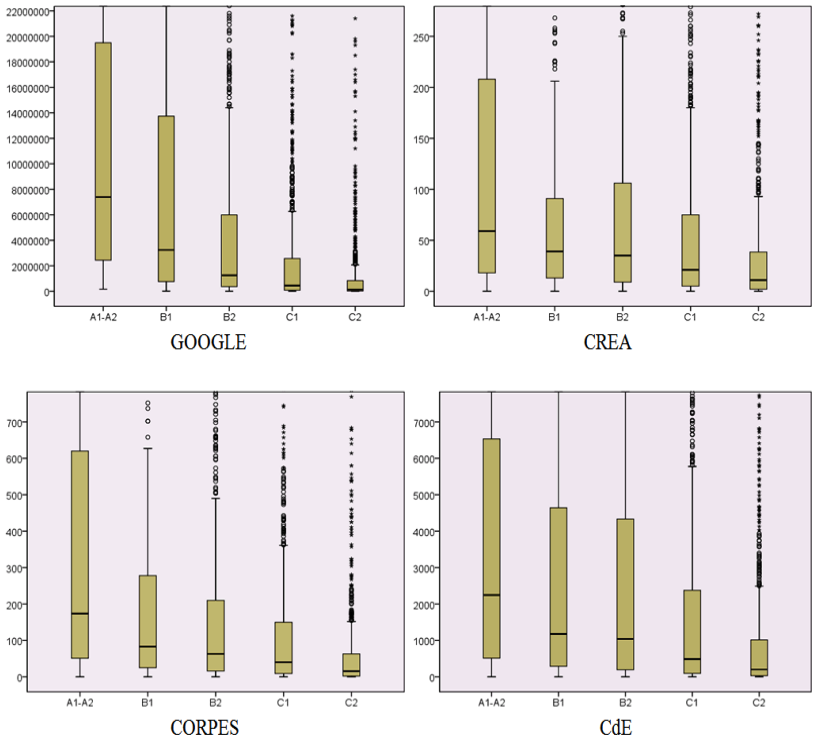
Como se aprecia en los gráficos de la Figura 1, en los que se incluye la media de frecuencias por nivel y por corpus, las frecuencias no presentan una distribución simétrica (que se da cuando la media, moda y mediana coinciden), sino que existe una gran asimetría, en este caso de tendencia paretiana, es decir, que el 20% de las observaciones arrojan un valor superior al 80% restante de la muestra. Estas cifras confirman la ley de Zipf (1935), que establece que en cualquier conjunto analizado, un reducido número de unidades léxicas presenta altos índices de frecuencia, mientras que un gran número son poco frecuentes.

Para confirmar si esta ley se cumple también en el caso de las SF, tomamos como ejemplo el nivel B2 en cada corpus. Así, si sumamos las ocurrencias del 20% de las más frecuentes y el resultado lo comparamos con la suma de las ocurrencias del 80% restante, se confirma ampliamente esta tendencia: *Google* (20% = 4753397000 / 80% = 1248905894), CREA (20% = 86837 / 80% = 25992), CORPES (20% = 188379 / 80% = 49642), CdE (20% = 5634815 / 80% = 901609). Además, hemos comprobado que dicho carácter se mantiene sistemáticamente en la distribución de las frecuencias en cada uno de los niveles y en todos los corpus.

Por lo que se refiere a la posibilidad de establecer puntos de corte en las frecuencias para la nivelación de estas unidades, en casos con este tipo de distribución se considera más apropiado tomar como referencia la moda o la mediana, más que otros estadísticos como la media o los cuartiles extremos, ya que estos últimos podrían distorsionar la descripción de la distribución del conjunto. Teniendo en cuenta que en nuestro estudio no podemos tomar como referencia la moda, ya que en muchos casos no hay repeticiones en el número de ocurrencias de las SF en los corpus, nos fijaremos en la mediana que, al representar el punto medio de la distribución de las frecuencias, nos parece un punto adecuado como valor central de referencia de cada nivel.

En la Figura 3, en la que representamos gráficamente la distribución de frecuencias por nivel en cada corpus, podemos apreciar tanto la distribución paretiana a la que venimos haciendo referencia, como otro hecho igualmente relevante para lo que estamos tratando: aunque en cada nivel parece existir un rango de frecuencias agrupado en torno al valor central de la mediana, se percibe claramente una amplia superposición de los otros límites (mínimo, máximo y cuartiles extremos) con los de los niveles anteriores y posteriores.

