REVISTA SIGNOS. ESTUDIOS DE LINGÜÍSTICA ISSN 0718-0934
 56(112) 282-305

Sofisticación y Diversidad como Medidas de Complejidad Léxica para Determinar el Perfil Colocacional de Textos Académicos en Español

Sophistication and Diversity as Lexical Complexity Measures to Identify the Collocational Profile of Spanish Academic Texts

Eleonora Guzzi
UNIVERSIDADE DA CORUÑA
ESPAÑA
eleonora.guzzi@udc.es

Margarita Alonso Ramos
UNIVERSIDADE DA CORUÑA
ESPAÑA
margarita.alonso@udc.es

Recibido: 07-XII-2021 / **Aceptado:** 04-VII-2022
DOI: 10.4067/S0718-09342023000200282

Resumen

En este estudio proponemos utilizar en el ámbito de las colocaciones los parámetros de sofisticación y diversidad, normalmente restringidos a unidades léxicas monoverbales. Para ello, presentamos una metodología que mide ambos parámetros en textos académicos en español. El objetivo final es comprobar su eficacia para determinar el perfil colocacional de dichos textos, en línea con el *Lexical Profile* (Laufer & Nation, 1995) utilizado para textos en inglés. Entendemos la 'sofisticación' como una propiedad de las colocaciones más prototípicas del discurso académico que se caracterizan por ser contrastivamente menos frecuentes en la lengua general que en dicho discurso. Para establecer los niveles de sofisticación, se implementan dos métodos: uno basado en criterios de frecuencia y otro en índices de Información Mutua (IM). Por otro lado, con 'diversidad' nos referimos a la proporción de colocaciones repetidas en un texto y lo calculamos a través de una adaptación de la medida TTR (Guiraud, 1954). Para comprobar la efectividad de los métodos se clasifican las colocaciones extraídas de 48 textos de un corpus académico de expertos y uno de noveles. Los resultados obtenidos demuestran que los noveles no se caracterizan por utilizar menos colocaciones sofisticadas del discurso académico, sino por una mayor inclusión de colocaciones de baja sofisticación (más propias de la lengua general), así como por un menor valor de diversidad colocacional.

Palabras Clave: Perfil colocacional, sofisticación colocacional, diversidad colocacional, complejidad léxica, discurso académico.

Abstract

In this study the parameters of sophistication and diversity, normally restricted to single lexical units, are proposed in the field of collocations. For this purpose, a methodology

that measures these two parameters in Spanish academic texts is presented. The aim is to test its effectiveness in determining the collocational profile of academic texts, in line with the well-known 'Lexical Profile' (Laufer & Nation, 1995). 'Sophistication' is understood as the property of collocations that are contrastively less frequent in general language and more prototypical in academic discourse. To classify collocations into different levels of sophistication, two methods are implemented: the first based on frequency criteria and the second on the Mutual Information (MI). On the other hand, 'diversity' refers to the number of repeated collocations in a text and is calculated using an adaptation of the TTR measure (Guiraud, 1954). To test the effectiveness of these methods, collocations extracted from 48 texts of an academic expert corpus and a novice corpus are analyzed. Results show that novices are not characterized by using fewer sophisticated collocations of academic discourse, but by a greater inclusion of low sophistication collocations in their texts (more typical of the general language), as well as by a lower value of collocational diversity.

Keywords: Collocational profile, collocational sophistication, collocational diversity, lexical complexity, academic discourse.

INTRODUCCIÓN

La calidad de la escritura de los textos de estudiantes está directamente relacionada con el conocimiento del vocabulario. Para medir la complejidad lingüística de los textos y evaluar su calidad, en el campo de la lingüística cuantitativa se ha introducido el concepto de 'riqueza léxica' o 'complejidad léxica' (Read, 2000; Malvern, Richards, Chipere & Durán, 2004; Bulté & Housen, 2012) que se ciñe a la cantidad y calidad del vocabulario utilizado por un hablante o en una muestra de lengua. En el mundo hispánico, los estudios utilizan más el término 'riqueza léxica' para tratar cálculos de frecuencias y de disponibilidad (Ávila, 1988; López-Morales, 2010; Capsada Blanch & Torruella Casañas, 2017). En el ámbito anglófono, la complejidad léxica se ha empleado para evaluar el conocimiento del vocabulario que poseen los estudiantes de inglés como L2 (Tidball & Treffers-Daller, 2008; Lu, 2012; Bulté & Housen, 2012; Kyle & Crossley, 2015). Para ello, se han propuesto análisis cuantitativos en los que se combinan múltiples parámetros como la 'diversidad léxica', la 'sofisticación léxica' y la 'densidad léxica', que proporcionan indicios sobre la repetición de palabras en un texto, el uso de palabras poco frecuentes y más difíciles de adquirir, y el uso de palabras con contenido léxico, respectivamente.

Independientemente de la metodología aplicada, el análisis de la competencia léxica de los estudiantes resulta imprescindible para poder diseñar algún sistema de evaluación como los incluidos en las certificaciones de español como segunda lengua con fines académicos (Mendoza, 2015), como, por ejemplo, EXELEEA[1] (Examen de Español como Lengua Extranjera para el Ámbito Académico). Se han propuesto numerosos estudios sobre el español con fines académicos (Vázquez, 2000, 2005; Parodi, 2010; Montolío, 2014), pero son escasos los análisis del léxico que emplean los estudiantes en sus producciones escritas. El discurso académico comprende un vocabulario más específico que los estudiantes no siempre conocen o saben emplear

de forma apropiada, y que se aprende por medio de la experiencia en el ámbito. Específicamente, dentro de este vocabulario, las combinaciones léxicas desempeñan un papel fundamental en el proceso de adquisición de la competencia léxica de una lengua (Paquot, 2018), ya que no solo es importante conocer las palabras, sino también sus combinaciones en cada contexto. Como consecuencia, dada la escasez de estudios que se centran en la complejidad de combinaciones léxicas en el español académico para nativos, en contraste con los numerosos estudios sobre inglés académico para no nativos, en el presente artículo presentamos un método para medir la complejidad léxica de las 'colocaciones' empleadas por estudiantes nativos en sus producciones académicas escritas.

Para este propósito, definimos las colocaciones, dentro del marco de la Lexicografía Explicativa y Combinatoria (Mel'čuk, 2012), como combinaciones léxicas que están formadas por una 'base' y un 'colocativo', cuyos elementos tienden a coocurrir, como, por ejemplo, 'plantear hipótesis' o 'arrojar resultados'. La selección de las dos unidades léxicas que las componen se lleva a cabo de formas distintas: la base se selecciona semánticamente, mientras que el colocativo se selecciona léxicamente (Alonso-Ramos, 2010). Puesto que nos centramos en el estudio de la complejidad únicamente de las colocaciones de entre el resto de posibles combinaciones léxicas existentes (locuciones, fórmulas, entre otras), nos referiremos al concepto de 'complejidad fraseológica' (Paquot, 2019) como 'complejidad colocacional', y la analizaremos a través de dos parámetros: la 'sofisticación colocacional' y la 'diversidad colocacional'.

