e-ScriptaRomanica
vol. 5 (2018): 39-60
MONOGRÁFICO

http://www.e-scripta.uni.lodz.pl/ojs/index.php/escriptaromanica | ISSN 2392-0718
Received: 05/04/2016 | Accepted: 31/07/2016
http://dx.doi.org/10.18778/2392-0718.05.05

Frecuencia léxica y secuenciación del vocabulario en lecturas graduadas del español[1]

Lexical frequency and vocabulary sequencing in Spanish graded readers

Marcos García Salido
Universidade da Coruña, España
Orsolya Vincze
Universidade da Coruña, España
Ana Orol González
Universidade da Coruña, España
Margarita Alonso Ramos
Universidade da Coruña, España

Resumen: El presente artículo estudia la distribución de palabras y colocaciones presentes en un corpus de lecturas graduadas del español a través de los distintos niveles de aprendizaje. El objetivo principal es verificar si se da una correlación entre la frecuencia léxica tal como se registra en un corpus del español general y la distribución de los elementos del vocabulario (formas univerbales y pluriverbales) en textos de diferente nivel, de manera que los elementos infrecuentes sean más numerosos conforme el nivel sube. Esta correlación no se puede dar por supuesta en las lecturas graduadas en español, pues un repaso a la bibliografía relevante indica que se ha dado prioridad a factores distintos a la selección del vocabulario en la creación de este tipo de materiales (en concreto, al componente gramatical).

Palabras clave: nivelación del vocabulario, nivelación de colocaciones, frecuencia léxica.

Abstract: This article examines the distribution of words and collocations in a corpus of Spanish graded readers across different levels of proficiency. The main aim is to verify whether there is any relation between lexical frequency as registered in a general corpus of Spanish and the distribution of vocabulary items (single words and collocations) in texts of different levels, such that there is an increase of infrequent items as the proficiency level rises. Such a relation cannot be taken for granted in the case of Spanish graded readers, since a review of the literature suggests that factors other than vocabulary selection (namely, grammatical features) have been given priority in creating texts for a given proficiency level.

Keywords: vocabulary grading, collocations grading, lexical frequency.

[1] El trabajo presentado en este artículo ha sido financiado en parte por el MINECO y los fondos FEDER a través del proyecto de investigación FFI2011-30219-C02-01, así como por la Xunta de Galicia, por medio del contrato de investigación postdoctoral POSA/2013/191, y el Ministerio de Educación, mediante el contrato para la Formación del Profesorado Universitario AP2010-4334.

1. Introducción

Uno de los criterios fundamentales para decidir qué ítems del léxico deberían incluirse en un programa de enseñanza de una lengua extranjera/segunda es la frecuencia de uso. La razón subyacente es que la inversión de tiempo en la enseñanza de ítems léxicos frecuentes resulta rentable, puesto que estos representan una gran proporción de las formas empleadas en cualquier texto y es probable que los aprendices se encuentren con ellos a menudo (Martínez 2013: 187; Nation 2001: 13 y ss.). La frecuencia de uso es también un factor que se debe tener en cuenta al decidir la progresión con la que se presentará el vocabulario. Autores como Sinclair y Renouf (1985: 154-155) y Nation (2001: 20-21) insisten en la importancia de adquirir un dominio considerable de ítems léxicos frecuentes antes de lanzarse a aprender formas menos frecuentes (o que representan una proporción menor de las formas usadas en los textos)[2]. Con respecto a las unidades pluriverbales en general, una serie de investigadores también ha defendido que la frecuencia debe tenerse en cuenta al planear tanto su selección como la secuencia en la que se presentarán, junto con otros factores como la composicionalidad o la transparencia (véanse, por ejemplo, Nation 2001: 329 o Martinez 2013)[3].

En el caso particular de la investigación dedicada a la enseñanza de español como lengua extranjera (ELE, en adelante), también se ha considerado valioso el papel de la frecuencia de uso en la planificación de qué vocabulario incluir y en qué orden presentarlo en los programas docentes. Así, el *Plan curricular del Instituto Cervantes* (1997-2015), que a veces se toma como referencia para determinar los contenidos lingüísticos que han de incluirse en materiales ELE, afirma que la frecuencia de uso —junto con la *rentabilidad comunicativa*— ha sido un criterio fundamental para decidir qué vocabulario se incluye en cada uno de los niveles de referencia (véanse también Gómez Molina, 2004: 797 e Izquierdo, 2004: 321 y ss.). También en el caso de las colocaciones, varios investigadores han señalado la utilidad de la frecuencia como guía en la elección de las que han de tratarse en las diferentes etapas del proceso de aprendizaje (Ferrando Aramo, 2012: 360-361; Higueras, 2006: 30). A pesar del relativo consenso en cuanto a la importancia de la frecuencia de uso en la elección y secuenciación del vocabulario, en el ámbito ELE no está del todo claro hasta qué punto la frecuencia léxica está supeditada a otro tipo de criterios o si la información relativa a este aspecto que manejan los autores de programaciones, manuales, etc. se basa en datos de corpus o en otras fuentes (por ejemplo, la frecuencia percibida; *vid. infra*).

El propósito de este artículo, así pues, es determinar si existe alguna relación entre la frecuencia léxica, determinada a partir de un corpus del español general, y la secuenciación del vocabulario en materiales ELE. Con este

[2] Sus enfoques no son, con todo, idénticos: Sinclair y Renouf conceden una especial importancia a conocer no solo formas frecuentes, sino *usos* frecuentes (pares de forma y significado que otros autores considerarían *unidades léxicas*). Nation, por su parte, toma como referencia el concepto de familia léxica (*vid. infra*).

[3] Si reinterpretamos este argumento en términos de adquisición, esperaríamos que el vocabulario frecuente se manejase antes que el infrecuente. Ciertos estudios de corpus (Crossley *et al*., 2014) y tests de vocabulario (González Fernández y Schmitt, 2015) muestran, sin embargo, que, a pesar de que existe cierta correlación entre la frecuencia y la adquisición del vocabulario, esta no es demasiado fuerte, al menos en niveles bajos.

propósito se ha usado un corpus de fragmentos de lecturas graduadas como muestra representativa del vocabulario al que están expuestos los aprendices de ELE durante su proceso de instrucción. Esta colección de lecturas graduadas está disponible en la web con el título de *Lecturas paso a paso* (Instituto Cervantes, 2000-2015)[4].

El estudio se estructura de la manera siguiente. En el apartado 2 se describe brevemente el papel que tiene la frecuencia en la secuenciación del léxico y se presentan los objetivos de la investigación con algo más de detalle. El apartado 3 se encarga de presentar diversos aspectos de la metodología adoptada, en concreto las características del corpus y nuestra concepción de las unidades de vocabulario —formas de palabra y colocaciones— que hemos estudiado. El apartado 4 presenta los resultados de examinar la diversidad léxica de las unidades de vocabulario consideradas de forma aislada y su distribución en las distintas secciones del corpus, establecidas a partir de los niveles de las lecturas que lo componen. El apartado 5 analiza la distribución de las colocaciones presentes en el corpus de acuerdo con los mismos parámetros. Por último, el apartado 6 presenta las conclusiones del estudio.