Figura 3. Distribución de frecuencias por nivel y por corpus

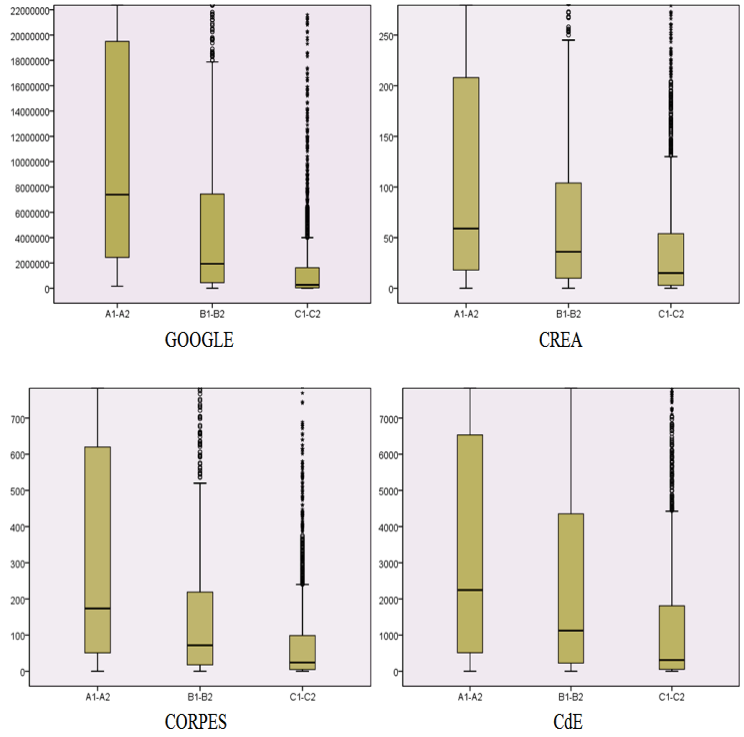


Esto, en principio, complicaría el intento de establecer puntos de corte precisos en la frecuencia para nivelar las SF, por lo que como hemos venido comentando, es necesario atender a otros criterios (distribución, rentabilidad comunicativa, composicionalidad, necesidades de los alumnos) para determinar el nivel concreto de una SF dada. No obstante, y a falta de análisis más profundos, creemos que nuestro estudio puede ofrecer orientaciones generales sobre la frecuencia de referencia en cada nivel.

En este sentido, ya Lindstromberg & Boers (2008) realizaron una propuesta semejante tomando como base la frecuencia de SF en *Google* y estableciendo tres niveles de frecuencia, alto, medio y bajo[18], macroniveles que nos parecen muy productivos como primera referencia a la hora de nivelar una SF, estimando como frecuentes las SF en el rango A1-A2, de un nivel medio de frecuencia las propias de B1-B2, y por último las de C1-C2, que presentarían una frecuencia baja, tal y como se aprecia en la Figura 4.

[18] Debemos aclarar que existen grandes diferencias entre los datos de frecuencia de las SF en *Google* recogidos por Lindstromberg & Boers (2008) y los arrojados por nuestro estudio, hecho que debemos atribuir sin duda al inmenso crecimiento de la información en Internet en los últimos años: según *World Wide Web Size* (https://www.worldwidewebsize.com/) e *Internet Live Stats* (https://www.internetlivestats.com/) solo en cinco años, entre 2008 y 2013, la cantidad de información creció un 900%, alcanzando en 2016 la cifra de un Zettabyte (1 zetabyte = dos billones de gigabytes), y se espera que para 2021 alcancemos los 3,3 Zettabytes.

Figura 4. Distribución de frecuencias por nivel y por corpus (macroniveles)



Con este propósito, esto es, para ofrecer un punto de referencia, que no de corte, para la nivelación de una SF dada, presentamos en la Tabla 2 los datos extraídos de nuestra muestra tomando como eje central la mediana de cada grupo, puesto que este dato estadístico, como hemos repetido, parece más fiable en distribuciones con tendencia paretiana, aunque también ofrecemos el dato de la media para poder comparar las diferencias entre ambas cifras.

Tabla 2. Frecuencias macroniveles

	GOOGLE		**CREA**		**CORPES**		**CdE**	
	mediana	**media**	**mediana**	**media**	**mediana**	**media**	**mediana**	**media**
A1-A2 (n=118)	7388000	26341796,6	59	288,73	174	658,65	2244,5	18777,28
B1-B2 (n=1286)	1932000	8711920,01	36	131,99	72	301,45	1125	7695,35
C1-C2 (n=1856)	271000	2801307,86	15	63,31	24,5	118,71	310,5	3798,16

De este modo, un profesor que necesite determinar el nivel en el que adscribir una determinada SF para presentarla y trabajarla en el aula, dispondrá de una cifra aproximada (la mediana) con la que, en función del recurso que emplee para conocer su frecuencia (*Google*, CREA, CORPES, CdE), podrá decidir si la expresión es propia de niveles iniciales (A1-A2), medios (B1-B2) o avanzados (C1-C2). Incluso, tomando como referencia los datos incluidos en la Tabla 1, podría

concretar algo más el nivel de la SF, siempre con las limitaciones que hemos expuesto.

4.3 Criterios complementarios a la frecuencia léxica

Como hemos podido ver, a pesar de no precisar un índice de frecuencia concreto para cada nivel, ya que como hemos avanzado estos datos deben tomarse como punto de referencia y no de corte, estimamos que pueden servir de base para la nivelación de las SF, siendo necesario complementarlos con otros criterios para la determinación más precisa del nivel de las SF que el profesor debe seleccionar para su trabajo en el aula, atendiendo igualmente a la rentabilidad comunicativa, los intereses y las necesidades de los alumnos.

Conscientes de todo ello, pero teniendo en cuenta los objetivos y dimensiones de este trabajo, y por tanto a modo de ejemplo y orientación, complementamos los datos de la frecuencia léxica absoluta con el análisis de las 100 SF más frecuentes atendiendo a otros criterios rentables comunicativamente como son los datos lexicométricos relativos a la frecuencia normalizada, la dispersión y la densidad léxica. Igualmente, atenderemos a aspectos lingüísticos como la estructura de las SF, así como la variación diatópica y la modalidad (oral/escrita). Finalmente, en consonancia con el enfoque nociofuncional adoptado por el PCIC, tendremos en cuenta un aspecto fundamental para la nivelación de estos exponentes léxicos como son los ámbitos temáticos en los que se emplean, relacionados con los descriptores de las escalas ilustrativas del MCER y las nociones específicas asignadas en el PCIC.

Para ello, nos serviremos de los datos obtenidos en el CORPES XXI, ya que este corpus presenta una base documental actual y permite consultas de estadística más avanzadas, así como distinciones de tipo diatópico y modal (oral y escrito). Igualmente, aunque por cuestiones de espacio no lo analizamos, este corpus posibilita otras consultas en función del tema, tipología textual y, en el caso de los documentos orales, permite distinguir en función del sexo, edad y nivel de estudios, con la interesante información de tipo sociolingüístico que esto puede aportar.