Con sofisticación colocacional nos referimos a la propiedad de las colocaciones que son más frecuentes en el discurso académico que en la lengua general y, por lo tanto, más prototípicas de dicho discurso. Partimos de la hipótesis de que los expertos muestran una mayor riqueza de vocabulario y un menor uso de colocaciones que se adscriben a la lengua general, debido a la experiencia en el ámbito de redacción académica, así como un mayor contacto con varios géneros textuales en el campo de la investigación. Por ejemplo, en textos académicos se prioriza el uso de verbos más específicos, como 'efectuar un análisis', frente a otros más genéricos, como 'hacer un análisis', o se hace uso de colocaciones más prototípicas de este discurso, como 'contrastar hipótesis' u 'obtener muestras'. Para establecer los niveles de sofisticación, se implementan dos métodos, uno basado en criterios de frecuencia y otro en índices de Información Mutua (IM). A su vez, nuestra segunda hipótesis plantea que los expertos utilizan una mayor diversidad de colocaciones académicas y que repiten con menor frecuencia el vocabulario. Para determinar si repiten frecuentemente determinadas colocaciones, complementamos el estudio de la sofisticación con la diversidad colocacional. Siguiendo este planteamiento, los parámetros que conforman la complejidad colocacional, aquí la sofisticación y la diversidad, permitirían identificar una parte del 'perfil colocacional' de textos académicos, entendido como una

representación de la 'competencia colocacional' del autor del texto, esto es, el conocimiento de las colocaciones propias de dicho discurso y la capacidad de utilizarlas correctamente.

Por consiguiente, los objetivos principales de este artículo son, por un lado, fijar los criterios para establecer las franjas de sofisticación para las colocaciones del discurso académico y, por otro lado, aplicar las medidas de sofisticación y diversidad colocacional que permitan contrastar el perfil de expertos, investigadores/as que publican artículos, y estudiantes noveles en sus trabajos de fin de grado/master. Las preguntas de investigación planteadas en relación con los objetivos son las siguientes:

1. ¿Cuál de los métodos propuestos es más eficiente para determinar la sofisticación?

2. ¿Los expertos escriben textos con colocaciones más sofisticadas que los noveles? ¿Producen textos con una mayor diversidad colocacional? ¿Existe correlación entre sofisticación y diversidad?

3. ¿La sofisticación colocacional varía dependiendo del dominio académico?

En la sección 1, empezaremos presentando el estado de la cuestión sobre la complejidad léxica. Seguidamente, en la sección 2, expondremos los corpus académicos y las colocaciones que hemos empleado para el estudio, así como la metodología para determinar la sofisticación y diversidad colocacional. En la sección 3, aplicaremos las medidas a una muestra de textos académicos cuyos resultados serán discutidos en la sección 4. Por último, expondremos las conclusiones derivadas del estudio y las líneas futuras.

1. Estudios sobre complejidad léxica

En las últimas décadas, se ha analizado la complejidad léxica principalmente en estudios sobre inglés como lengua extranjera para valorar los niveles de competencia léxica de los aprendices de una segunda lengua (Malvern et al., 2004; Yu, 2010; Crossley, Salsbury & McNamara, 2012). A pesar de que una gran parte de los estudios se centran en textos de L2, consideramos que estos parámetros pueden ser extrapolados a los textos de noveles, puesto que también están adquiriendo una competencia léxica, pero, en este caso, académica.

Los parámetros en los que nos centraremos son la diversidad y la sofisticación. La diversidad léxica se trata de un índice de frecuencia que se emplea para contabilizar la proporción de palabras únicas de un texto –*types*– frente a todas las palabras empleadas –*tokens*– (McCarthy & Jarvis, 2010). Una de las fórmulas más conocidas para medir la diversidad léxica es la TTR, del inglés *Type-Token Ratio* (Guiraud, 1954), que divide el número de *types* entre el número de *tokens*. Se han aplicado otras medidas, como el MWF (*Mean Word Frequency*; Tweedie & Baayen, 1998), que calcula la ratio de

tokens/types, o el RTTR (*Root Token-Type Ratio*; Guiraud, 1954), entre otras. Sin embargo, como ya se ha señalado (McCarthy & Jarvis, 2010), estas medidas no tienen en cuenta la longitud de los textos y presentan un bajo nivel de representatividad: cuantas más palabras tiene un texto, más aumenta el número de *tokens*, pero disminuye relativamente el número de *types*, debido a que tendemos a repetir determinadas palabras para mantener la cohesión del texto. Para resolver este problema, se han propuesto otras medidas, como la MSTTR (*Mean Segmental Type-Token Ratio*; Johnson, 1944) o la MATTR (*Moving-Average-Type-Token-Ratio*; Covington & McFall, 2010), que ha sido validada como la medida más estable y sensible a los efectos de la longitud del texto, así como la *vocd-D* o la MTLD (McCarthy & Jarvis, 2010), que reducen el efecto del tamaño segmentando los textos en pequeños fragmentos.

Por otra parte, la sofisticación léxica caracteriza al conjunto de palabras que son poco frecuentes y que indican, por consiguiente, un nivel más avanzado de conocimiento de la lengua (Bardel, Gudmundson & Lindqvist, 2012). Read (2000: 200) la define del siguiente modo: "The use of technical terms and jargon as well as the kind of uncommon words that allow writers to express their meanings in a precise and sophisticated manner". En este sentido, las palabras frecuentes son más sencillas de aprender y producir que las infrecuentes (Laufer & Nation, 1995) y se aprenden de forma más temprana durante el proceso de adquisición de una lengua (Bardel et al., 2012). Como consecuencia, los estudiantes que poseen una escasa competencia léxica, utilizan menos vocabulario sofisticado (Laufer & Nation, 1995). Para identificar las palabras sofisticadas frente a las no sofisticadas, la mayoría de estudios recurren al contraste de frecuencia con un corpus de referencia (Laufer & Nation, 1995), donde las palabras menos frecuentes en el corpus de referencia se consideran sofisticadas (Lu, 2012; Kyle & Crossley, 2015). Las principales fórmulas empleadas para obtener el valor de sofisticación léxica son la LS1 (*Lexical Sophistication* 1) y la LS2, que calculan la proporción de palabras sofisticadas frente al total de unidades léxicas de un texto. La LS1 utiliza como unidad de medida los *types*, mientras que la LS2 utiliza los *tokens*. También existen otras medidas, como la VS1 (*Verb Sophistication 1)*, la VS2 (*Verb Sophistication 2*) o la CVS1 (*Corrected Verb Sophistication 1*).

A pesar de que la sofisticación léxica se considere un componente fundamental en la competencia escrita, la mayoría de estudios se han centrado en evaluar la sofisticación de unidades léxicas monoverbales (Paquot, 2018) y encontramos sólo algunos que tratan la sofisticación fraseológica en colocaciones (Orol González & Alonso-Ramos, 2013; Paquot, 2018, 2019). Orol González y Alonso-Ramos (2013) consideran sofisticadas las colocaciones que presentan una base poco frecuente, es decir, un máximo de 3 ocurrencias por millón de palabras (siguiendo a Almela, Cantos, Sánchez, Sarmiento & Almela, 2005), comparándolas con el corpus de referencia *esTenTen* (Kilgariff & Renau, 2013). En el estudio de Paquot (2019), en cambio, se proponen dos métodos a partir de las colocaciones de un corpus de

aprendices de inglés como L2. En el primer caso, se contrastan con las colocaciones que aparecen en la *Academic Collocation List* (ACL, Ackermann & Chen, 2013), y se considera que una colocación es sofisticada si se encuentra en las 2.000 colocaciones académicas de dicha lista. En el segundo caso, se calcula el valor de IM de las colocaciones identificadas en los textos y se consideran sofisticadas las colocaciones con un valor de IM alto (>3). Siguiendo estos criterios, se aplican las fórmulas LS1 y LS2 para comparar los textos de cada nivel calculando la proporción de colocaciones sofisticadas.