2. El papel de la frecuencia en la secuenciación del vocabulario de materiales ELE

Como se decía en la introducción, se reconoce ampliamente que la frecuencia léxica debería desempeñar un papel en la planificación de la enseñanza del vocabulario. Una manera de obtener información sobre la frecuencia de un ítem léxico, ya sea una forma univerbal o una secuencia pluriverbal, es observar sus apariciones en un corpus, es decir, su frecuencia de corpus. Hay que tener en cuenta que esta medida no representa directamente la exposición de ningún individuo concreto a datos lingüísticos, como han notado diversos autores (Hoey, 2005: 14; González-Fernández y Schmitt, 2015)[5]. Es, así pues, útil distinguir entre la *frecuencia léxica* de una determinada forma, esto es, la frecuencia con la que un ítem del vocabulario se presenta en el input del individuo X y la *frecuencia de corpus* de ese ítem. Aun teniendo esta precaución en mente, la frecuencia de corpus parece la aproximación más objetiva posible a la frecuencia léxica. Sin embargo, no está claro hasta qué punto la frecuencia de corpus ha venido determinando la inclusión y secuenciación del vocabulario a través de las distintas etapas del aprendizaje en el ámbito de ELE. En el caso de las colocaciones esto probablemente es una consecuencia de la falta de estudios de frecuencia colocacional que apunta Ferrando Aramo (2012: 360). Por lo que respecta a las formas léxicas univerbales, hace aproximadamente un siglo que tenemos a nuestra disposición listas de frecuencia (*cf.* Izquierdo, 2004: 333), que, por otro lado, hoy en día pueden ser extraídas de corpus modernos con relativa facilidad. A pesar de ello, uno de los documentos más influyentes en el diseño de programaciones y materiales ELE (Instituto Cervantes, 1997-2015) prefiere apoyarse en la intuición de profesionales experimentados en la enseñanza, antes que en datos de corpus, para determinar la frecuencia y rentabilidad

[4] http://cvc.cervantes.es/aula/lecturas/

[5] Nadie ha estado expuesto a todos los textos que se incluyen en un corpus determinado ni ningún corpus contiene todo el input que recibe una persona concreta.

del vocabulario que en él se incluye (véase Instituto Cervantes, 1997-2015: Introducción a la sección 9)[6].

Por lo que toca a las lecturas graduadas, puesto que normalmente se las considera un instrumento para que los aprendices practiquen la lectura extensiva sin que se encuentren con demasiadas palabras y estructuras que no puedan entender dado su nivel de aprendizaje, las listas de frecuencia léxica parecen en principio una herramienta adecuada para controlar el vocabulario que se incluirá en un determinado nivel. En el ámbito del inglés como lengua extranjera (ILE), parece que esto se viene llevando a cabo desde hace tiempo: de acuerdo con Nation (2001: 15), la lista de palabras de alta frecuencia de Michael West (1953) se usó para la confección de las colecciones más tempranas de lecturas graduadas. La historia de este tipo de materiales en el mundo de ELE es bastante más breve (de acuerdo con San Mateo Valdehíta [2005: 25], las primeras datan de finales de los años ochenta) y a juzgar por lo que se dice en la bibliografía sobre el tema, las listas de frecuencia no han desempeñado un papel tan relevante como en la tradición anglófona. Así, Báez Montero y Suárez Rodríguez (2012: 122), al analizar diversas colecciones de lecturas graduadas en español, encuentran que los contenidos apropiados para cada nivel se deciden sobre la base de consideraciones de tipo gramatical: según las autoras, hay cierta tendencia a considerar el tiempo y el modo verbal como criterios niveladores. De forma similar, San Mateo Valdehíta (2005: 41), dando cuenta del diseño de uno de estos textos, afirma que su principal preocupación fue la de ajustar la dificultad de las estructuras gramaticales al nivel de los destinatarios. En cuanto al vocabulario, la autora menciona que, ya que el texto se dirige a lectores de nivel avanzado, se han incluido formas pertenecientes a los registros coloquial y formal, así como "jerga juvenil" (*ibid*.: 47).

Si bien de lo anterior se desprende que la gramática ha tenido un papel preponderante para decidir qué contenidos se incluirán en un nivel determinado, ciertas editoriales hacen explícitos sus criterios en relación con la selección del léxico. Así, Santillana y SGEL ofrecen indicaciones específicas[7] para alguna de sus colecciones de lecturas graduadas, aclarando que se ha controlado el número de palabras léxicas diferentes en cada nivel, de manera que la diversidad léxica se incrementa conforme el nivel sube. En los prólogos de las lecturas graduadas de SGEL se señala, además, que se ha controlado la proporción de palabras de cada texto cubiertas por formas pertenecientes a diferentes bandas de frecuencia (*muy alta, alta, moderada, baja* y *muy baja*). Estas bandas se establecieron a partir de las frecuencias léxicas registradas en el CREA (Real Academia Española, s. f.) y mediante una metodología

[6] En concreto en la sección citada se afirma: "Como criterios de selección, siempre partiendo de la apreciación intuitiva basada en la experiencia docente, han primado la frecuencia y la rentabilidad comunicativa" (http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/niveles/09_nociones_especificas_introduccion.htm).

[7] Esta información es accesible en las páginas http://www.santillanaele.com/catalogo/lecturas-graduadas/leer-espanol-adultos y http://ele.sgel.es/ficheros/productos/downloads/Facil%20lectura%20Introduccion_400.pdf. Cabe notar que la colección de SGEL no está entre las que componen las *Lecturas paso a paso* del Instituto Cervantes.

desarrollada para el *Gran Diccionario de Uso del Español Actual* (Sánchez Pérez, 2001). Aun así, los términos empleados para hacer referencia a las unidades de vocabulario consideradas (*palabras, palabras léxicas*) son hasta cierto punto ambiguos ya que permiten que el término *palabra* se interprete como forma de palabra (o *type*) o como un *lema* de una entrada de diccionario[8], de manera que no podemos formarnos una idea absolutamente concreta de la diversidad léxica de cada nivel.

Cuadro 1. Número de *palabras* diferentes en las lecturas graduadas de Santillana y SGEL

SGEL	**Nivel**	**INICIAL**		**INTERMEDIO**		**AVANZADO**	
	Nº palabras	500-1200		1200-2000		2000-3000	
Santillana	**Nivel**	**1***	**2***	**3**	**4**	**5**	**6**
	Nº palabras	400	700	1000	1500	2000	2000

* Hay dos niveles adicionales dedicados a lectores infantiles *(principiantes)* que contienen respectivamente 100 y 200 palabras diferentes

En resumen, aun cuando en la investigación sobre ELE se considera la frecuencia un factor importante para determinar qué vocabulario deberían adquirir los aprendices y en qué orden, en la práctica, salvo excepciones como las que suponen ciertas colecciones de Santillana y SGEL, parece haber existido cierta reticencia a utilizar listas de frecuencia o datos procedentes de corpus, dando primacía bien a la gramática bien a la frecuencia percibida para decidir qué contenidos se incluyen en cada nivel. En el caso concreto de las lecturas graduadas, el vocabulario aparece en ocasiones supeditado a la gramática como criterio nivelador. Considerando el panorama descrito, este artículo pretende dilucidar si existe alguna relación entre la frecuencia de corpus y la secuenciación con la que el vocabulario se presenta a los aprendices de ELE. Este objetivo se justifica si tenemos en cuenta que no parece haberse prestado especial atención a los datos de corpus en la distribución del vocabulario a través de distintos niveles, de modo que no puede darse por sentada una correlación entre ambos aspectos. Además, de demostrarse tal correlación, esto significaría que los datos de corpus —y en concreto la frecuencia— podrían utilizarse como un criterio fácilmente exportable para la nivelación del vocabulario de nuevos materiales ELE. En lo que sigue, trataremos de establecer:

i) si el vocabulario que se presenta a los aprendices de ELE a través de los sucesivos niveles de aprendizaje es diferente en términos de su frecuencia en un corpus de español general, de modo que ítems de frecuencia baja se introducen después de haber tratado las formas más frecuentes y

ii) si tales diferencias pueden también observarse en las colocaciones que se incluyen en los materiales estudiados.