Como hemos adelantado, un criterio que puede complementar la frecuencia absoluta de las SF es la denominada "frecuencia normalizada", esto es, el número de casos por millón de palabras, ya que dicho dato puede ayudarnos a tener una visión más clara y precisa de la frecuencia absoluta, pues esta última puede moverse en cifras muy amplias. Así, la frecuencia normalizada de las 100 SF tomadas como muestra para este análisis oscila entre el 1,25 de "moción de censura" (C2) y el 56,01 de "derechos humanos" (B2), aunque el promedio es de 10,06. De modo general, se aprecia cierta tendencia a una menor frecuencia normalizada en SF de niveles superiores, lo que seguiría confirmando la utilidad de la frecuencia para la nivelación de estas expresiones: "moción de censura" (C2: 1,25), "entidad bancaria" (C1: 2,14), "abogado defensor" (B2: 2,15), "estación de servicio" (C1: 2,68), "libertad de prensa" (B2: 2,68), "alto el fuego" (C2: 2,98), "policía municipal" (C1: 3,36), "poder legislativo" (B2: 3,54), "elecciones municipales" (B2: 3,66), "banda sonora" (B2: 3,74), "justicia social" (B2: 3,91), "consejero

delegado" (C2: 4,16). Igualmente, aunque en menor grado, las SF de un nivel más bajo tienen tendencia a presentar una frecuencia normalizada mayor: "correo electrónico" (A1: 15,18), "materia prima" (B1: 17,59), "página web" (A1: 20,51), "seguridad social" (A2: 20,65), "recursos naturales" (B1: 22,07), "primer ministro" (A2: 23,80), "partido político" (A2: 25,33), "cambio climático" (B1: 25,38), "comunidad autónoma" (B1: 28,18), "medios de comunicación" (B1, 49.44).

Otro concepto clave en la estadística lexicométrica a la que acudimos es la dispersión, esto es, la relación entre la frecuencia absoluta y el número y tipo de documentos, ya que una unidad léxica puede tener una mayor frecuencia que otra debido a que se repite muchas veces en un grupo reducido de textos, por lo que su rentabilidad comunicativa puede ser menor. Para obtener un índice más comprensible y útil, hemos dividido el número de casos de aparición de cada SF entre el número de documentos en los que aparece, por lo que estaríamos atendiendo a algo semejante a la "densidad léxica", es decir, el número de palabras diferentes que se encuentran en cada texto. Aquí, las cifras oscilan entre 1,15 ("rueda de prensa", B2), y 4,51 ("aceite de oliva", B1), apreciándose una tendencia similar a la que hemos comentado para la frecuencia normalizada, esto es, el número de casos por documento es inversamente proporcional al nivel. Así, a mayor nivel, menor número de casos por documento: "rueda de prensa" (B2: 1,5), "entidad bancaria" (C1: 1,19), "banda sonora" (B2: 1,23), "obra maestra" (C1: 1,26), "abogado defensor" (B2: 1,26), "medida de seguridad" (B2: 1,27), "justicia social" (B2: 1,29), "tercera edad" (B2: 1,30), "en vivo" (C1: 1,31); mientras que a menor nivel, más número de casos por documento: "relaciones sexuales" (B1: 1,98), "partido político" (A2: 2,05), "seguridad social" (A2: 2,16), "persona mayor" (A2: 2,22), "comunidad autónoma" (B1: 2,22), "recursos naturales" (B1: 2,33), "cuarto de baño" (A1: 2,41), "vino blanco" (A1: 2,60), "hacer el amor" (B1: 2,62), "cambio climático" (B1: 2,83), "aceite de oliva" (B1: 4,51).

De este modo, tanto la frecuencia normalizada como el número de casos por documento también pueden ser índices que contribuyan a completar y modular la frecuencia absoluta, confirmando la correlación inversa entre los datos lexicométricos relativos a la frecuencia y el nivel de competencia que debe asignarse a la SF en cuestión.

Por lo que se refiere a los aspectos puramente lingüísticos, y tomando como referencia la clasificación de Corpas (1996), quien, partiendo de una concepción más amplia de la fraseología, las divide en colocaciones, locuciones y enunciados fraseológicos, la inmensa mayoría de estas 100 SF son colocaciones ("abogado defensor", "ama de casa", "banda sonora", "cambio climático", "derechos humanos", "energía solar", "libertad de expresión", "página web", "proyecto de ley", "sistema operativo", "tiempo libre", etc.), con un número muy reducido de locuciones ("alto el fuego", "dar a luz", "en directo", "en vivo", "encogerse de hombros", "hacer el amor") y ningún enunciado fraseológico. Esto podría confirmar las opiniones que defienden relegar las SF más complejas a niveles más avanzados, aunque repetimos que podrían enseñarse desde los primeros estadios de

aprendizaje si atendemos a las necesidades comunicativas de los alumnos y a la rentabilidad comunicativa de estas SF, y no a su complejidad formal o semántica.