Con el fin de determinar estos índices automáticamente y medir el perfil léxico de un texto, algunos autores han propuesto herramientas en línea que han servido como referencia para este estudio. Entre ellas destacan el *Lexical Frequency Profile* (LFP) de Laufer y Nation (1995), que ofrece un perfil léxico de los textos tras medir la proporción de palabras de baja frecuencia frente a las de alta frecuencia de textos de aprendices de inglés como L2, el *Lexical Oral Production Profile* (LOPP) propuesto por Bardel et al. (2012), que calcula la sofisticación léxica de textos orales de italiano y francés como L2; el *L2 Lexical Complexity Analyzer*, de Ai y Lu (2010), que aplica varios parámetros de complejidad léxica o la herramienta *Tool for the Automatic Analysis of Lexical Sophistication*, (TAALES; Kyle & Crossley, 2015), que obtiene puntuaciones para 135 índices relacionados con la frecuencia, los n-gramas, la dispersión, entre otros. En esta línea, las medidas empleadas en este estudio para el cálculo de la complejidad colocacional pretenden ser un punto de partida para la creación de una herramienta que identifique automáticamente el perfil colocacional de textos académicos y que, a su vez, pueda ser empleada como recurso en el ámbito de certificaciones de español con fines académicos.

2. Metodología

En las secciones 2.1. y 2.2. expondremos los corpus y las colocaciones académicas que hemos empleado para el estudio. En la sección 2.3, explicaremos cómo determinamos las franjas de sofisticación colocacional, siguiendo criterios de frecuencia como primer método (2.3.1.) y valores de IM, como segundo método (2.3.2.); y en la sección 2.4, presentaremos un método para el cálculo de la diversidad colocacional.

2.1. Corpus

Nos basamos en dos corpus académicos, HARTA-Expertos-Plus (HEP) y HARTA-Noveles (HN). El corpus académico HARTA-Expertos-Plus contiene 21.068.482 palabras procedentes de 3.870 artículos de investigación: 19.043.390 palabras proceden del corpus académico-científico *Iberia* (Ahumada, Zamorano, García & Lara, 2011) y 2.025.092 palabras provienen del corpus HARTA-Expertos (Alonso-Ramos, García-Salido & Garcia, 2017), formado por textos de la sección en español del *Spanish-English Research Articles Corpus* (SERAC; Pérez-Llantada, 2014) y

por artículos recogidos de otras revistas científicas. HEP es un corpus etiquetado, lematizado y analizado sintácticamente. Se divide en cuatro dominios principales, que a su vez están subdivididos en disciplinas, siguiendo la estructura del SERAC: Artes y Humanidades (AH), Biología y Ciencias de la Salud (BM), Ciencias Físicas e Ingeniería (CF) y Ciencias Sociales y Educación (CS). Como corpus de referencia, a su vez utilizado para establecer las franjas de sofisticación, seleccionamos el corpus de lengua general *esTenTen* (Kilgariff & Renau, 2013), que contiene 20.306.642.991 palabras procedentes de textos de lengua general de páginas web, como periódicos digitales, páginas de Wikipedia, blogs, etc. Por último, contamos con el corpus *HARTA-Noveles* (HN; Villayandre Llamazares, 2019), que contiene 176 textos, repartidos entre trabajos de final de grado (TFG) y de final de máster (TFM) de varias universidades españolas y que cuenta con 2.230.153 palabras. Los textos están clasificados siguiendo la estructura de HEP y se emplean para analizar el perfil colocacional de los textos de noveles y llevar a cabo una comparativa con los textos de expertos.

2.2. Extracción de colocaciones académicas

Las colocaciones académicas utilizadas para determinar el perfil colocacional se han obtenido de forma semiautomática. En primer lugar, se procesó sintácticamente el corpus HEP para extraer colocaciones de 5 relaciones de dependencias sintácticas (N + N, V + Obj, Suj + V, N + Adj, V + Obl), a partir de una lista de nombres académicos (García-Salido, 2021). Seguidamente, se realizó una extracción automática a partir del corpus HEP, basada en medidas de asociación estadísticas (*log-likelihood*, IM, entre otras), y criterios de frecuencia (≥5 ocurrencias) (Alonso-Ramos, García-Salido, Garcia & Guzzi, 2019). En tercer lugar, un grupo de anotadores refinó manualmente los candidatos de colocaciones extraídos automáticamente para obtener un listado final de colocaciones académicas. En este estudio, nos centramos únicamente en el análisis de la relación sintáctica V + Obj, que conforman un total de 1.911 colocaciones del discurso académico.

2.3. Establecimiento de las franjas de sofisticación colocacional

Para establecer las franjas de sofisticación colocacional, que se emplearán para medir las colocaciones extraídas de textos de los corpus HEP y HN, implementamos dos métodos: el primero se basa en el contraste de frecuencias y el test estadístico *log-likelihood* (LL) y, el segundo, en valores de IM.

2.3.1. Primer método: Franjas por frecuencias

Contrastamos las frecuencias de las colocaciones entre *esTenTen* y HEP para crear franjas que agrupan colocaciones en una escala de mayor a menor nivel de sofisticación. Para asignar un grado de sofisticación a las colocaciones, consideramos que las colocaciones utilizadas por los expertos, como, por ejemplo, 'refutar una hipótesis', tienen un mayor nivel de sofisticación por ser más prototípicas del discurso

académico y menos frecuentes en la lengua general. En este sentido, si una colocación presenta una diferencia significativa de frecuencia entre el corpus de referencia y el corpus académico, y en el corpus de referencia su frecuencia se aproxima a 0, se ubicará en una franja de sofisticación más alta.

En primer lugar, anotamos las frecuencias absolutas y normalizadas, por 1 millón de palabras, del corpus *esTenTen* y del corpus académico HEP de las 1.911 colocaciones. Tras obtener los datos, separamos las colocaciones que presentan un mayor número de ocurrencias en el corpus de lengua general y establecimos la primera franja (F1) asociada a un nivel de sofisticación baja. A pesar de que partimos de colocaciones únicamente académicas, las colocaciones del discurso académico se encuentran en un continuum entre la lengua general y la lengua académica (Hyland & Tse, 2007), en contraste con las colocaciones especializadas (L'Homme, 2000). Colocaciones como 'hacer trabajo' o 'hacer sugerencia' no se consideran sofisticadas por no ser prototípicas del discurso académico; sin embargo, pueden igualmente utilizarse, aunque con menos frecuencia, para hacer referencia a determinados procesos o actividades académico-científicas.

