Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 83-94 recibido 31-03-2022 revisado 20-05-2022 aceptado 23-05-2022

Selección de colocaciones académicas en español a través de un filtro de interdisciplinariedad

Selecting Spanish academic collocations using a filter of interdisciplinarity

Eleonora Guzzi, Margarita Alonso Ramos
Universidade da Coruña, CITIC
eleonora.guzzi@udc.es, margarita.alonso@udc.gal

Resumen: En este artículo se propone una metodología para compilar una lista de colocaciones académicas con base nominal que se integran en una herramienta léxica (Alonso-Ramos, García-Salido y Garcia, 2017). Para ello, establecemos un filtro que mide la interdisciplinariedad de los nombres académicos a partir de los cuales se extraen las colocaciones (García-Salido, 2021), con el fin de mantener los nombres frecuentes y bien distribuidos en distintas disciplinas académicas, y descartar aquellos que se adscriben a la terminología o que son más característicos de la lengua general. Utilizamos tres criterios: (1) el IDF (Jones, 1972); (2) el análisis de la distribución de colocaciones; (3) el contraste con listas de vocabulario académico inglés. Los resultados muestran que estos criterios son útiles para identificar los nombres prototípicos del discurso académico y permiten filtrar la lista de colocaciones académicas. No obstante, persiste el problema de cómo tratar la desambiguación semántica en relación con las diferentes disciplinas.

Palabras clave: discurso académico, interdisciplinariedad, colocaciones académicas.

Abstract: In this paper a methodology to compile a list of noun-based academic collocations that feed a lexical tool (Author, 2017) is proposed. To do so, a filter that measures the interdisciplinarity of academic nouns from which collocations are extracted (García-Salido, 2021) is established. This filter is applied to include nouns that are frequent and homogeneously distributed across different academic disciplines, and discard those ascribed to terminology or are more prototypical of general language. Three criteria were used: (1) the IDF (Jones, 1972); (2) an analysis of collocation distributions; (3) a contrast with vocabulary lists of academic English. Results show that these criteria are useful for identifying prototypical nouns of academic discourse and allow for filtering the list of academic collocations. However, the problem regarding how to deal with semantic disambiguation in different disciplines is still present.

Keywords: academic discourse, interdisciplinarity, academic collocations.

1 Introducción

Uno de los principales objetivos dentro del ámbito de las lenguas con fines académicos ha sido el de proporcionar listas de vocabulario académico para ser utilizadas como recursos pedagógicos e integradas en la enseñanza de la escritura académica. Este vocabulario incluye unidades y combinaciones léxicas que son específicas del género académico, pero no son terminológicas. A su vez, se caracterizan por ser más frecuentes en el discurso académico que en la lengua general o en otros géneros. Siguiendo a Tutin (2007a) podemos definir el vocabulario académico como aquel vocabulario que hace referencia a los procedimientos y actividades científicas del discurso científico, esenciales en la argumentación y en la estructuración de los textos académicos (Drouin, 2007; Paquot y Bestgen, 2009).

Hasta el momento, se han propuesto varias listas de unidades léxicas académicas especialmente para el inglés y el francés: *Academic Vocabulary Lis*t (AVL, Gardner y Davies, 2013), *Academic Word List* (AWL, Coxhead, 2000), *Academic Keyword List* (AKL, Paquot, 2007, *French Cross-disciplinary Scientific Lexicon* (Hatier et al., 2016), *Lexique*

ISSN 1135-5948. DOI 10.26342/2022-69-7

Scientifique Transdisciplinaire (LST, Drouin, 200)7, entre otras. Por otro lado, dentro de las listas de combinaciones léxicas destacan la *Academic Collocation List* (ACL, Ackermann y Chen, 2013) y la *Academic English Collocation List* (Lei y Liu, 2018), centradas específicamente en las colocaciones, así como la lista de palabras y colocaciones académicas empleadas para la herramienta lexicográfica *Collocaid* (Frankenberg-Garcia et al., 2019). En el ámbito del español, se ha propuesto recientemente una lista de unidades léxicas académicas, la *Spanish Academic Key Word List* (SpAKWL, García-Salido 2021), que incluye 1.239 lemas de nombres, adjetivos, verbos y adverbios. A partir de los nombres de esta lista, también se ha extraído una primera versión de colocaciones académicas con base nominal que se integran en la *Herramienta de Ayuda a la Redacción de Textos Académicos* (HARTA; Alonso-Ramos, García-Salido y Garcia, 2017; disponible en: http://www.dicesp.com:8083/).