En ocasiones, dependiendo del contexto y las necesidades de enseñanza, al profesor de ELE le será útil y necesario disponer de datos sobre la distribución diatópica de las SF susceptibles de ser seleccionadas. En este caso, los corpus lingüísticos también suponen una inestimable fuente de información, ya que podemos distinguir entre los exponentes que tienen una frecuencia de uso semejante en los dos grandes ámbitos diatópicos de nuestro idioma (España e Hispanoamérica) como "centro comercial" (A1: 7,10/8,71)[19], "clase política" (C1: 4,74/4,81), "clase social" (B2: 6,72/6,94), "correo electrónico" (A1: 16,83/14,42), "dar a luz" (B2: 6,08/6,06), "energía solar" (B2: 5,40/5,03), "entidad bancaria" (C1: 2,79/2,14), "ministro de educación" (B1: 6,59/6,09), "ropa interior" (B1: 5,81/5,28). Asimismo, podemos comprobar que ciertas expresiones son más comunes en España, aunque su frecuencia en Hispanoamérica puede ser igualmente más o menos alta, por lo que se trataría más bien de tendencias en su distribución geolectal: "aceite de oliva" (B1: 27,43/8,74), "comunidad autónoma" (B1: 80,92/0,65), "consejero delegado" (C2: 10,37/1,02), "cuarto de baño" (A1: 12,02/1,59), "encogerse de hombros" (C1: 11,81/4,78), "en directo" (C1: 21,27/6,06), "fuerzas de seguridad" (C1: 10,43/5,41), "primer ministro" (A2: 33,16/19,03), "puesto de trabajo" (B1: 20,57/8,74), "teléfono móvil (A1: 23,55/6,38), "tipo de interés" (C2: 11,80/0,95). Finalmente, hallamos SF de mayor frecuencia en Hispanoamérica, aunque con presencia igualmente variable en España: "derechos humanos" (B2: 27,18/70,64), "educación superior" (B2: 3,90/19,79), "en vivo" (C1: 4,68/16,79), "fuerzas armadas" (C2: 11,35/25,13), "libre comercio" (C2: 4,13/13,72), "recursos naturales" (B1: 7,35/29,64), "teléfono celular" (C1: 0,37/0,07), "tipo de cambio" (C2: 2,12/9,73).

De la misma forma, podemos obtener datos relativos a la modalidad (oral y/o escrita) en la que es empleada una expresión, y así encontramos SF con una frecuencia de uso similar en ambas modalidades[20]: "ama de casa" (A1: 4,00/4,49), "cámara de comercio" (B2: 3,37/3,36), "clase política" (C1: 4,78/4,71), "educación pública" (B2: 4,92/4,71), "iglesia católica" (B1: 10,37/11,45), "institución pública" (C2: 7,28/7,18), "justicia social" (B2: 3,92/3,36), "policía nacional" (C1:

[19] Entre paréntesis, indicamos el nivel según el PCIC y la frecuencia normalizada en España/frecuencia normalizada en Hispanoamérica.

[20] Incluimos entre paréntesis el nivel establecido en el PCIC y la frecuencia normalizada en la modalidad escrita/frecuencia normalizada en la modalidad oral. No obstante, debemos tomar estos datos con cierta cautela, ya que el CORPES XXI incluye un porcentaje mucho más amplio de textos escritos (el 90% del total) que de textos orales (el 10% del total), aunque la frecuencia normalizada se refiere al número de apariciones por millón de palabras en cada modalidad, por lo que estimamos que dicho índice puede tomarse como referencia. Por otro lado, también debemos tener en cuenta que entre los géneros discursivos orales presentes en este corpus (debate, discurso, entrevista, magacines y variedades, retransmisiones deportivas, sorteos y concursos, tertulia y otros) no se encuentra la conversación coloquial, género discursivo de suma importancia en la enseñanza y aprendizaje de lenguas, y que podría servir para una distinción más precisa entre el registro formal y el informal o coloquial en el uso de las SF.

13,19/13,25), "sentido del humor" (B1: 5,56/4,04), "tipo de cambio" (C2: 7,17/6,51); SF con una tendencia mayor en la modalidad escrita: "abogado defensor" (B1: 2,17/0,22), "artes plásticas" (B2: 4,90/0,22), "crisis financiera" (B2: 5,44/1,79), "educación superior" (B2: 14,51/3,81), "libre comercio" (C2: 10,53/3,81), "materia prima" (B1: 17,75/6,29), "obra de arte" (B1: 12,76/2,24), "poder político" (B2: 7,11/2,02), "recursos naturales" (B1: 22,34/4,04), "tercera edad" (B2: 4,31/1,79); y expresiones con mayor frecuencia en lo oral: "comunidad autónoma" (B1: 21,23/499,61), "formación profesional" (B2: 4,94/35,49), "medios de comunicación" (B1: 48,88/87,38), "ministro de defensa" (B1: 6,24/52,79), "partido político" (A2: 24,75/64,69), "poder judicial" (B2: 12,86/97,27), "puesto de trabajo" (B1: 12,35/39,76), "rueda de prensa" (B2: 16,71/53,01), "seguridad social" (A2: 18,86/141,52), "tipo de interés" (C2: 4,33/26,28).

Finalmente, atendiendo los ámbitos temáticos en los que se emplean estas SF, los relacionaremos con los descriptores de las escalas ilustrativas del MCER y las nociones específicas asignadas en el PCIC. De este modo, por lo que se refiere a la riqueza del léxico, ya hemos comentado que al A1 le corresponde un repertorio léxico básico, tanto de palabras como de frases aisladas, propio de situaciones concretas, mientras que al A2, también con un repertorio limitado, le corresponde el léxico de entornos cotidianos y habituales para cumplir necesidades concretas y cotidianas. Esto se evidencia en nuestro estudio de las 100 SF más frecuentes del total del inventario de expresiones recopiladas, ya que en estos niveles iniciales los temas más frecuentes, según el repertorio de nociones específicas del PCIC, encontramos temas básicos y cotidianos como la vivienda, los viajes, el alojamiento y el transporte ("aire acondicionado", A1; "cuarto de baño", A1), el trabajo ("ama de casa", A1), las compras, las tiendas y los establecimientos ("centro comercial, A1), la identidad personal ("correo electrónico", A1; "tarjeta de crédito", A1), las características físicas ("ojos azules, A1), la información y los medios de comunicación ("página web", A1, "correo electrónico", A1; "teléfono móvil", A1), la alimentación ("vino blanco", A1), las actividades artísticas ("obra de teatro", A2), el gobierno, la política y la sociedad ("partido político", A2; "primer ministro", A2), los servicios ("persona mayor", A2; "seguridad social", A2), el ocio, el tiempo libre y el entretenimiento ("tiempo libre", A2).