Seguidamente, calculamos el *log-likelihood* de las colocaciones más frecuentes en la lengua académica para separar las colocaciones que muestran una diferencia significativa entre el corpus académico y la lengua general de las que no presentan una diferencia significativa. Tras obtener los valores de cada colocación, establecimos el punto de corte en 6,63 (p <0,01) y agrupamos en la franja 2 (F2) las colocaciones que se encuentran por debajo del punto de corte, asociadas a un nivel de sofisticación moderado. Finalmente, a partir de las colocaciones que presentaron una diferencia significativa, establecimos las dos últimas franjas de sofisticación en función de su frecuencia normalizada en la lengua general: las colocaciones con una frecuencia ≥1, es decir, que son relativamente frecuentes en la lengua general, se agrupan en la franja 3 (F3), considerada como sofisticación alta, y las colocaciones con una frecuencia <1 en la lengua general se agrupan en la franja 4 (F4), asociada a una sofisticación muy alta. Como resultado, obtuvimos cuatro franjas de sofisticación, como podemos observar en la Tabla 1:

Tabla 1. Franjas de sofisticación por *log-likelihood* y frecuencia.

FRANJAS	TIPO DE SOFISTICACIÓN	FRECUENCIA	Nº DE COLOCACIONES	PORCENTAJES
1	Baja	esTenTen > HEP	365	19,09%
2	Moderada	esTenTen < HEP no dif. significativa	636	33,28%
3	Alta	dif. sign. (esTenTen ≥1)	625	32,71%
4	Muy alta	dif. sign. (esTenTen <1)	285	14,92%

Tomando como referencia estas franjas de sofisticación, aplicamos las fórmulas LS1, en línea con Paquot (2019), que calcula la proporción de *types* sofisticados entre *types* totales del texto, para identificar la proporción de colocaciones sofisticadas que aparecen en los textos analizados. En relación con la distinción entre *type* y *token*, es necesario detenernos en cómo contabilizamos las colocaciones. En este estudio, llevamos a cabo una distinción entre 'lema colocacional' (Orol González & Alonso-Ramos, 2013) y *token*, es decir, la forma base que representa todas las formas flexionadas o variantes morfológicas de una colocación. Este planteamiento contrasta con la visión tradicional de *type*, ya que este no representa en una única ocurrencia dos formas distintas de una misma colocación. Por ejemplo, en 'daremos importancia' y 'dieron importancia', identificamos 2 *tokens* y 2 *types*, pero únicamente 1 lema colocacional. Para calcular la LS1 y la diversidad colocacional (vid. 2.4.), utilizaremos las fórmulas 'lemas colocacionales sofisticados/lemas colocacionales' y 'lemas colocacionales/tokens colocacionales', respectivamente. Con este primer análisis, contrastamos los datos de los textos para observar si existen diferencias significativas de sofisticación considerando los valores obtenidos a partir de la LS1 de la franja 1 y 3-4 conjuntamente (colocaciones poco sofisticadas –F1– y muy sofisticadas –F3 y F4–). Las fórmulas utilizadas se presentan en la Tabla 2:

Tabla 2. Fórmulas empleadas para la sofisticación V+Obj con el primer método.

	Descripción	**Fórmula**
LS1vobj F1	Frecuencia (LS1 - franja 1)	Lemas Franja1/Lemas totales
LS1vobj F3-4	Frecuencia (LS1 - franja 3-4)	Lemas Franja3+4/Lemas totales

2.3.2. Segundo método: Franjas por Información Mutua

En relación con el segundo análisis, seguimos a Paquot (2019) para identificar la sofisticación a través de la IM y fijamos el valor ≥3 como valor estándar. Esta medida podría resultar más eficiente en el contraste de colocaciones de nativos frente a no nativos, ya que separa las combinaciones léxicas con un bajo nivel de idiomaticidad de las colocaciones fraseológicas (Durrant & Schmitt, 2009; Granger & Bestgen, 2014; Paquot, Gries & Yoder 2020); no obstante, adaptamos esta medida para identificar las colocaciones más sofisticadas que se encontrarían por encima del valor establecido. Seguidamente, dividimos las colocaciones en tres franjas: las colocaciones que presentan un valor <3 (F1), con una sofisticación baja; las colocaciones que presentan un valor entre 3-6 (F2), asociadas a una sofisticación moderada; y las que presentan un valor >6 (F3), que corresponde a una sofisticación alta, como se muestra a continuación (Tabla 3):

Tabla 3. Franjas de sofisticación por IM.

FRANJAS	TIPO DE SOFISTICACIÓN	VALOR DE IM	Nº DE COLOCACIONES	PORCENTAJES
1	Baja	<3	1.201	63%
2	Moderada	3-6	650	34%
3	Alta	>6	59	3%

Por medio de este valor, calculamos la proporción de colocaciones de cada grupo de IM, centrándonos, al igual que en el caso anterior, en los valores obtenidos a partir de la LS1 de sofisticación baja (F1) y sofisticación alta (F2-3). A continuación, se muestran las fórmulas utilizadas en este segundo análisis (Tabla 4):

Tabla 4. Fórmulas empleadas para la sofisticación colocacional V+Obj con el segundo método.

	Descripción	**Fórmula**
LS1 vobj IM F1	IM (LS1 - franja 1)	Lemas IMFranja1/Lemas totales
LS1 vobj IM F2+3	IM (LS1 - franja 2+3)	Lemas IMFranja2+3/Lemas totales

2.4. Análisis de la diversidad colocacional

Para calcular la diversidad en los textos de nuestros corpus, empleamos una de las fórmulas más tradicionales propuestas en la bibliografía, la TTR (Guiraud, 1954). Dicha fórmula, en este caso denominada LTR, que se muestra en la Tabla 5, calcula el ratio (R) de los lemas colocacionales (L) entre los tokens (T) de las colocaciones:

Tabla 5. Fórmula empleada para la diversidad colocacional.

	Descripción	**Fórmula**
DC vobj	Diversidad colocacional V + Obj	Lemas/Tokens

Debido a que nuestro propósito es identificar los lemas y los *tokens* de colocaciones utilizados en cada texto, sin tener en cuenta todas las palabras del texto, empleamos la medida LTR, a pesar de que los textos de expertos y noveles presenten tamaños distintos. Un resultado que se aproxime a 1 representaría que todas o casi todas las colocaciones utilizadas en un texto son distintas.

3. Cálculo de la sofisticación y diversidad colocacional en textos de expertos y noveles

Con el fin de determinar la sofisticación y diversidad colocacional de los textos, analizamos 24 textos del corpus de expertos y 24 textos del corpus de noveles (*N=48*), 6 textos de cada uno de los 4 dominios, respectivamente, aplicando las distintas fórmulas de sofisticación y diversidad colocacional. Este análisis se concibe como un estudio piloto debido al tamaño reducido de la muestra. A continuación, presentamos un ejemplo de análisis de un texto del corpus de noveles (Tabla 6):

Tabla 6. Ejemplo de análisis de un texto del corpus HN.