En estos estudios se han aplicado varios criterios estadísticos para identificar automáticamente el vocabulario interdisciplinar en el discurso académico. Sin embargo, los criterios pueden diferir entre las listas de diferentes lenguas por los rasgos léxicos de las mismas. Por ejemplo, como apuntan Cobb y Horst (2004), existe un continuum entre el vocabulario académico y no académico en lenguas como el francés o el español debido a que, entre otras razones, el vocabulario greco-latino que es característico de los textos académicos, también se integra en dominios no académicos. Sin embargo, esto no sucede en inglés, donde los términos greco-latinos tienen mucha más presencia en el discurso académico. En el caso del español, además, aunque una palabra se utilice en la lengua general, se podría incluir en las listas académicas si también es frecuente en el discurso académico, debido a que, por un lado, los sentidos de las palabras académicas pueden diferir de los utilizados en los textos que no son académicos (Gilquin, Granger y Paquot, 2007), como, por ejemplo, *trabajo* (‘composición científica o literaria’ frente a *trabajo* como ‘empleo’, en la lengua general) y, por otro lado, porque dichas palabras pueden formar parte de combinaciones léxicas que son más exclusivas de los textos académicos (García-Salido, 2021), como la colocación *dato cuantitativo* con la palabra *dato,* frente a *dato personal.* El objetivo final a la hora de compilar las listas de vocabulario académico debería ser crear un balance entre la especificidad del vocabulario y su productividad. Esto es, se debería apuntar a la inclusión de unidades y combinaciones léxicas que son productivas para la redacción de textos académicos y que pueden estar en la intersección, por ejemplo, entre la lengua general y la lengua académica, pero que excluyen lo que es propio de otros géneros.

Otras dificultades a la hora de recoger este vocabulario argumentadas en Hyland y Tse (2007) son la falta de atención a la posible polisemia y a la variedad disciplinar. La polisemia implica que distintos sentidos sean utilizados en distintas disciplinas académicas: por ejemplo, la palabra *volumen* puede significar ‘capacidad’ en Física, ‘ejemplar de libro’ en Biblioteconomía o ‘frecuencia acústica’ en Ingeniería. Por otra parte, la variedad disciplinar se asocia a la posibilidad de que determinadas unidades y combinaciones léxicas pueden ser académicas, pero pueden utilizarse con más frecuencia en algunas disciplinas.

En definitiva, no existe un consenso común que apunte hacia una generalización de criterios sobre cómo determinar la interdisciplinariedad y obtener vocabulario característico de textos académicos. Como consecuencia, este trabajo se propone con un doble objetivo: por un lado, ofrecer una lista más refinada de nombres académicos en español y, por otro lado, utilizar la lista para filtrar las colocaciones académicas con base nominal que se integran en HARTA. Para alcanzar estos objetivos, seguimos una metodología basada en el refinamiento de los nombres de la *SpAKWL*, que implica el descarte de aquellos adscritos a una disciplina específica o que son significativamente más recurrentes en la lengua general.

A continuación, en la sección 2 presentamos el proceso de extracción de la *SpAKWL* y de las colocaciones académicas en español; en la

sección 3 presentamos la metodología empleada para identificar los nombres de la *SpAKWL* que son interdisciplinares y filtrar las colocaciones que contienen dichos nombres; en la sección 4 exponemos los resultados obtenidos a partir de los distintos análisis; y, en la sección 5, discutimos los resultados, seguidos de las conclusiones finales.

2 Descripción de los datos: SpAKWL y colocaciones académicas

Este estudio parte de la lista de palabras académicas del español *SpAKWL* (García-Salido, 2021), extraída siguiendo criterios de especificidad y de distribución a partir de un corpus académico, HARTA-Expertos (HE, Alonso-Ramos, García-Salido y Garcia, 2017). HE contiene 413 artículos científicos publicados, cuya mayoría proviene de la sección en español del corpus *Spanish-English Research Articles Corpus* (SERAC; Pérez-Llantada, 2014), y suma un total de 2.025.092 palabras. Está divido en 4 dominios principales (Artes y Humanidades, Biología y Medicina, Ciencias Sociales y Ciencias Físicas e Ingeniería) y 12 subdominios, siguiendo la estructura del SERAC. El proceso de tokenización y lematización se llevó a cabo mediante LinguaKit (Garcia y Gamallo, 2016) y el de etiquetación con FreeLing (Padró y Stanilovsky, 2012). Para analizar sintácticamente el corpus con dependencias universales (Nivre et al., 2016) se utilizó UDPipe (Straka, Hajic y Straková, 2016).

Para determinar la especificidad e identificar las palabras específicas del ámbito académico frente a un corpus de referencia, se empleó el test estadístico log-likelihood a partir de las frecuencias absolutas de HE y del corpus de referencia, en este caso, la sección de ficción del corpus de narrativa LEXESP (Sebastián-Gallés et al., 2000), de 5 millones de palabras. Como criterio de distribución, se seleccionaron aquellas palabras con ocurrencias en los 4 dominios y el 10% de palabras con una distribución más homogénea en términos de DP (Deviation of Proportions, Gries, 2008). La lista de palabras académicas resultante cuenta con 1.239 lemas que se corresponden con nombres, verbos, adverbios y adjetivos.