3. Descripción del corpus, unidades de vocabulario y colocaciones

Esta sección trata tres aspectos metodológicos. En primer lugar, describimos el corpus que se ha usado en este estudio. En segundo lugar, damos cuenta de

[8] Por ejemplo, *cantaba* y *cantó* son dos *types* pertenecientes a un mismo lema *cantar* (*vid. infra* ap. 3.2).

qué unidades de vocabulario hemos considerado. Por último, se da una definición del concepto de colocación empleado.

3.1. Características del corpus

El corpus está formado por los textos que se presentan en la web del Instituto Cervantes (2000-2015) bajo el título Lecturas paso a paso. Los textos proceden de diferentes editoriales (Coloquio, Difusión, Edelsa, Edinumen, Santillana, SGEL y SM) y de colecciones dedicadas a la elaboración de lecturas graduadas, por lo que pensamos que pueden ser representativos de los tipos de material a los que han tenido acceso profesores y aprendices de español durante los últimos años —y a los que todavía pueden acceder de forma gratuita vía web—. Como se desprende de lo visto en el apartado 2, los criterios de las diferentes colecciones no tienen por qué ser coincidentes, pero esto redunda en la representatividad de la muestra utilizada, pues es probable que, durante su instrucción, los aprendices hayan tenido acceso a materiales de diversa procedencia.

El hecho de que los textos estén respaldados por una institución de reconocido prestigio en el campo y que sean accesibles a través de la web es un elemento más a favor de considerarlos representativos del input que puede recibir un aprendiz de español. Por último, el tamaño moderado de esta colección la hace asequible a la anotación manual, lo que ha sido esencial para la identificación de las colocaciones que contiene según nuestro marco teórico. Ha de tenerse en cuenta, además, que esta muestra se toma como representativa del input que reciben los aprendices, pero que los datos de frecuencia tanto de vocabulario univerbal como de colocaciones se han extraído de un corpus de español general (el esTenTen11; Kilgarriff y Renau, 2013).

El Instituto Cervantes ha unificado los diferentes sistemas de nivelación empleados por cada editorial, a veces distintos dependiendo incluso de la colección (*vid.* Apéndice), en uno que distingue únicamente tres niveles: inicial, intermedio y avanzado. No hay indicaciones explícitas de los criterios que se manejan para incluir cada uno de los textos en uno u otro nivel más allá de que la clasificación responde a la dificultad de los textos (vid. Instituto Cervantes, 2000-2015)[9]. Además, excepcionalmente se pueden encontrar textos pertenecientes a manuales especializados en el nivel avanzado, es decir, textos que originalmente no se concibieron como lecturas graduadas. El corpus contiene un total de 70124 palabras repartidas en tres subcorpus correspondientes al sistema de tres niveles empleado por el Instituto Cervantes. Se puede ver la composición de cada uno en el Cuadro 2.

Cuadro 2. Composición de los tres subcorpus de *Lecturas paso a paso*

	INICIAL	INTERMEDIO	AVANZADO
types	3639	3561	5063
tokens	27987	16029	26108

[9] http://cvc.cervantes.es/aula/lecturas/

De las cifras se desprende que los tres subcorpus no tienen un tamaño homogéneo: la muestra correspondiente al nivel intermedio es considerablemente más pequeña que las otras dos. Debido a esta diferencia, comparar el número de *types* distintos que contiene cada sección es poco revelador con respecto a su diversidad léxica. En el apartado 4 se tratará este problema en más detalle y se propondrá una solución.

3.2. El concepto de unidad de vocabulario

Los estudios centrados en el aprendizaje del vocabulario pueden usar diversas opciones para tratar con la frecuencia léxica. Normalmente, se manejan los conceptos de *token* (cada ocurrencia de una determinada forma de palabra en un texto, corpus, etc.), *type* (la forma de palabra en sí, sin prestar atención a si ocurre una o más veces en un determinado corpus) y *lema* (una forma que normalmente agrupa todas las posibles variantes flexivas de una determinada palabra). A estos, hay que añadir en cierta tradición anglófona el concepto de *familia léxica* (*word family*) desarrollado por Nation, que incluye, además de las variantes flexivas, las formas derivadas de un determinado lema ("a headword, its inflected forms, and its closely related derived forms" [Nation, 2001: 8]) y que ha sido aplicado en un buen número de estudios de ILE (por ejemplo, en Laufer y Nation, 1995). Nation justifica la conveniencia de usar las familias léxicas como unidad de referencia pues, según el autor, reflejan especialmente bien la carga que supone aprender vocabulario nuevo: una vez que un aprendiz ha dominado los sistemas de flexión y derivación del inglés, el esfuerzo que supone adquirir formas flexionadas o derivadas de un lexema ya conocido se considera mínimo.

La conveniencia de este concepto a la hora de estudiar el aprendizaje de una lengua segunda ha sido puesto en tela de juicio por Bogaards (2001: 322–323), que señala algunos problemas que entraña el uso de las familias léxicas. En primer lugar, la idea de familia léxica obvia el esfuerzo que supone la adquisición de palabras polisémicas. En segundo lugar, una forma representativa de una familia léxica puede corresponder a una unidad léxica entera, pero también puede ser parte de una (p. ej., *defender un supuesto, por supuesto*). En tercer lugar, Bogaards menciona casos que, a pesar de que podrían mantener relaciones de derivación o composición (sus ejemplos son *grace* ['garbo, salero'], *gracious* ['amable', 'magnánimo'], *graceless* ['feo']; *saucepan* ['sartén'], *saucebox* ['persona rijosa']), no parece razonable considerar miembros de la misma familia léxica debido a las diferencias semánticas que existen entre ellos. Algunas de estas críticas pueden hacerse extensivas a unidades más tradicionales manejadas en lingüística de corpus, como las citadas *type* y lema, ya que, si no ha habido un proceso de desambiguación semántica, son formas que, en muchos casos, corresponden a más de un significado. El concepto de *unidad léxica* (*grosso modo*, un par formado por un significado bien delimitado y una única forma de palabra o una secuencia plurivebal) resuelve muchos de estos problemas, puesto que posibilita tomar en cuenta unidades pluriverbales, relaciones de polisemia, etc. Bogaards (2001), de hecho, propone la adopción del concepto de unidad léxica en el contexto de la investigación en L2 como un sustituto a los de *palabra* o *familia léxica*. El autor sigue la definición de Cruse (1986), aunque el concepto, tratado de una manera más o menos similar, es también capital

dentro de la Teoría Sentido-Texto (TST; Mel'čuk et al., 1995). Para las unidades léxicas univerbales, la TST usa el término de *lexema,* mientras que para las plurivebales el de *frasema* (para más detalles sobre la relación del concepto unidad léxica en Cruse y en la TST, véanse Cruse 1986: capítulo 2 y Mel'čuk et al., 1995: 55 y ss.). Desgraciadamente, debido a las limitaciones que imponen los softwares más habituales para el manejo de grandes corpus, hemos tenido que conformarnos con utilizar *tokens, types* y *lemas* para el estudio de las formas univerbales del vocabulario de nuestros corpus.

3.3. Colocaciones

En cuanto al concepto de *colocación* que adoptamos en este trabajo, cabe señalar que seguimos la perspectiva fraseológica presente en la Teoría Sentido-Texto al considerar una colocación como una combinación binaria de unidades léxicas, una de las cuales funciona como *base* y la otra como su *colocativo*. El hablante elige libremente la base para transmitir un determinado significado. La elección del colocativo está guiada léxicamente por la base, es decir, es la identidad léxica de la base —y no solo su semántica— la que determina la elección del colocativo. Se explica así por qué lexemas semánticamente afines no son compatibles con los mismos colocativos (p. ej.: *dar/*hacer un paseo; *dar/hacer una excursión;* vid. Mel'čuk, 1996, 2012, entre otros).