Por su parte, en el nivel B1 los descriptores se refieren al léxico propio de la mayoría de los temas de la vida diaria relativos a la familia, aficiones e intereses, trabajo viajes y hechos de la actualidad, mientras que el B2 incluye los asuntos relacionados con la especialidad del aprendiente y sobre temas más generales. De nuevo, estos descriptores son confirmados en nuestro estudio, pues se observa una clara especialización en las temáticas más básicas, con una fuerte tendencia a incluir expresiones sobre política y cuestiones sociales: gobierno, política y sociedad ("guerra civil", B1; "guerra mundial", B1; "historia del arte", B1; "jefe de estado", B1; "ministro de defensa", B1; "abogado defensor", B2; "campaña electoral", B2; "elecciones generales", B2; "justicia social", B2; "poder legislativo", B2), economía e industria ("materia prima", B1; "puesto de trabajo", B1; "recursos naturales", B1; "cámara de comercio", B2; "comercio exterior", B2; "crisis

financiera", B2), geografía y naturaleza ("cambio climático", B1; "desarrollo sostenible", B2), salud e higiene ("centro de salud", B1; "efectos secundarios", B2), individuo, dimensión física, dimensión perceptiva y anímica ("hacer el amor", B1; "relaciones sexuales", B1; "mala suerte", B1; "sentido del humor", B1; "dar a luz", B2), servicios ("servicios sociales", B1; "medida de seguridad", B2; "tercera edad", B2), actividades artísticas ("artes plásticas", B2; "banda sonora", B2; "primer plano", B2), ciencia y tecnología ("base de datos", B2; "disco duro", B2), educación ("educación pública", B2; "educación superior", B2; "formación profesional", B2), información y medios de comunicación ("libertad de expresión", B2; "libertad de prensa", B2; "rueda de prensa", B2).

Finalmente, el nivel C1 presenta más riqueza y precisión mediante el buen dominio de un amplio repertorio léxico, en el que se incluye el uso de expresiones idiomáticas en registros formales y coloquiales. Por su parte, el C2, con la variedad y calidad propia de un repertorio léxico muy amplio que incluye expresiones idiomáticas y coloquiales y los aspectos connotativos del significado, es semejante al del hablante nativo culto. Muestra de ello son las numerosas SF sobre gobierno, política y sociedad ("clase política", C1; "código penal", C1; "estado de derecho", C1; "guerra fría", C1; "alto el fuego", C2; "fuerzas armadas", C2; "institución pública", C2; "moción de censura", C2; "proyecto de ley", C2) y cuestiones más específicas en campos como la información y los medios de comunicación ("en directo", C1; "teléfono celular", C1), las actividades artísticas ("en vivo", C1; "obra maestra", C1), los servicios ("entidad bancaria", C1; "entidad financiera", C1; "fuerzas de seguridad", C1; "policía municipal", C1), la ciencia y tecnología ("sistema operativo", C1), el trabajo ("consejero delegado", C2) o la economía e industria ("libre comercio", C2; "tipo de cambio", C2; "tipo de interés", C2).

Esto confirma de nuevo la adecuada asignación del nivel en el PCIC en función, como hemos repetido, del criterio nociofuncional que lo caracteriza, y que por tanto, la selección léxica por niveles debe comenzar atendiendo las escalas descriptivas del MCER relativas a riqueza y dominio léxico, concretándose dicha escala, desde un enfoque nociofuncional, en los listados de nociones específicas que presenta el PCIC. No obstante, como dichos repertorios no son exhaustivos, para que el profesor de ELE pueda enriquecerlos con otras SF y adaptarlos a sus necesidades y a las de sus alumnos, hemos comprobado que el índice de frecuencia es útil, completándolo con otros criterios, siempre con la vista puesta en la rentabilidad comunicativa, tanto desde el punto de vista lexicométrico como lingüístico.

5 Conclusiones

Partiendo de la base de la importancia de las SF en el desarrollo de la competencia léxica de los estudiantes de ELE, de la centralidad y transversalidad del léxico en el desarrollo de la competencia comunicativa, y en la misma dirección en la que apuntan diversos especialistas, consideramos que no es adecuada la recomendación tanto del MCER como del PCIC de retrasar el aprendizaje de estas SF hasta los niveles avanzados, ya que debido a su alta rentabilidad comunicativa su enseñanza desde los niveles iniciales, en consonancia también con las recomendaciones del

enfoque léxico, contribuiría sin duda a mejorar la fluidez y el nivel de comprensión de los aprendientes.

En cuanto a las conclusiones de nuestro trabajo, y a la luz de los resultados presentados, estimamos que podemos volver a confirmar, como ya hicimos en Martos & Contreras (2018), pero ahora con mayor profusión de datos y evidencias, que la nivelación realizada por el PCIC de las SF es bastante precisa a pesar de basarse en un criterio intuitivo.