Título del texto		*La regulación del dinero político*
Dominio		Ciencias Sociales
Nº colocaciones y palabras del texto		46 colocaciones; 11.834 palabras
M1	**LS1 Sofisticación baja**	0,17
M1	**LS1 Sofisticación alta**	0,63
M2	**LS1 Sofisticación baja**	0,46
M2	**LS1 Sofisticación alta**	0,57
Diversidad colocacional (LTR)		0,68

Una vez obtenidos los datos de los 48 textos, comprobamos la homogeneidad de las varianzas con el test de Levene (p>0,05) y los supuestos de normalidad con la prueba de Shapiro-Wilk para aplicar las medidas estadísticas que se explican a continuación, obteniendo como resultado una varianza constante y una distribución normal (p>0,05). En primer lugar, contrastamos las medias entre expertos y noveles en relación con cada parámetro de sofisticación y diversidad, cuyos resultados se explican en la sección 4.1. En segundo lugar, empleamos un modelo de regresión logística binaria para comprobar qué variables de sofisticación (baja-alta con ambos métodos) y diversidad (*tokens* colocacionales, lemas colocacionales y LTR) pueden predecir si un texto es experto o novel. Los resultados obtenidos a partir de dicho test se presentan en la sección 4.2. En último lugar, utilizamos la prueba de ANOVA de un factor, seguida de la prueba Tukey como test *posthoc* para la comparación de la sofisticación colocacional entre los dominios de cada grupo y mostramos los resultados correspondientes en la sección 4.3. Para los test estadísticos utilizados, establecimos el nivel de significación en 0,05. Todos los análisis se llevaron a cabo con Jamovi[2]. En la Tabla 7 se muestran los datos descriptivos de ambos grupos (Expertos = E; Noveles = N):

Tabla 7. Medias, desviacion estándar y varianza para las medidas de complejidad colocacional.

	Grupos	**LS1 SBaja**	**LS1 SMod**	**LS1 SAlta**	**LS1 SBaja IM**	**LS1 SAlta IM**	**Div**
Media	E	0,128	0,237	0,627	0,627	0,388	0,833
	N	0,228	0,19	0,58	0,598	0,463	0,79
DE	E	0,0556	0,064	0,015	0,015	0,0613	0,131
	N	0,0806	0,0589	0,0316	0,0618	0,032	0,121
Varianza	E	0,00309	0,00409	2,25E-04	2,25E-04	0,00376	0,014
	N	0,00649	0,00347	1,00E-03	0,00383	0,00102	0,017

4. Resultados y discusión

A continuación, presentamos los resultados obtenidos a partir del análisis cuantitativo: en la sección 4.1. mostramos las medias de los parámetros de sofisticación y diversidad para observar el contraste entre los textos de expertos y noveles; en la sección 4.2. exponemos los parámetros que podrían caracterizar a los

textos de cada grupo, derivados del test de regresión logística binaria; y, en la sección 4.3. los resultados obtenidos a partir del análisis de la sofisticación por dominios.

4.1. Contraste de medias de sofisticación y diversidad entre expertos y noveles

Con respecto al primer método (M1), basado en frecuencias, podemos inferir que la media de colocaciones de sofisticación alta es semejante en expertos (0,627) y en noveles (0,598), pero que los expertos utilizan un menor porcentaje de colocaciones de sofisticación baja: 0,128 de expertos frente a 0,228 de noveles. Algunos ejemplos de las colocaciones de sofisticación alta encontradas en los textos de expertos podrían ser 'aplicar criterio' o 'adquirir competencia', y en los textos de noveles otras como 'proporcionar conocimiento' o 'incrementar resistencia'. En relación con las colocaciones de sofisticación baja, en expertos encontramos ejemplos como 'contener información' o 'tener interés', y en los textos de noveles algunas como 'formar parte' o 'tomar decisión'. En los Gráficos 1 y 2 es posible observar la diferencia de medias:

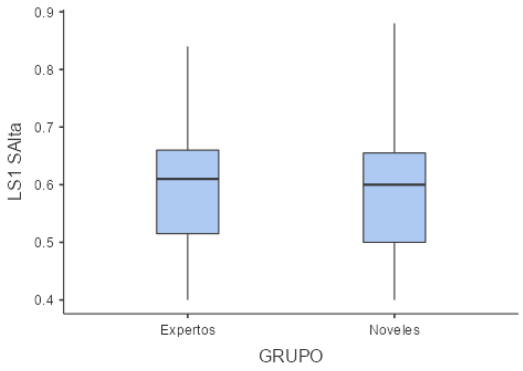

Gráfico 1. Sofisticación alta (M1).

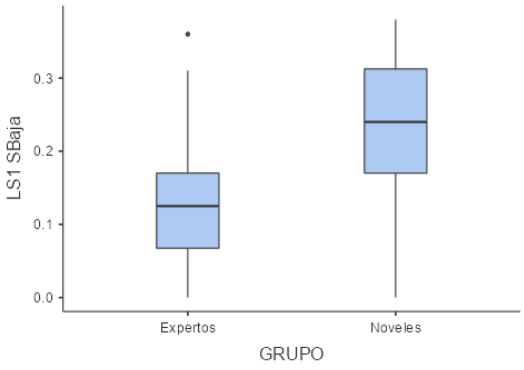

Gráfico 2. Sofisticación baja (M1).

En relación con el segundo método (M2), basado en IM, no parecen existir diferencias entre expertos y noveles ni en las medias sofisticación alta (Gráfico 3) ni en las medias de sofisticación baja (Gráfico 4):

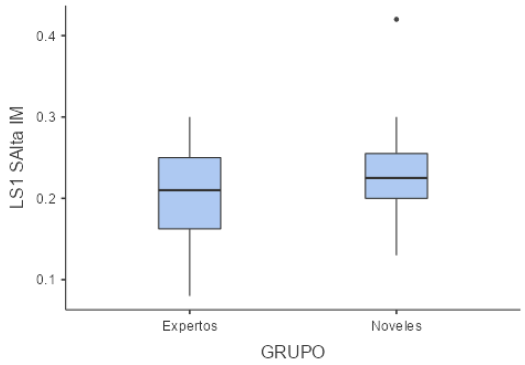

Gráfico 3. Sofisticación alta (M2).

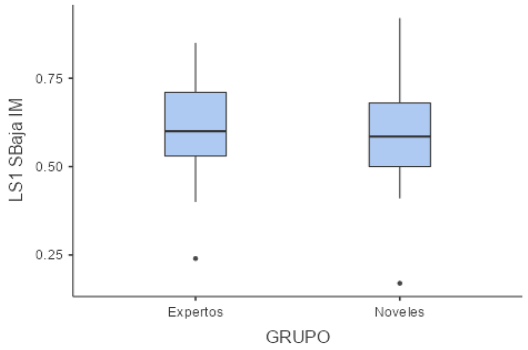

Gráfico 4. Sofisticación baja (M2).

Podemos concluir que en esta muestra de textos los expertos utilizan un menor número de colocaciones de baja sofisticación y que el método que determina la diferencia es el método 1, pues el método 2, tal y como indicamos en la sección 2, parece ser más eficiente para el contraste de colocaciones empleadas por nativos y no nativos.