De acuerdo con este enfoque, los criterios tenidos en cuenta para determinar si una cierta combinación podía o no considerarse colocación fueron (siguiendo a Mel'čuk, 2012: 38–39):

(a) La composicionalidad semántica de la combinación. Las colocaciones son composicionales, de manera que su significado puede ser descompuesto en partes más pequeñas y estas partes pueden relacionarse con cada miembro de la colocación. Así, no hemos tenido en cuenta expresiones pluriverbales no composicionales que a veces se incluyen bajo la etiqueta de colocación, como locuciones (p. ej., *tomar el pelo*) o cuasi-locuciones (e.j., *puente aéreo,* donde el significado 'aéreo' permanece, pero el sentido del conjunto no se deduce a partir de la suma de 'puente'+ 'aéreo').

(b) La elección de los miembros de la colocación. De acuerdo con la perspectiva fraseológica adoptada y como se ha indicado ya, la base de la colocación se elige libremente, mientras que la elección del colocativo está restringida en función de la base. Además, este puede ser una unidad léxica que aparece en otros contextos (p. ej., *decir* en *decir mentiras*) o puede ser exclusiva del contexto de la colocación (p. ej., el significante *robot* en *retrato robot* transmite un significado que solo presenta en el contexto de la colocación: 'hecho por la policía combinando partes de fotografías de acuerdo con la declaración de un testigo, etc.')

Las colocaciones de *Lecturas paso a paso* se han identificado manualmente de acuerdo con los criterios fraseológicos presentados. Este enfoque implica que nos hemos limitado a un subconjunto relativamente específico de combinaciones y que no hemos tenido en cuenta otras que se incluyen en ciertos trabajos que también estudian el fenómeno de las colocaciones, tales como sintagmas en los que uno de los constituyentes es una forma gramatical (por ejemplo, *this* en *this study*, ítem de la lista de

colocaciones elaborada por Durrant, 2009) o expresiones no composicionales (*for good* ['para siempre, definitivamente'] se define como colocación semánticamente opaca en Nation, 2001)[10]. Además, cabe señalar que la frecuencia de coocurrencia de los miembros de una determinada combinación en *Lecturas paso a paso* no ha sido relevante para su identificación como colocación.

4. Distribución del vocabulario en los distintos niveles de aprendizaje

Esta sección examina la distribución del vocabulario en nuestros tres diferentes subcorpus atendiendo exclusivamente a formas univerbales. La división en subcorpus se ha hecho de acuerdo con los tres niveles —inicial, intermedio y avanzado— que distingue el Instituto Cervantes en las *Lecturas paso a paso*. Como se ha dicho, las lecturas son una compilación de textos procedentes de diferentes colecciones y elaborados con diferentes criterios. Por ello, el primer paso ha sido determinar si los tres subcorpus eran diferentes en cuanto a la distribución de su vocabulario. Con tal fin se han tenido en cuenta dos aspectos: (i) su diversidad léxica, medida en términos de los cocientes *type*/*token* y lema/*token* y (ii) la composición de los subcorpus en cuanto a la frecuencia que presenta su vocabulario en el español general. Nuestras expectativas de partida eran que (a) la diversidad léxica se iría incrementando conforme subiera el nivel de aprendizaje y (b) que el uso de ítems de frecuencia léxica baja sería más limitado cuanto más bajo fuese el nivel. Las páginas que siguen dan cuenta de hasta qué punto se cumplen estas expectativas.

Los cocientes entre *types* y *tokens* y lemas y *tokens* pueden considerarse medidas de la diversidad léxica de un texto: un cociente de 1 significaría que cada *type* o lema aparece una sola vez en cada texto, mientras que valores más cercanos a cero indicarían que el texto está compuesto por frecuentes repeticiones de un pequeño conjunto de elementos. Ahora bien, la comparación de corpus de diferente tamaño por medio de estas medidas es problemática, ya que el volumen del corpus las condiciona. Con conjuntos pequeños de datos, la tasa de incremento de *types* y lemas es similar a la de los *tokens*, pero conforme el tamaño del corpus va aumentando el incremento de aquellos se ralentiza (cf. Biber 1993: 250; Sánchez y Cantos-Gómez, 1997: 261), de modo que los corpus pequeños tienen en general cocientes *type*/*token* o *lema*/*token* más altos que los más grandes. Este fenómeno complica la comparación de unos y otros. Para remediar en lo posible este problema hemos utilizado secciones de igual tamaño (10000 tokens) para cada uno de nuestros subcorpus. En el cuadro 3 se presentan los resultados de la comparación.

[10] De acuerdo con el enfoque adoptado, las locuciones o *idioms* no son un subgrupo dentro del conjunto de las colocaciones. Ambas se consideran elementos fraseológicos, pero las colocaciones son composicionales, a diferencia de las locuciones (véanse Mel'čuk, 2012, 2015).

Cuadro 3. Cocientes *type/token* y lema/*token* de los tres subcorpus

LEVEL	TYPE/TOKEN	LEMA/TOKEN
inicial	0,179	0,142
intermedio	0,217	0,166
avanzado	0,258	0,160

En lo que toca a los *types*, la diversidad léxica de los tres subcorpus se incrementa con el nivel de aprendizaje, tal como se esperaba. Sin embargo, en el caso de los lemas, el nivel intermedio muestra un cociente algo más alto que el avanzado. Parece, por tanto, que si bien el nivel avanzado presenta un número algo menor de lemas distintos por cada mil palabras que el intermedio, cada lema está representado por una variedad mayor de *types* (esto es, de formas flexivas distintas). De la comparación de estas dos medidas, cabe concluir que el nivel avanzado se distingue por su mayor complejidad morfológica, aun si su diversidad léxica es algo menor que la observada en el nivel intermedio.

El segundo criterio empleado para determinar las diferencias en la distribución de vocabulario en los tres subcorpus es la frecuencia que las formas incluidas en cada uno de ellos presentan en un corpus de español general. Esto se puede llevar a cabo o bien comparando listas de palabras agrupadas y ordenadas de acuerdo con su frecuencia, o bien estableciendo qué proporción de cada texto cubren los ítems de dichas listas (lo que en el ámbito anglosajón se ha dado en llamar *lexical frequency profile; vid.* Laufer y Nation, 1995). Dichas listas están normalmente dispuestas en grupos de mil elementos ordenados de mayor a menor frecuencia.

Para el presente estudio, extrajimos los 5000 *types* o formas de palabra más frecuentes en el corpus *esTenTen11* de español europeo, un corpus compuesto de textos de la web y que contiene más de dos mil millones de palabras. Comparamos, en primer lugar, las coincidencias entre esta lista y las listas de *types* de los distintos subcorpus de *Lecturas paso a paso*[11]. Los resultados de la comparación pueden verse en el Cuadro 4.

Cuadro 4. Comparación de los 5000 *types* más frecuentes en *esTenTen11* con los *types* presentes en *Lecturas paso a paso*

		INICIAL	INTERMEDIO	AVANZADO	MARGINAL
1K	Frec.	658	627	762	2047
	% columna	18,08	17,61	15,05	
	Resid. est	**2,05**	1,34	**-2,86**	

[11] Hemos descartado de las listas ciertas abreviaturas (p.ej. *a., o.,* etc.) que coinciden con *types* muy frecuentes (preposición *a*, conjunction *o,* etc.) debido a las distorsiones que causaban en el "perfil léxico" de los subcorpus, de modo que las listas finales pueden contener algo menos de 1000 ítems por franja. Tanto la comparación de listas como el perfil de los subcorpus se han hecho con la ayuda del software AntWordProfiler (Anthony, 2014).