Igualmente, creemos haber contribuido a reforzar la pertinencia de la frecuencia como criterio para establecer la nivelación de las SF y, en este mismo sentido, entendemos que *Google* puede ser considerada una herramienta adecuada para dicho fin, a la par que los corpus lingüísticos disponibles en la actualidad para el español. No obstante, somos conscientes de que los rápidos y continuos cambios que se producen en el volumen de información en la Red podrían imposibilitar la replicabilidad de los resultados, algo totalmente necesario a la hora de conferir robustez y validez a cualquier investigación. No obstante, como matiza González Fernández (2017: 132), este hecho "no es sino un reflejo de la naturaleza dinámica del lenguaje. Si la realidad cambia, los resultados de los trabajos deben reflejar esta evolución, lo que constituirá no un error en las conclusiones de trabajos anteriores, sino nuevos resultados". Además, como también señala este mismo autor, las herramientas tecnológicas con las que contamos actualmente pueden paliar esta posible deficiencia, pues es posible almacenar la información en bases de datos y recuperarla en cualquier momento para su comparación con los datos obtenidos en nuevas investigaciones. Como ejemplo de lo que decimos, hemos podido comprobar las apreciables diferencias en cuanto a las cifras presentadas en nuestro trabajo y las que recogen Lindstromberg y Boers (2008) en su estudio sobre la frecuencia de las SF en *Google*, diferencias que no restan validez a la clasificación realizada por estos investigadores.

Como hemos podido comprobar a la hora de realizar nuestra propuesta de nivelación, la especial distribución de las frecuencias de las SF en los distintos niveles, que como hemos visto confirma la ley de Zipf (1935), hace ineludible sumar la frecuencia a otros criterios, pues tal y como apuntaba Alvar (2005: 19), "la información sobre la frecuencia de uso no puede ser un valor rígido al que debamos sujetarnos, pues hay otros factores que influyen en el aprendizaje de las palabras", entre otros la dispersión en el corpus en el que se consulte (criterio este que también hemos verificado en parte en nuestro estudio), la rentabilidad y productividad, la opacidad o transparencia, los factores culturales, los intereses de los alumnos, las propias producciones de los estudiantes, o incluso las intuiciones de los hablantes nativos.

No obstante, la selección léxica por niveles debe estar guiada por los contenidos de las escalas descriptivas del MCER relativas a riqueza y dominio léxico, concretándose dicha escala con el repertorio de nociones específicas que presenta el PCIC, basado en un enfoque nociofuncional, que prima las habilidades comunicativas, las funciones, los temas y las situaciones que se quieran enseñar, antes que las características estructurales de las unidades léxicas.

Finalmente, la consideración de esos otros posibles criterios constituye una de las evidentes limitaciones de este trabajo, que pueden servir de base para futuras investigaciones, en las que, igualmente y para completar la base documental, sería necesario contabilizar las ocurrencias de otras configuraciones sintácticas de las SF, así como contemplar los casos de posibles cambios en el orden sintáctico de sus constituyentes y el efecto de fenómenos como la variación lingüística en la forma y el uso de las SF. No obstante, como el repertorio propuesto por el PCIC no es exhaustivo, para que el profesor de ELE pueda completarlo con otras SF, hemos comprobado que el índice de frecuencia es útil, conjugándolo con otros criterios, siempre con la vista puesta en la rentabilidad comunicativa y las necesidades de los alumnos, tanto desde el punto de vista lexicométrico como lingüístico.

Para ello, y en este sentido, hemos podido comprobar que los corpus lingüísticos son una excelente fuente de información que pone a nuestra disposición un amplio abanico de datos relativos a todos estos aspectos: frecuencias, dispersión, densidad, ámbitos temáticos, tipología textual, variedades diatópicas, modalidad,etc.

Referencias bibliográficas

Aguilar Ruiz, Manuel José (2013), "Notas sobre las posibilidades de aprendizaje de español mediante unidades fraseológicas", *MarcoELE,* 17. Disponible en: https://bit.ly/338dCOg

Albelda Marco, Marta (2011), "Rentabilidad de los corpus discursivos en la didáctica de lenguas extranjeras", en Javier de Santiago, Hanne Bongaerts, Jorge Juan Sánchez & Marta Seseña (coords), *Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE*. Salamanca: ASELE, 83-95. Disponible en: https://goo.gl/5oEXpR

Alvarado Ortega, Mª Belén (2010), *Las fórmulas rutinarias del español: teoría y aplicaciones.* Frankfurt am Main: Peter Lang.

Alvar Ezquerra, Manuel (2005), "La frecuencia léxica y su utilidad en la enseñanza del español como lengua extranjera", en Mª Auxiliadora Castillo Carballo (Coord.), *Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: deseo y realidad*. *Actas del XV Congreso Internacional de la Asele*. Sevilla: Universidad de Sevilla, 19-39. Disponible en: https://bit.ly/2XDRcDi

Baralo, Marta (2005), "La competencia léxica en el Marco común europeo de referencia", *Carabela,* 58:27-48.

Baralo, Marta (2007): "Adquisición de palabras: redes semánticas y léxicas", en *Actas del Foro de español internacional: Aprender y enseñar léxico,* 384-399. Disponible en: https://bit.ly/2LJhgen

Battaner Arias, Paz & Carmen López Ferrero (2019), *Introducción al léxico, componente transversal de la lengua.* Madrid: Cátedra.

Benigno, Verónica, Olivier Kraiff, Francis Grossmann & Antonino Velez (2016), "La notion de collocation fondamentale: Une étude de corpus", *Cahiers de Lexicologie*, 108(1):125-146.

Bolaños Cuéllar, Sergio (2015), "La lingüística de corpus: perspectivas para la investigación lingüística contemporánea". *Forma y Función*, 28(1):31-54.