Con respecto a los índices de diversidad colocacional, aplicamos la fórmula LTR a las colocaciones de los textos analizados para comprobar la hipótesis planteada, esto es, que los textos de expertos repiten un menor número de colocaciones en un mismo texto. A continuación, se muestran las medias de *tokens* colocacionales (Gráfico 5) y lemas colocacionales (Gráfico 6), que se normalizaron por 1.000 palabras en cada texto debido al distinto tamaño de los mismos (6.401 en expertos frente a 15.650 en noveles), y de LTR (Gráfico 7):

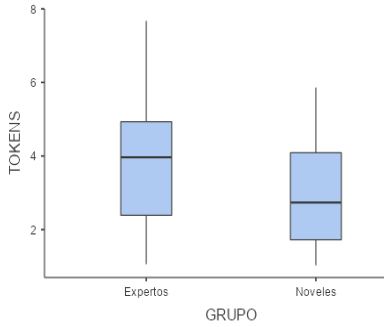

Gráfico 5. Medias de tokens colocacionales.

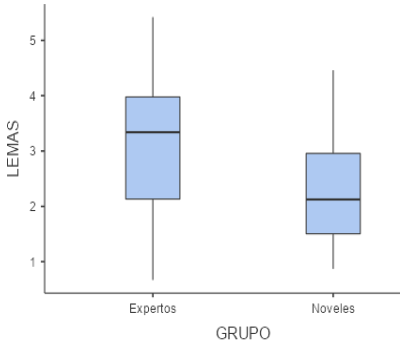

Gráfico 6. Medias de lemas colocacionales.

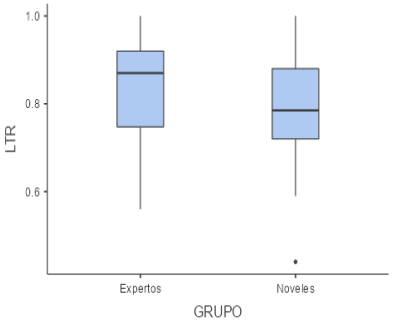

Gráfico 7. Medias de diversidad colocacional (LTR).

Como podemos observar, se produce un ligero incremento de media de diversidad colocacional en los textos de expertos frente a los de noveles, así como en la proporción de lemas y de tokens colocacionales.

4.2. Selección de variables relevantes para determinar la clasificación de los textos académicos

Para comprobar qué variables de sofisticación y diversidad pueden predecir si un texto es experto o novel, calculamos una regresión logística binaria. Una vez observadas las medias de cada grupo para cada parámetro (sección 4.1.), calculamos el coeficiente de Pearson para corroborar que la diversidad y sofisticación no tienen correlación en los textos de expertos (r = 0,188, p=0,37) ni en textos de noveles (r = 0,04, p=0,83), es decir, un porcentaje alto de colocaciones sofisticadas no implica un valor alto de diversidad colocacional ni viceversa.

Tras aplicar varios modelos de regresión eliminando aquellas variables que claramente no eran significativas (p. ej., sofisticación alta), seleccionamos el modelo que incluye las variables independientes de tokens colocacionales, LTR (diversidad) y LS1 SBaja (sofisticación baja). De acuerdo con las observaciones presentadas en la sección anterior (4.1.), de las tres variables predictivas, la LTR (p=0,040) y LS1 SBaja (p=0,008) fueron significativas, mientras que la variable de tokens colocacionales no fue significativa (p=0,072). Este modelo explicaría el 38,5% (R^2 de Nagelkerke) de la varianza en expertos, lo que indica que tiene un escaso valor predictivo, pues, a pesar de que las variables significativas son predictivas, sería necesario complementar el modelo con otras variables para aumentar su predictibilidad. Considerando todo lo expuesto, clasificaríamos correctamente el 77% de los casos con una sensibilidad del 75% y especificidad del 79%. Si observamos las medias marginales estimadas de LTR (Gráfico 8) y LS1 SBaja (Gráfico 9), se podría predecir que un texto es experto (GRUPO=1) cuanto mayor es el valor de diversidad colocacional (LTR) y menor es el de sofisticación baja (LS1 SBaja):

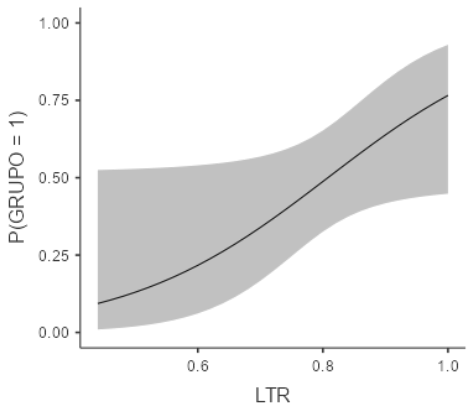

Gráfico 8. Media marginal estimada para LTR.

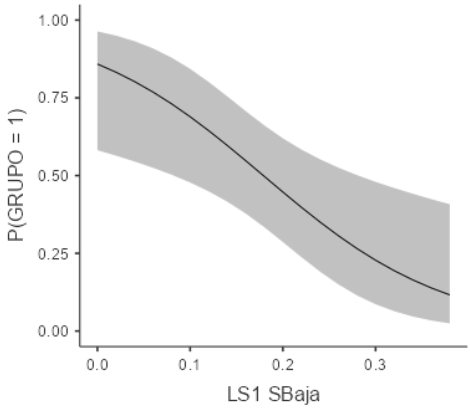

Gráfico 9. Media marginal estimada para LS1 SBaja.

Podemos concluir, pues, que las diferencias significativas destacan sobre todo en el contraste de colocaciones de sofisticación baja utilizadas por ambos grupos, con una correlación negativa con los expertos. El modelo podría resultar poco predictivo debido al número de variables predictivas incluidas, así como los expertos presentan pocas colocaciones de sofisticación baja y los noveles, en algunos casos, utilizan el mismo promedio que los expertos, pero, en otros casos, utilizan una gran cantidad de colocaciones de sofisticación baja en comparación tanto con expertos como con el resto de noveles, esto es, existe mucha variabilidad entre ellos.

4.3. Análisis de la sofisticación por dominios

Con el fin de no sólo observar esta variabilidad dentro del grupo de noveles, sino que también determinar si existen diferencias entre los dominios de expertos, realizamos un contraste de sofisticación entre los cuatro dominios del corpus HEP y del corpus HN (Biología y Ciencias de la Salud, Ciencias Sociales, Artes y Humanidades, y Ciencias Físicas e Ingeniería). En el caso de los expertos, observamos que los dominios de Artes y Humanidades y Ciencias Físicas presentan una proporción más alta de colocaciones de sofisticación alta (LS1 SAlta) (Gráfico 10) y que el dominio de Artes y Humanidades, a su vez, presenta un mayor porcentaje de colocaciones de la franja de sofisticación baja (LS1 SBaja), junto con el dominio de Ciencias Sociales (Gráfico 11):

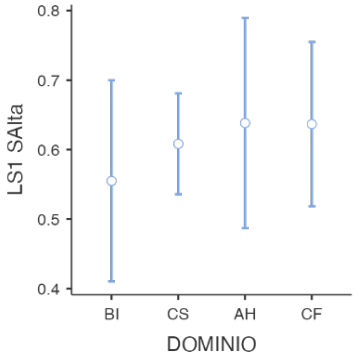

Gráfico 10. LS1 SAlta en expertos por dominio.