2K	Frec.	377	334	475	1186
	% columna	10,36	9,38	9,38	
	Resid. est	1,34	-0,56	-0,66	
3K	Frec.	288	243	377	908
	% columna	7,91	6,82	7,45	
	Resid. est	1,13	-1,27	0,11	
4K	Frec.	238	214	297	749
	% columna	6,54	6,01	5,87	
	Resid. est	1,05	-0,24	-0,70	
5K	Frec.	188	181	285	654
	% columna	5,17	5,08	5,63	
	Resid. est	-0,44	-0,65	0,91	
OTROS	Frec.	1890	1962	2867	6719
	% columna	51,94	55,1	56,63	
	Resid. est	**-2,32**	0,25	**1,76**	
MARGINAL		3639	3561	5063	12263

De la comparación de las listas de vocabulario extraídas del *esTenTen* y de los tres subcorpus de *Lecturas paso a paso* podemos concluir que nuestra expectativa inicial se cumple parcialmente. Aunque las diferencias no son muy grandes, puede observarse que conforme sube el nivel, la proporción de *types* pertenecientes a la banda de los mil más frecuentes va decreciendo, mientras que la proporción de *types* que no se encuentran dentro de la lista de los cinco mil más frecuentes se va incrementando. En las franjas intermedias, sin embargo, no se observa una progresión tan clara.

Para saber si el vocabulario utilizado en los tres subcorpus es significativamente diferente en relación a su frecuencia en nuestro corpus de referencia hemos aplicado la prueba del chi cuadrado a los datos de frecuencias presentados en el Cuadro 4[12]. El valor global para el conjunto de

[12] A pesar de su uso tradicional en lingüística para comparar diferentes corpus, la prueba del chi cuadrado está cada vez más cuestionada en este ámbito (véase Bestgen, 2014). Uno de los problemas que presenta su aplicación es la sospecha de que se está violando la asunción de independencia entre las observaciones. Parece razonable pensar que la pertenencia a un mismo texto incide en la probabilidad de aparición de una determinada forma, de modo que las observaciones pertenecientes a él no son independientes entre sí (p. ej. la probabilidad de un término médico es más alta en un texto médico que en uno que trate de arquitectura). Con todo, aquí no comparamos ocurrencias de formas de palabra concretas en distintos corpus, sino ocurrencias de ítems pertenecientes a una determinada franja de frecuencia en distintas listas (Cuadro 4) o corpus (Cuadros 5 y 6). Dado que se ha observado que la distribución de formas frecuentes e infrecuentes sigue patrones similares en cualquier texto (Zipf, 1935), no está claro hasta qué punto la pertenencia a un mismo texto de varias observaciones significa que estas no sean independientes entre sí.

los datos (χ^2=31,297, p<0,001) permite afirmar que hay diferencias significativas en cuanto a la presencia de vocabulario más o menos frecuente dependiendo del nivel de las *Lecturas* considerado. Para determinar cuáles son las casillas que más contribuyen a esa diferencia nos hemos fijado en los residuos estandarizados de cada casilla[13], siguiendo la propuesta de Oakes y Farrow (2007). Destacamos en negrita las que presentan diferencias más grandes con respecto a la frecuencia esperada. En primer lugar, observamos que la diferencia más notable se da en las formas pertenecientes a las mil más frecuentes del español general que recoge el subcorpus avanzado. Encontramos después las formas infrecuentes (fuera de la lista) en el subcorpus inicial. En ambos casos la presencia de formas pertenecientes a estas franjas de frecuencia es menor de la que sería de esperar. A continuación vienen las formas muy frecuentes en el subcorpus inicial, que, por el contrario, tienen una presencia mayor del valor esperado. Les siguen, ya a una cierta distancia, las formas infrecuentes en el subcorpus avanzado, con una presencia también mayor a la esperable. Las diferencias en el resto de los casos son menos pronunciadas. Tenemos, pues, que nuestras expectativas con respecto a la distribución del vocabulario se vuelven a confirmar en los casos extremos (subcorpus inicial y avanzado y franjas de formas o bien muy frecuentes, o bien relativamente raras, en la medida en que no aparecen en las listas). Se podría decir que en las lecturas de nivel inicial se "abusa" de formas de frecuencia muy alta, mientras que se restringe el uso de formas de frecuencia baja. En las de nivel avanzado tenemos la situación opuesta.

Si en las líneas precedentes hemos comparado los tres subcorpus teniendo en cuenta las listas de vocabulario que se usa en cada uno de ellos (esto es, considerando cada forma una sola vez independientemente de sus repeticiones), atenderemos ahora a qué proporción de las ocurrencias (*tokens*) que forman cada texto corresponden a cada una de las franjas de frecuencia que distinguimos. Se ha considerado que la presencia de una proporción alta de formas infrecuentes es un signo de *sofisticación léxica* (Read, 2000: 203–204) y la esperamos, por tanto, en textos dirigidos a aprendices de nivel avanzado, antes que en textos de nivel inicial o intermedio. Los resultados de este análisis se muestran en el Cuadro 5.

Cuadro 5. Tokens de *Lecturas paso a paso* correspondientes a los *types* de *esTenTen11* ordenados por franjas de frecuencia

		INICIAL	INTERMEDIO	AVANZADO	MARGINAL
1K	Frec.	19343	10555	17618	47516
	% columna	69,11	65,85	67,48	
	Resid. Est	**2,75**	**-2,94**	-0,55	
2K	Frec.	2084	1064	1717	4865
	% columna	7,45	6,64	6,58	

[13] Los residuos estandarizados se obtienen restando a la frecuencia observada la frecuencia esperada y dividiendo por la raíz de esta última —(O-E)/√E—. El resultado nos indica qué diferencia hay con respecto a la frecuencia esperada en una escala estandarizada (comparable, por tanto, independientemente del tamaño de los valores observados).

	Resid. Est	3,23	-1,44	-2,22	
3K	Frec.	1154	601	1020	2775
	% columna	4,12	3,75	3,91	
	Resid. Est	1,40	-1,32	-0,40	
4K	Frec.	762	459	724	1945
	% columna	2,72	2,86	2,77	
	Resid. Est	-0,51	0,68	-0,01	
5K	Frec.	689	362	737	1788
	% columna	2,46	2,26	2,82	
	Resid. Est	-0,92	**-2,31**	**2,76**	
OTROS	Frec.	3955	2,988	4292	11235
	% columna	14,13	18,64	16,44	
	Resid. Est	**-7,90**	**8,29**	1,69	
MARGINAL		27987	16029	26108	70124

De nuevo se aplicó la prueba del chi cuadrado para determinar si las diferencias en cuanto a la composición de los tres subcorpus eran estadísticamente significativas. El resultado fue nuevamente positivo (χ^2=186,25, p<0,000). Atendiendo a los residuos estandarizados de cada casilla nos encontramos, sin embargo, con un panorama diferente al que se nos presentaba al comparar las listas de vocabulario de los tres subcorpus. Aquí son los resultados del subcorpus intermedio los que presentan diferencias más grandes con respecto a los valores esperados. Así, este subcorpus muestra una desviación mayor que los otros dos con respecto a las ocurrencias correspondientes a formas fuera de la lista, que tienen una presencia mayor a la esperada. La siguiente desviación más grande se da en los empleos correspondientes al mismo tipo de formas en el subcorpus de nivel inicial, si bien, en este caso cubren una proporción de texto menor a lo esperable. Las formas muy frecuentes (primera franja) se usan con mucha profusión en nivel inicial, pero el subcorpus que más se aparta de lo esperado en la infrautilización de estas formas es el de nivel intermedio y no el avanzado, como cabría suponer. En este último subcorpus, sin embargo, aparecen sobrerrepresentadas las ocurrencias correspondientes a formas relativamente infrecuentes (quinta franja), que en el inicial tienen una frecuencia menor a la esperada, pero no tanto como en el nivel intermedio. Cabe pensar que parte de la errática distribución del subcorpus de nivel intermedio se debe a su diferencia de tamaño con respecto a los otros dos.