Briz Gómez, Antonio & Marta Albelda Marco (2009), "Estado actual de los corpus de lengua española hablada y escrita: I+D", en *El español en el mundo. Anuario 2009*, Madrid: Instituto Cervantes. Disponible en: https://bit.ly/2txdcBo

Bustos Gisbert, Jose María (2001), "Definición de glosarios léxicos del español: niveles inicial e intermedio", *Enseñanza*, 19:35-72.

Buyse, Kris (2017), "Los corpus como herramientas de aprendizaje del léxico", en VV. AA., *Enseñar léxico en el aula de español. El poder de las palabras*, Barcelona: Difusión, 121-141.

Casares, Julio (1950/1992), *Introducción a la lexicografía moderna*, Madrid: CSIC.

Consejo de Europa (2002), *Marco común europeo de referencia para las lenguas aprendizaje, enseñanza, evaluación.* Madrid: Anaya y CVC. Disponible en: https://bit.ly/2IkNKo4

Contreras, Narciso M. & Fermín Martos (2020), "Las secuencias formulaicas en la enseñanza del español como LE/L2. La frecuencia léxica como criterio de nivelación", *Porta Linguarum*, 33:111-127.

Corpas Pastor, Gloria (1996), *Manual de fraseología española*, Madrid: Gredos.

Corpas Pastor, Gloria (ed.) (2003), *Diez años de investigación de fraseología: análisis sintáctico-semánticos, contrastivos y traductológicos,* Madrid: Iberoamericana.

Coseriu, Eugenio (1981), *Lecciones de Lingüística general*, Madrid: Gredos.

Crida, Carlos & Arianna Alessandro (Eds.) (2019), *Innovación en fraseodidáctica. Tendencias, enfoques y perspectivas*, Berlin: Peter Lang.

Cruz Piñol, Mar (2014), "Veinte años de tecnologías y ELE. Reflexiones en torno a la enseñanza del español como lengua extranjera en la era de Internet", *MarcoELE,* 19. Disponible en: https://bit.ly/2Kc8901

Cruz Piñol, Mar (2017), *Lingüística de corpus y enseñanza de español como 2/L*. Madrid: Arco Libros.

Erman, Britt & Beatrice Warren (2000), "The idiom principle and the open-choice principle", *Text*, 20, 87-120.

Ettinger, Stefan (2008), "Alcances e límites da fraseodidáctica. Dez preguntas clave sobre o estado actual da investigación", *Cadernos de Fraseoloxía Galega* 10, 95-127.

Ferrando Aramo, Verónica (2012), *Aspectos teóricos y metodológicos para la compilación de un diccionario combinatorio destinado a estudiantes de E/LE*, Tesis doctoral, Tarragona: Universitat Rovira i Virgili,. Disponible en: https://bit.ly/34c6wtm

García Salido, Marcos (2017), "La frecuencia de corpus como criterio para nivelar colocaciones léxicas", *Études Romanes de Brno,* 38(2):29-49.

García Salido, Marcos & Margarita Alonso Ramos (2018), "Asignación de niveles de aprendizaje a las colocaciones del Diccionario de Colocaciones del español", *Revista Signos. Estudios de Lingüística,* 51(97):153-174.

Gómez Molina, Jose Ramón (2004), "Los contenidos léxico-semánticos", en Jesús Sánchez Lobato & Isabel Santos Gargallo (Eds.), *Vademécum para la formación de profesores, Enseñar español como segunda lengua (L2)/lengua extranjera (LE)*. Madrid: SGEL, 789-810.

González Fernández, Adela (2017), "La web como corpus: un esbozo", *Lengua y Habla*, 21, 126-150.

González Rey, M.ª Isabel (2012), "De la didáctica de la fraseología a la fraseodidáctica", *Paremia*, 21, 67-84.

Higueras García, Marta (2004), "Internet en la enseñanza del español" en Jesús Sánchez Lobato & Isabel Santos Gargallo (Eds.), *Vademécum para la formación de profesores. Enseñar español como segunda lengua (L2)/lengua extranjera (LE)*. Madrid: SGEL, 1067-1077.

Instituto Cervantes (2006), *Plan Curricular del Instituto Cervantes. Niveles de referencia para el español*. Madrid: Instituto Cervantes, Biblioteca Nueva. Disponible en: https://bit.ly/2IWhigd

Lennon, Paul (1998), "Approaches to the teaching of idiomatic language", *IRAL*, XXXVI/1:11-30.

Leontaridi, Elini, Marina Ruiz Morales & Natividad Peramos Solert (2011)," Las unidades fraseológicas del español: su enseñanza y adquisición en la clase de ELE", en *Actas de las Jornadas de Formación del Profesorado en la Enseñanza de ELE y la Literatura Española Contemporánea*. Grecia: Universidad Aristóteles de Salónica, 187-206. Disponible en: https://bit.ly/2Od4wLN

Lewis, Michael (1993), *The Lexical Approach. The state of ELT and a way forward*. London: Teacher Training.

Lewis, Michael (1997), *Implementing the Lexical Approach: Putting theory into practice*. London: Teacher Training.

Lindstromberg, Seth & Frank Boers (2008), *Teaching chunks of language: From noticing to remembering*. Londres: Helbling Languages.

López Vázquez, Lucía (2011), "La competencia fraseológica en los textos de los manuales de ELE de nivel superior", en Javier de Santiago Guervós et al. (Eds.), *Del texto a la lengua: la aplicación de los textos a la enseñanza-aprendizaje del español L2-LE*. Salamanca: ASELE, 531-542. Disponible en: https://bit.ly/2quqd1a

Martín Noguerol, María (2012), "*¿Qué se dice en español cuando…?* Las fórmulas rutinarias y las situaciones sociales de comunicación en los niveles iniciales", en Óscar Abenójar (Ed.), *Actas del III simposio internacional de didáctica del español para extranjeros*. Argel: Instituto Cervantes, 57-64. Disponible en: https://goo.gl/PZcsdB

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

Martos, Fermín & Narciso Contreras (2018), "El empleo de corpus para el aprendizaje de secuencias formulaicas en ELE/L2. La frecuencia de uso en el nivel B2 del PCIC", *CHIMERA. Romance Corpora and linguistic Studies* 5(1):1-26.