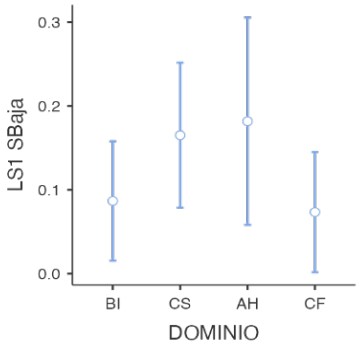

Gráfico 11. LS1 SBaja en expertos por dominio.

Sin embargo, las diferencias no son estadísticamente significativas (p=0,6 y p=0,09). Contrariamente a los textos de expertos, en los textos de noveles es el dominio de Biología y Ciencias de la Salud el que presenta más colocaciones sofisticadas (Gráfico 12), pero la diferencia significativa (p=0,006) la encontramos en relación con la sofisticación baja entre Biología y Ciencias de la Salud, y Artes y Humanidades (p=0,010). Como se muestra en el Gráfico 13, Artes y Humanidades es el dominio con mayor número de colocaciones de sofisticación baja:

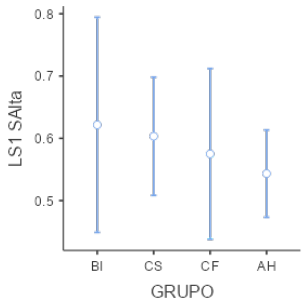

Gráfico 12. LS1 SAlta en noveles por dominio.

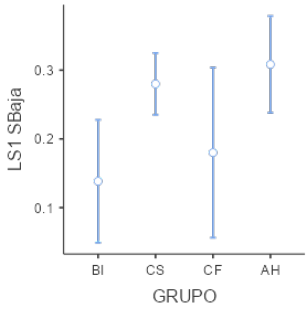

Gráfico 13. LS1 SBaja en noveles por dominio.

En este sentido, podríamos destacar el uso frecuente de colocaciones de sofisticación baja en el dominio de Artes y Humanidades en noveles, donde se aprecia una diferencia significativa con Biología y Ciencias de la Salud, pero no encontramos diferencias estadísticamente significativas para determinar qué grupo utiliza colocaciones más sofisticadas. A este respecto, es interesante resaltar que el dominio de Artes y Humanidades presenta la mayor proporción de colocaciones de sofisticación baja en noveles y en expertos. Este resultado puede ser debido a que, en los textos de este dominio, se utiliza una lengua académica más inteligible y menos técnica, es decir, más asociada a la lengua común. De ahí que los hablantes no expertos en las materias puedan leer y comprender con más facilidad un texto de este dominio en comparación a un texto, por ejemplo, de Ingeniería, que presenta más vocabulario terminológico y específico de los procesos y actividades que se realizan en dicha disciplina.

En definitiva, podemos constatar que los noveles ya conocen las colocaciones específicas del discurso académico, pero siguen utilizando en sus textos las colocaciones de la lengua general. La solución para que un estudiante escriba como un experto no es la de incorporar colocaciones más sofisticadas, ya que en el proceso de escritura de su trabajo final de carrera o máster ya las tiene adquiridas, sino la de desterrar de su repertorio las colocaciones poco sofisticadas o más comunes de la lengua general. En este sentido, la sofisticación y la diversidad, calculadas siguiendo la

metodología presentada en este estudio, permiten identificar algunas de las carencias léxicas de los noveles en el ámbito académico. Considerando lo mencionado anteriormente, debemos advertir que las medidas estadísticas, en determinados contextos, pueden resultar todavía insuficientes para evaluar la calidad y el uso del vocabulario de un estudiante debido a la falta de validez como modelo teórico (Jarvis, 2013). Dicho esto, para refinar el proceso automático de clasificación de colocaciones en franjas de sofisticación, se podría incorporar una validación humana con el fin de determinar si, además de la frecuencia, otros factores pueden afectar a la clasificación de una colocación. Esto es especialmente relevante si se pretende tomar el estudio del perfil colocacional como material pedagógico o para la evaluación de textos académicos, y podría ir en línea con lo propuesto por Simpson-Vlach y Ellis (2010) quienes combinan las medidas estadísticas con el FTW (*Formulae Teaching Worth*), que es una predicción de cómo los instructores evalúan la relevancia y validez de cada unidad.

CONCLUSIONES

En este estudio, hemos presentado una metodología para identificar la sofisticación y la diversidad colocacional de textos del discurso académico contrastando textos de expertos y de noveles. Para calcular la sofisticación, hemos aplicados dos métodos, uno basado en cálculos de frecuencia y otro en índices de IM. Siguiendo el primero, una colocación se clasifica en una franja de sofisticación más alta cuanto menos frecuente sea en la lengua general y más propia sea del discurso académico y, en el segundo caso, cuanto más valor de IM posea. Los datos obtenidos a partir del primer método, seleccionado como el método más representativo, indican que la diferencia entre los textos de expertos y de noveles no estriba en el uso de colocaciones más sofisticadas por parte del primer grupo, sino por el mayor uso de colocaciones poco sofisticadas por parte de los noveles. Con respecto a las colocaciones de sofisticación baja, es importante destacar el mayor uso de colocaciones de esta franja en los textos de Artes y Humanidades, que puede estar asociado a que en este dominio se da un menor uso al vocabulario especializado y un mayor uso al vocabulario de la lengua general. Por último, se han observado diferencias relacionadas con la diversidad colocacional, dado que son los expertos quienes repiten menos colocaciones del discurso académico en un mismo texto.

Este estudio ha permitido obtener una clasificación de las colocaciones académicas en español según su sofisticación y ha ofrecido datos relevantes en cuanto al uso de las colocaciones académicas en dos grupos con un distinto nivel de competencia léxica. Los parámetros aplicados y la clasificación de colocaciones podrían ser los primeros elementos a integrar en una futura herramienta en línea que automatice el cálculo de la complejidad colocacional, similar a otras como el *Lexical Frequency Profile,* y que pueda resultar especialmente útil en el ámbito de certificaciones de español con fines académicos.

REFERENCIAS BIBLIOGRÁFICAS

Ackermann, K. & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL): A Corpus-Driven and Expert-Judged Approach. *Journal of English for Academic Purposes*, *12*(4), 235-247.

Ai, H. & Lu, X. (2010). *A Web-Based System for Automatic Measurement of Lexical Complexity*. Ponencia presentada en el 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA.

Ahumada, I., Zamorano, J. P., García, E. D. R. & Lara, I. A. (2011). Design and Development of Iberia: A Corpus of Scientific Spanish. *Corpora*, *6*(2), 145-158.

Almela, R., Cantos, P., Sánchez, A., Sarmiento, R. & Almela, M. (2005). *Frecuencias del español. Diccionario y Estudios Léxicos y Morfológicos*. Madrid: Universitas, S.A.

Alonso-Ramos, M. (2010). No importa si la llamas o no colocación, descríbela. En C. Mellado Blanco, P. Buján Otero, C. Herrero Kaczmarek, N. M. Iglesias Iglesias & A. Mansilla Pérez (Eds.), *La fraseología del S. XXI: Nuevas perspectivas de la fraseología del S.XXI* (pp. 55-80). Berlín: Frank & Timme.