Lo que se ha observado con respecto a la distribución del vocabulario en *Lecturas paso a paso* puede resumirse como sigue:

a) Si se atiende a la diversidad léxica de los tres subcorpus por medio del cociente *type*/*token*, se obtiene que aquella aumenta a la par del nivel de aprendizaje, como sería de esperar. Sin embargo, si es el cociente lema/*token*

la medida empleada, encontramos una diversidad algo mayor en el nivel intermedio que en el avanzado. Todo ello sugiere que el nivel avanzado es el más complejo en cuanto a la morfología, mientras que el nivel intermedio presenta una diversidad de vocabulario algo mayor que el anterior.

b) Si comparamos las listas de formas de palabra que componen el corpus de referencia y el de aprendices, encontramos diferencias significativas en los casos extremos: entre los subcorpus inicial y avanzado y entre formas muy frecuentes (empleadas menos de lo esperado en el avanzado, más en el inicial) y las que están fuera de la lista de las cinco mil más habituales (empleadas menos de lo esperado en el nivel inicial, algo más en el avanzado).

c) Si comparamos las ocurrencias correspondientes a las formas de distinta frecuencia en los tres subcorpus, tenemos que en el nivel inicial las de formas muy frecuentes (1k) cubren proporciones de texto mayores de lo esperado y las infrecuentes (fuera de la lista) proporciones muy por debajo de lo que cabría esperar, pero es el nivel intermedio, y no en el avanzado, donde las ocurrencias de estas formas inusuales tienen mayor representación.

Si bien estos resultados se ajustan solo parcialmente a nuestra hipótesis inicial, parecen congruentes hasta cierto punto con lo que se ha visto en el apartado 2 con respecto a los criterios manejados en la elaboración de algunas lecturas graduadas. Así, el hecho de que los subcorpus intermedio y avanzado presenten una variedad similar de lemas, pero que sea el segundo el que despliega una mayor variedad de *types*, revela que es la diversidad morfológica lo que realmente marca la diferencia entre estos dos niveles. Esto podría ser indicativo de la preponderancia de la gramática como criterio nivelador.

5. Distribución de las colocaciones de *Lecturas paso a paso* según su frecuencia

En esta sección se estudia la distribución de las colocaciones contenidas en los textos de los tres niveles de *Lecturas paso a paso* en relación con la frecuencia que presentan en un corpus general del español. Al igual que en el caso de las formas univerbales, nuestra hipótesis de partida es que las colocaciones menos frecuentes serán introducidas en niveles más avanzados que las frecuentes, ya que, por un lado, parece existir relativo consenso en que el vocabulario frecuente merece atención prioritaria y, por otro, parece razonable asumir que el vocabulario de las lecturas graduadas está de alguna manera controlado. Antes de examinar si esta hipótesis se verifica en los datos, se explicará brevemente cómo se ha determinado la frecuencia de cada colocación.

Tras extraer manualmente del corpus de lecturas graduadas todas las colocaciones identificadas de acuerdo con los criterios expuestos en el apartado 2.4, se determinó su frecuencia en el corpus *esTenTen11*. El punto de partida para cada una de las búsquedas fue el lema de la colocación en cuestión —es decir, un lema complejo que incluía el lema de la base y el del colocativo—. Así, si el corpus contenía, por ejemplo, la colocación *suena el teléfono,* obtuvimos la frecuencia de esta colocación en concreto buscando las ocurrencias correspondientes a los lemas *sonar* y *teléfono* en aquellos contextos que encajasen con la misma relación sintáctica (SUJETO-VERBO) de la

colocación buscada. Para ello se utilizaron consultas en lenguaje CQL que se servían de la anotación morfológica del corpus (para más detalles véase Vincze y Alonso Ramos, 2013).

Una vez obtenida la frecuencia de las colocaciones de nuestro corpus, las agrupamos en diferentes franjas de frecuencia. Estas resultaron de ordenarlas según su frecuencia por millón de acuerdo con una escala logarítmica (de modo similar a lo propuesto por Van Heuven et al., 2014): el primer grupo incluye colocaciones por debajo de 0,1 ocurrencias por millón de palabras (p.m.), el segundo colocaciones de frecuencia igual o mayor a 0,1 p.m. y menores a 1 p.m., y así sucesivamente.

En el cuadro 6 puede verse la distribución de las colocaciones pertenecientes a las distintas franjas en los tres subcorpus de lecturas graduadas. Puesto que la mayoría de las colocaciones solo presentaba una ocurrencia en cada subcorpus[14], se descartaron las (infrecuentes) repeticiones de una misma colocación. La comparación es, por tanto, entre listas de colocaciones de cada subcorpus y no entre todas sus ocurrencias. Además, solo siete presentaban frecuencias por encima de 100 p.m. —una en el subcorpus avanzado, una en el inicial y cinco en el intermedio—. La escasez de este tipo de colocaciones de frecuencia muy alta probablemente está en relación con el hecho de que las colocaciones, en general, muestran frecuencias más bajas que las formas univerbales. Debido a su escaso número, estas siete colocaciones se excluyeron de la comparación, con lo que esta se ha llevado a cabo sobre cuatro franjas de distinta frecuencia.

Cuadro 6. Distribución de colocaciones agrupadas de acuerdo con su frecuencia en *esTenTen11* en los subcorpus de *Lecturas paso a paso*

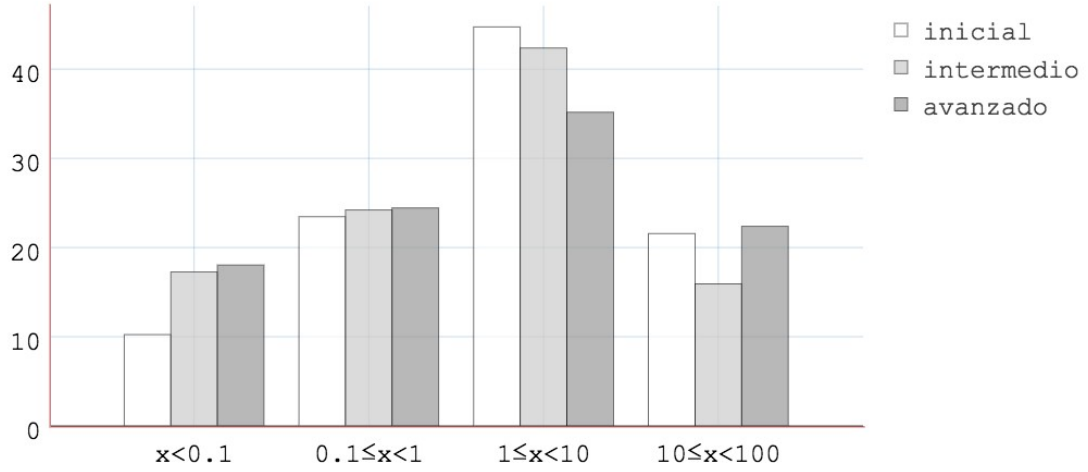
	FRANJA	INICIAL	INTERMEDIO	AVANZADO	MARGINAL
x<0.1	Frec.	38	55	79	172
	% columna	10,24	17,52	18,04	
	Resid. est	**-2,49**	1	1,45	
0.1≤x<1	Frec.	87	76	107	270
	% columna	23,45	24,20	24,43	
	Resid. est	-0,23	0,06	0,16	
1≤x<10	Frec.	166	133	154	453
	% columna	44,74	42,36	35,16	
	Resid. est	1,33	0,56	-1,70	
10≤x<100	Frec.	80	50	98	228
	% columna	21,57	15,92	22,37	
	Resid. est	0,53	-1,72	0,96	
MARGINAL		371	314	438	1123