McGee, Iain (2008), "Word frequency estimates revisited: A response to Alderson (2007)", *Applied Linguistics,* 29(3):509-514.

Moreno Teva, Inmaculada (2012), *Las secuencias formulaicas en la adquisición de español L2*. Tesis doctoral. Suecia: Universidad de Estocolmo. Disponible en: https://bit.ly/2OdGyQH

Muñoz-Basols, Javier (2015), "Enseñanza del lenguaje idiomático", en Javier Gutiérrez-Rexach (Ed.), *Enciclopedia lingüística hispánica*, vol. 2. London: Routledge, 442-453.

Nation, Paul (1990), *Teaching and learning vocabulary*. Boston: Heinle an Heinle.

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

Olímpio de Oliveira, M.ª Eugenia (2006), "Fraseología y enseñanza de español como lengua extranjera", *RedELE*, 5. Disponible en: https://bit.ly/33bVnHP

Penadés Martínez, Inmaculada (1999), *La enseñanza de las unidades fraseológicas*. Madrid: Arco/Libros.

Pérez Serrano, Mercedes (2015), "Tratamiento de la combinatoria léxica en documentos de referencia y curriculares: el caso del MCER y del PCIC", *Revista de Investigación Lingüística*, 18:213-232.

Pérez Serrano, Mercedes (2017), *La enseñanza-aprendizaje del vocabulario en ELE desde los enfoques léxicos*. Madrid: Arco/Libros.

REAL ACADEMIA ESPAÑOLA, *Corpus de referencia del español actual* (CREA). Disponible en: http://www.rae.es/recursos/banco-de-datos/crea

REAL ACADEMIA ESPAÑOLA, *Corpus del Español del Siglo XXI* (CORPES). Disponible en: http://www.rae.es/recursos/banco-de-datos/corpes-xxi

Rojo, Guillermo (2016), "Los corpus textuales del español", en Javier Gutiérrez-Rexach (Ed.), *Enciclopedia lingüística hispánica*. London: Routledge, 285-296.

Ruiz Gurillo, Leonor (1997), *Aspectos de fraseología teórica española*. Valencia: Universidad, Anejo XXIV de *Cuadernos de Filología*.

Ruiz Gurillo, Leonor (2000), "La fraseología", en Antonio Briz & Grupo VAL.ES.CO. (Coords.), *¿Cómo se comenta un texto coloquial?* Barcelona: Ariel, 169-189.

Sánchez Rufat, Anna (2011), "Léxico gramaticalizado y lengua formulaica: algunas precisiones al enfoque léxico", *Sintagma,* 23:85-98.

Sánchez Rufat, Anna (2017), "Estrategias para la enseñanza de secuencias formulaicas en el aula de español como lengua extranjera", en Dimitrinka. Georgeva Níkleva (Ed.), *La formación del profesorado de español como lengua extranjera. Necesidades y tendencias*, Bern: Peter Lang, 257-282.

Saracho Arnáiz, Marta (2016), "Una metodología para la enseñanza-aprendizaje de fraseología en ELE". *Boletín de ASELE*, 55:17-31.

Serradilla Castaño, Ana (2014), "La fraseología en el aula de ELE: nuevos enfoques y propuestas didáctica", *RedELE* (volumen monográfico), VV.AA. (eds.), *¿Qué necesitamos en el aula de ELE?: reflexiones en torno a la teoría y la práctica*, 73-98. Disponible en: https://bit.ly/2D8wcvh

Sinclair, John (1991), *Corpus, Concordance and Collocation*. Oxford: Oxford University Press.

Sinclair, John (2004), "Corpus and text: Basic principles", en Martin Wynne (Ed.), *Developing linguistic corpora: A guide to good practice*. Oxford: Oxbow Books, 1-16. Disponible en: https://bit.ly/2XGFtUG

Sinclair, John & Renouf, A. (1985), "A lexical learning syllabus for language", en Ronald Carter & Michael McCarthy (Eds.), *Vocabulary and Language Teaching* Londres/Nueva York: Longman, 140-160.

Szyndler, Agnieszka (2015), "La fraseología en el aula de E/LE: ¿un reto difícil de alcanzar? Una aproximación a la fraseodidáctica", *Didáctica. Lengua y Literatura*, 27:197-216.

Timofeeva, Larissa (2013), "La fraseología en la clase de lengua extranjera: ¿misión imposible?", *Onomázein*, 28:320-336.

Velázquez Puerto, Karen (2018), *La enseñanza-aprendizaje de fraseología en ELE*. Madrid: Arco/Libros.

Villayandre Llamazares, Milka & Laura Llanos Casado (2017), "Los corpus electrónicos en la clase de español: reflexiones y aplicaciones", en Beatriz Peña & Ana Mª. Aguilar (Coords.), *Didáctica de la lengua y la literatura. Buenas prácticas docentes.* Tomo I , Madrid: ACCI, 52-79.

Wray, Alison (2002), *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Zipf, George (1935), *The psycho-biology of language*. Cambridge: MIT Press.

Zuluaga, Alberto (1980), *Introducción al estudio de las expresiones fijas,* Tesis doctoral inédita, Tübingen: Max Hueber, Verlag.