Alonso-Ramos, M., García-Salido, M. & Garcia, M. (2017). Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations. En I. Kosem, J. Kallas, C. Tiberius, S. Krek, M. Jakubíček & V. Baisa (Eds.), *Proceedings of eLex 2017 conference* (pp. 571-586). Leiden: The Netherlands.

Alonso-Ramos, M., García-Salido, M. G., Garcia, M. & Guzzi, E. (2019). Identification of Cross-Disciplinary Spanish Academic Collocations for a Lexical Tool. En I. Kosem & T. Zingano Kuhn (Eds.), *Book of abstracts of eLex 2019 conference: Electronic lexicography in the 21st century: Smart lexicography* (pp. 51-52). Sintra: Portugal.

Ávila, R. (1988). Lengua hablada y estrato social: Un acercamiento lexicoestadístico. *Nueva Revista de Filología Hispánica*, *36*(1), 131-148.

Bardel, C., Gudmundson, A. & Lindqvist, C. (2012). Aspects of Lexical Sophistication in Advanced Learners' Oral Production: Vocabulary Acquisition and Use in L2 French and Italian. *Studies in Second Language Acquisition*, *34*(2), 269-290.

Bulté, B. & Housen, A. (2012). Defining and Operationalising L2 Complexity. En A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA* (pp. 21-46). Ámsterdam: John Benjamins.

Capsada Blanch, R. & Torruella Casañas, J. (2017). Métodos para medir la riqueza léxica de los textos. Revisión y propuesta. *Verba. Anuario Galego de Filoloxía,* 44*,* 347-408.

Covington, M. A. & McFall, J. D. (2010). Cutting the Gordian Knot: The Moving-Average Type–Token Ratio (MATTR). *Journal of Quantitative Linguistics*, *17*(2), 94-100.

Crossley, S., Salsbury, T. & McNamara, D. (2012). Predicting the Proficiency Level of Language Learners Using Lexical Indices. *Language Testing*, *29*(2), 243-263.

Durrant, P. & Schmitt, N. (2009). To What Extent do Native and Non-Native Writers Make Use of Collocations? *International Review of Applied Linguistics in Language Teaching*, *47*(2), 157-177.

García-Salido, M. (2021). *Compiling an Academic Vocabulary List of Spanish*. DOI: https://doi.org/10.13140/RG.2.2.27681.33123.

Granger, S. & Bestgen, Y. (2014). The Use of Collocations by Intermediate vs. Advanced Non-Native Writers: A Bigram-Based Study. *International Review of Applied Linguistics in Language Teaching*, *52*(3), 229-252.

Guiraud, P. (1954). *Les Caractères Statistiques du Vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Hyland, K. & Tse, P. (2007). Is There an Academic Vocabulary? *TESOL Quarterly*, *41*(2), 235-253.

Jarvis, S. (2013). Capturing Diversity in Lexical Diversity. *Language Learning,* 63, 87-106.

Johnson, W. (1944). Studies in Language Behavior: A Program of Research. *Psychological Monographs*, *56*(2), 1-15.

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

Kyle, K. & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. *Tesol Quarterly*, *49*(4), 757-786.

Laufer, B. & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics, 16*(3), 307-322.

López-Morales, H. (2010). Los índices de riqueza léxica y la enseñanza de lenguas. En F. J. de Santiago Guervós, H. Bongaerts, J. J. Sánchez Iglesias & M. Seseña Gómez (Eds.), *Del texto a la lengua: La aplicación de los textos a la enseñanza-aprendizaje del español L2-LE* (pp. 15-28). Salamanca: Asociación para la Enseñanza del Español como Lengua Extranjera.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, *96*(2), 190-208.

L'Homme, M. C. (2000). Understanding Specialized Lexical Combinations. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *6*(1), 89-109.

Malvern, D., Richards, B., Chipere, N. & Durán, P. (2004). *Lexical Diversity and Language Development. Quantification and Assessment*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

McCarthy, P. M. & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods,* 42, 381-392.

Mel'čuk, I. (2012). Phraseology in the Language, in the Dictionary, and in the Computer. *Yearbook of phraseology*, *3*(1), 31-56.

Mendoza, A. (2015). La selección de las tareas de escritura en los exámenes de lengua extranjera destinados al ámbito académico. *Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas*, 18*,* 106-123.

Montolío, E. (Dir.). (2014). *Manual de escritura académica y profesional* (2 vols.). Barcelona: Ariel.

Orol González, A. & Alonso-Ramos, M. (2013). A Comparative Study of Collocations in a Native Corpus and a Learner Corpus of Spanish. *Procedia-Social and Behavioral Sciences*, 95, 563-570.

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights from a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly, 15*(1), 29-43.

Paquot, M. (2019). The Phraseological Dimension in Interlanguage Complexity Research. *Second Language Research*, *35*(1), 121-145.

Paquot, M., Gries, S. T. & Yoder, M. (2020). Measuring Lexicogrammar. En P. Brinke & T. Brunfatus (Eds.), *The Routledge Handbook of Second Language Acquisition and Language Testing* (pp. 223-232). Nueva York: Routledge.

Parodi, G. (Ed.) (2010). *Academic and Professional Discourse Genres in Spanish*. Ámsterdam/Filadelfia: John Benjamins.

Pérez-Llantada, C. (2014). Formulaic Language in L1 and L2 Expert Academic Writing: Convergent and Divergent Usage. *Journal of English for Academic Purposes,* 14, 84-94.

Read, J. (2000). *Assessing Vocabulary*. Cambridge: CUP.

Simpson-Vlach, R. & Ellis, N. C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, *31*(4), 487-512.

Tidball, F. & Treffers-Daller, J. (2008). Analysing Lexical Richness in French Learner Language: What Frequency Lists and Teacher Judgements Can Tell us About Basic and Advanced Words1. *Journal of French language studies*, *18*(3), 299-313.

Tweedie, F. J. & Baayen, R. H. (1998). How Variable May a Constant Be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, *32*(5), 323-352.

Vázquez, G. (Coord.). (2000). *Guía didáctica del discurso académico escrito ¿Cómo se escribe una monografía?* Madrid: Edinumen.

Vázquez, G. (Coord.). (2005). *Español con fines académicos: De la comprensión a la producción de textos*. Madrid: Edinumen.

Villayandre Llamazares, M. (2019). HARTA de noveles: Un corpus de español académico. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos, 5*(1), 131-145.

Yu, G. (2010). Lexical Diversity in Writing and Speaking Task Performances. *Applied linguistics*, *31*(2), 236-259.

AGRADECIMIENTOS

Este estudio ha sido posible gracias a la financiación del Ministerio de Ciencia e Innovación (PID2019-109683GB-C21); al soporte de la Xunta de Galicia a través de la ayuda ED431C 2020/11; del Centro de Investigación de Galicia do Sistema Universitario de Galicia "CITIC", financiado por la Xunta de Galicia y la Unión Europea (FEDER GALICIA 2014-2020), con la ayuda ED431G 2019/01 (80% FEDER, 20% Consellería de Educación, Universidade e Formación Profesional"); y del Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020.

NOTAS

[1] Para más información véase: https://www.cepe.unam.mx/certificaciones/exeleaa

[2] The jamovi project (2021). *jamovi* (Version 1.6) [Computer Software]. Recuperado de: https://www.jamovi.org