[14] El cociente lema/token de los tres subcorpus estaba por encima de 0,98 en los tres casos.

http://www.e-scripta.uni.lodz.pl/ojs/index.php/escriptaromanica | ISSN 2392-0718
Received: 20/03/2016 | Accepted: 12/04/2016

La presencia de colocaciones pertenecientes a distintas franjas de frecuencia muestra diferencias significativas dependiendo del nivel considerado, como indica la prueba del chi cuadrado (χ^2=18,625, p=0,005). Es especialmente notable el caso de las colocaciones más infrecuentes (por debajo de 0,1 p.m.) en el nivel inicial. Las diferencias en cuanto a la distribución se pueden apreciar más claramente en la Fig. 1.

Figura 1. Presencia de colocaciones según su frecuencia en los tres niveles de *Lecturas graduadas*



El gráfico muestra que las colocaciones raras (con frecuencia por debajo de 0,1 p.m.) aparecen menos en la lista de colocaciones de las lecturas de nivel inicial que en el resto. Las diferencias en la segunda franja de frecuencias no son marcadas (véanse los residuos estandarizados correspondientes en el Cuadro 6), pero la tendencia es la misma: menor presencia en el nivel inicial que en el intermedio y en este con respecto al avanzado. La tendencia cambia de sentido en las colocaciones pertenecientes a la tercera franja (las penúltimas más frecuentes), con una proporción más alta en la lista correspondiente al nivel inicial que en la del intermedio y mayor en esta, a su vez, que en las de avanzado. En la franja de colocaciones más frecuentes no hay una progresión clara y es en el nivel avanzado donde aparecen mejor representadas. La diferencia más patente (como se aprecia atendiendo a los residuos estandarizados; Cuadro 6) se encuentra, pues, en la escasa presencia de colocaciones muy infrecuentes en el nivel inicial, con lo que podemos concluir que este tipo de combinaciones se evita en textos dirigidos a aprendices de español en los estadios más tempranos del proceso.

Estos resultados podrían considerase en cierta medida sorprendentes si tenemos en cuenta que carecemos de evidencias positivas de que se haya controlado la frecuencia colocacional en las lecturas. Una hipótesis plausible es que los autores de estos materiales evitan combinaciones inusuales en textos dirigidos a aprendices noveles, pero tampoco se puede descartar que

la escasez de tales colocaciones en esos textos sea consecuencia de la menor presencia en el nivel inicial de formas de palabra con frecuencia baja que se verificaba más arriba.

6. Conclusiones

Considerando los datos repasados en este artículo, se puede dar una respuesta parcialmente afirmativa a las preguntas planteadas en el apartado 2. Se puede observar cierta relación entre el nivel de los aprendices a los que van dirigidas las lecturas graduadas estudiadas y la frecuencia tanto de formas univerbales como de colocaciones incluidas en ellas. Nuestra expectativa inicial era que los elementos infrecuentes —formas de palabra en general y también colocaciones— apareciesen en niveles más avanzados que los frecuentes, de manera que la proporción de los primeros fuese aumentando conforme avanzase el nivel. Esta expectativa se cumple cuando comparamos las listas de formas de palabra utilizadas en cada subcorpus. No obstante, la diversidad de lemas del nivel intermedio es mayor que la de los otros dos, así como la proporción de ocurrencias de formas ausentes entre las cinco mil más frecuentes del español. Debido a las diferencias de tamaño entre la muestra correspondiente al nivel intermedio y las de los niveles inicial y avanzado, debemos ser cautelosos a la hora de interpretar estos resultados inesperados.

Como se ha visto en el apartado 2, solo algunas editoriales con colecciones de lecturas graduadas reconocen haber controlado el vocabulario correspondiente a cada nivel por medio de datos de corpus. Además, la revisión de estudios sobre este tipo de textos muestra que se le ha dado un papel más prominente a la gramática que al vocabulario como criterio de nivelación. Esto casa con los datos observados a propósito de la diversidad de *types* y lemas, que revelan que la complejidad morfológica aumenta a través de los sucesivos niveles, pero no se incrementa de forma clara la diversidad léxica del nivel avanzado con respecto del intermedio. No puede descartarse, por lo tanto, que la presencia más escasa de formas de palabra y colocaciones infrecuentes en el nivel inicial sea el resultado de la intuición de los autores de lecturas graduadas con respecto a la frecuencia léxica, antes que del manejo de datos de corpus.

Lo dicho hasta aquí tiene una serie de consecuencias. En primer lugar, la distribución de vocabulario de los materiales de ELE examinados coincide hasta cierto punto con la que resultaría usando información procedente de un corpus —al parecer, independientemente de que se haya controlado este aspecto con la ayuda de este tipo de herramienta—. En segundo lugar y teniendo en cuenta esta correlación, la frecuencia de corpus es una información que se puede aplicar a la nivelación de nuevos materiales ELE (diccionarios, manuales, lecturas graduadas, etc.) para obtener unos resultados que respondan a criterios que parecen aplicarse ya en cuanto a la secuenciación del vocabulario, pero de manera más controlada y consistente. Con este mayor control cabe pensar que se evitarían ciertos hechos contradictorios observados en el estudio, como la falta de una progresión clara en la presencia de formas de palabras de frecuencias intermedias o de colocaciones de frecuencia elevada.

Quedan, con todo, diversas cuestiones pendientes de respuesta. Por ejemplo, si bien hemos observado una relación inversa entre la frecuencia de formas de palabra y colocaciones y el orden en el que se presentan en los materiales, no podemos saber si hay omisiones de elementos frecuentes. Además, dado que nuestro análisis del vocabulario no tenía en cuenta la polisemia, no podemos asegurar que los sentidos más frecuentes de una determinada forma hayan recibido un tratamiento prioritario en contraste con los menos frecuentes (algo sobre lo que advertían hace años Sinclair [1991: 112-113] y Sinclair and Renouf [1985]).

Otra pregunta que queda abierta es la efectividad de este tipo de secuenciación basada en la frecuencia y su posible combinación con otros criterios. A este respecto, estudios recientes (Crossley et al., 2014; González-Fernández y Schmitt, 2015) muestran que, mientras el conocimiento colocacional y la distribución del léxico en las producciones de aprendices a nivel avanzado presentan una elevada correlación en cuanto a su frecuencia con los datos de su input, dicha correlación no se aprecia claramente en aprendices de niveles más bajos.

Bibliografía

ALONSO RAMOS, M. (2012): "Explorando la frecuencia léxica para el Diccionario de colocaciones del español". *Cum corde et in nova grammatica: Estudios ofrecidos a Guillermo Rojo*. T. JIMÉNEZ JULIÁ, B. LÓPEZ MEIRAMA, V. VÁZQUEZ ROZAS Y A. VEIGA (eds.). Santiago de Compostela: Universidade de Santiago de Compostela, pp. 19–40.

BÁEZ MONTERO, I. Y B. SUÁREZ RODRÍGUEZ (2011): "Lectura analógica / lectura digital : 'el papel' de las lecturas graduadas en aprendices de E / LE". *La Red y sus aplicaciones en la enseñanza-aprendizaje del español como lengua extranjera*. C. HERNÁNDEZ GONZÁLEZ, A. CARRASCO SANTANA Y E. ÁLVAREZ RAMOS (eds.). Valladolid: ASELE, pp. 117–128.

BESTGEN, Y. (2014): "Inadequacy of the chi-squared test to examine vocabulary differences between corpora". *Literary and Linguistic Computing,* 29,2, pp. 164–170.

BIBER, D. (1993): "Representativeness in Corpus Design". *Literary and Linguistic Computing*, *8*,4, pp. 243–257.

BOGAARDS, P. (2001): "Lexical Units and the Learning of Foreign Language Vocabulary". *Studies in Second Language Acquisition* 23,3, pp. 321–343.

CROSSLEY, S., T. SALSBURY, A. TITAK & D. MCNAMARA (2014): "Frequency effects and second language lexical acquisition: Word types, word tokens, and word production". *International Journal of Corpus Linguistics,* 19, pp. 301–332.

CRUSE, D. A. (1986): *Lexical Semantics*. Cambridge: Cambridge University Press.

DURRANT, P. (2009): "Investigating the viability of a collocation list for students of English for academic purposes". *English for Specific Purposes*, 28,3, pp. 157–169.

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

GÓMEZ MOLINA, J. R. (2004): "Los contenidos léxico-semánticos". *Vademécum para la formación de profesores*. J. SÁNCHEZ LOBATO E I. SANTOS GARGALLO (dirs.): Madrid: SGEL, pp. 789–810.

GONZÁLEZ FERNÁNDEZ, B. Y N. SCHMITT (2015): "How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure". *ITL - International Journal of Applied Linguistics,* 166,1, pp. 94–126.

HIGUERAS, M. (2006): *Las colocaciones y su enseñanza en la clase de ELE*. Madrid: Arco Libros.

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

LAUFER, B. Y P. NATION (1995): "Vocabulary Size and Use: Lexical Richness in L2 Written Production". *Applied Linguistics,* 16,3, pp. 307–322.

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

MEL'ČUK, I. (1996): "Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon". *Lexical Functions in Lexicography and Natural Language Processing*. L. WANNER (ed.). Amsterdam/Philadelphia: John Benjamins, pp. 37–102.

— (2012): "Phraseology in the language, in the dictionary, and in the computer". *Yearbook of Phraseology*, 3, pp. 31–56.

— (2015): "Clichés, an understudied class of phrasemes". *Yearbook of Phraseology*, 6(1), pp. 35–54

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

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

OAKES, M. P. Y M. FARROW (2007): "Use of the chi-squared test to examine vocabulary differences in English language corpora representing seven different countries". *Literary and Linguistic Computing,* 22,1, pp. 85–99.

READ, J. (2000): *Assessing Vocabulary*. Cambridge: Cambridge University Press.

SÁNCHEZ, A. (2001): *Gran Diccionario de Uso del español actual*. Madrid: SGEL.

SÁNCHEZ, A. & P. CANTOS-GÓMEZ (1997): "Predictability of Word Forms (Types) and Lemmas in Linguistic Corpora. A Case Study Based on the Analysis of the CUMBRE Corpus : An 8-Million-Word Corpus of Contemporary Spanish". *International Journal of Corpus Linguistics,* 2,2, pp. 259–280.

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

SINCLAIR, J. M. Y A. RENOUF (1985): "A lexical learning syllabus for language". *Vocabulary and Language Teaching*. C. RONALD & M. MCCARTHY (eds.). London; New York: Longman, pp. 140–160.

VAN HEUVEN, W. J. B., P. MANDERA, E. KEULEERS & M. BRYSBAERT (2014): "Subtlex-UK: A new and improved word frequency database for British English". *Quarterly Journal of Experimental Psychology,* 67, pp. 1176–1190.

VINCZE, O. Y M. ALONSO RAMOS (2013): "Incorporating Frequency Information in a Collocation Dictionary: Establishing a Methodology". *Procedia - Social and Behavioral Sciences,* 95, pp. 241–248.

Sitografía

ANTHONY, L. (2014): *AntWordProfiler (Version 1.4.1*). Tokyo: Waseda University. <http://www.laurenceanthony.net/>

INSTITUTO CERVANTES (1997-2015): *Plan curricular del Instituto Cervantes. Niveles de referencia para el español*. <http://cvc.cervantes.es/ensenanza/biblioteca_ele/plan_curricular/>.

INSTITUTO CERVANTES (2000-2015): *Lecturas paso a paso*. <http://cvc.cervantes.es/aula/lecturas/>

REAL ACADEMIA ESPAÑOLA (s.f.): *Corpus de Referencia del Español Actual* (CREA). <http://corpus.rae.es/creanet.html>

SAN MATEO VALDEHÍTA, A. (2005): *Una lectura graduada narrativa para estudiantes de ELE de nivel avanzado*. Tesis de máster. http://www.mecd.gob.es/redele/Biblioteca-Virtual/2005/memoriaMaster/1-Semestre/SAN-MATEO-UCM.html

Apéndice: Composición del corpus de lecturas graduadas

Título original (Título en *Lecturas..*)	EDITORIAL	COLECCIÓN	NIVEL ORIGINAL	NIVEL *Lecturas*
El Sueño de Otto	Santillana		nivel 1	inicial
Gente que lee	Difusión		nivel 3	inicial
Amnesia	Edinumen	Lecturas graduadas	A1	inicial
El delfín	Coloquio			inicial
De viaje (Por soñar)	Santillana	Leer en español		inicial
Lejos de casa	Difusión	Lola Lago Detective	A2	inicial
Vuelo 505 con destino a Caracas	Difusión	Primera plana	B1	inicial
El hombre del bar	Santillana		nivel 2	inicial
El misterio de la llave	Santillana		nivel 1	inicial
Paisaje de otoño	Edinumen	Lecturas graduadas	A2	inicial
Doce a las doce	Edelsa	Para que leas	nivel 2	inicial
Un sueño muy extraño	SGEL		intermedio	intermedio
Lola	Edelsa	Para que leas	nivel 3	intermedio
El Secreto de Cristóbal Colón	Santillana		nivel 3	intermedio
Una morena y una rubia	Edelsa	Para que leas	nivel 3	intermedio
De fiesta en verano (La tomatina de Buñol)	Difusión	Aires de Fiesta	A2	intermedio
Memorias de septiembre	Edinumen	Lecturas graduadas	B1	intermedio
Muerte entre muñecos	Edinumen	Lecturas graduadas	B1	intermedio
Guantameras	Difusión	América Latina	A2	intermedio
Una mujer en apuros	SGEL			intermedio
De fiesta en otoño (En busca del oro carmesí)	Difusión	Aires de Fiesta	A2	intermedio
De fiesta en verano (El paso	Difusión	Aires de Fiesta	A2	intermedio

del fuego)				
Una música tan triste	Edinumen	Lecturas graduadas	C2	avanzado
Sobre Iberoamérica (Misiones guaraníes)	SM			avanzado
Una etiqueta olvidada	Difusión	Venga a leer	nivel 4	avanzado
Los polifacéticos	SM			avanzado
Los labios de Bárbara	Edinumen	Lecturas graduadas	C1	avanzado
Noventa y seis horas y media en ninguna parte (cuento chino)	Edelsa	Para que leas	nivel 4	avanzado
Antología de la poesía española hasta el siglo XIX (La poesía del romanticismo)	SGEL			avanzado
Do de pecho	Edelsa	Para que leas	nivel 5	avanzado
España escribe sobre Europa (París de la Belle epoque)	SM			avanzado
Congreso en Granada	Edelsa	Para que leas	nivel 5	avanzado
La última novela	Edinumen	Lecturas graduadas	C2	avanzado
La expedición de Kon Tiki	SM			avanzado