A partir de esta lista, se seleccionaron los nombres (n=602) que se emplearon como bases para extraer automáticamente una primera versión de colocaciones académicas, que están integradas en HARTA. Para este propósito, definimos las colocaciones, dentro del marco de la Lexicografía Explicativa y Combinatoria (Mel'čuk, 2012), como combinaciones léxicas con un significado composicional, que están formadas por una 'base', en este caso, un nombre, y un 'colocativo', y cuyos elementos tienden a coocurrir, como, por ejemplo, *alcanzar un objetivo*.

Para analizar la interdisciplinariedad de las bases y de las colocaciones académicas en español, partimos de las colocaciones ya integradas en HARTA y de un segundo grupo más numeroso de colocaciones, que fue extraído a partir de una ampliación del corpus HE, HARTA-Expertos-Plus (HEP). Este corpus contiene 21.068.482 palabras procedentes de 3.870 artículos de investigación: 19.043.390 palabras proceden del corpus académico-científico *Iberia* (Ahumada et al., 2011) y 2.025.092 palabras provienen del corpus HE. El corpus HEP se divide en los mismos cuatro dominios principales que HE, a su vez divididos en subdominios. Para los nuevos artículos, se aplicó el mismo proceso de tokenización, lematización y análisis sintáctico que se siguió para HE. Tras replicar los pasos seguidos para obtener la primera versión de colocaciones académicas, en primer lugar, se extrajeron colocaciones de 5 relaciones de dependencias sintácticas (N + N, V + Obj, Suj + V, N + Adj, N + Obl). En segundo lugar, se realizó una extracción automática, basada en medidas de asociación estadísticas (log-likelihood, Información Mutua, entre otras), y en criterios de frecuencia (≥5 ocurrencias) (Alonso-Ramos, García-Salido y Garcia, 2017). En tercer lugar, un grupo de anotadores refinó manualmente los candidatos extraídos automáticamente para obtener las colocaciones académicas, siguiendo criterios fraseológicos que se enmarcan dentro de la Teoría Sentido-Texto (Mel'čuk, 2012). A

pesar del refinamiento manual y de las medidas escogidas, tanto en la lista de nombres académicos como en las colocaciones seleccionadas se incluyen casos más asociados a la terminología, como, por ejemplo, la palabra *tejido* usada con el sentido de Biología 'cada uno de los agregados de células de la misma naturaleza' (DLE, s.m, def. 4), o la colocación *ingresar paciente,* que no son productivas para los varios dominios académicos.

3 Metodología

Con el fin de descartar los nombres especializados y los que son más frecuentes en la lengua general o en otros géneros, en primer lugar, aplicamos la medida IDF (Inverse Document Frequency, Jones 1972) a los 602 nombres de la lista *SpAKWL*. Esta medida se basa en el cálculo de la proporción de documentos que contiene un determinado término y se utiliza frecuentemente en el campo de la Extracción de la Información para identificar palabras clave, esto es, palabras específicas de un conjunto de textos que tienen un alto valor de IDF. Tras calcular el IDF de los nombres, ordenamos los valores de mayor a menor para visualizar en la posición más alta los candidatos más terminológicos. Calculamos la media de IDF (=0,659) y llevamos a cabo una revisión exhaustiva de los nombres situados por encima y por debajo de la misma para seleccionar el punto de corte, en este caso, la palabra *potencia* (IDF=0,900). En este estudio, el IDF nos ayudó a determinar precisamente las palabras que pueden ser descartadas (n=119) por ser menos interdisciplinares y más características de algunos artículos, que se corresponden con las ubicadas por encima del punto de corte (ver Anexo 1).

A partir del resultado obtenido, seguimos un proceso de filtrado dividido en diferentes fases para analizar con más profundidad los 119 nombres que se descartarían por el IDF. En la primera fase (F1), un grupo de anotadores clasificó los nombres con IDF ≥0,900 en dos grupos: los nombres que, tal y como indica esta medida, se descartan por ser especializados o porque pertenecen a la lengua general (G1) y los nombres que deben seguir un proceso de revisión posterior (G2). En la segunda fase (F2), en la que se incluyen los nombres clasificados en el G2, identificamos las colocaciones extraídas que contienen dichos nombres como bases para analizar su distribución y el número de colocativos con los que se combina. En este proceso, algunas bases se descartan, otras bases se reincluyen en la lista inicial y otro grupo de bases se selecciona para revisar en la siguiente fase. Por último, en la tercera fase (F3), contrastamos los equivalentes de los nombres en cuatro listas de inglés académico (*AVL, AWL, AKL*, y la lista de palabras académicas de *Collocaid*). En la Figura 1, exponemos el proceso seguido:

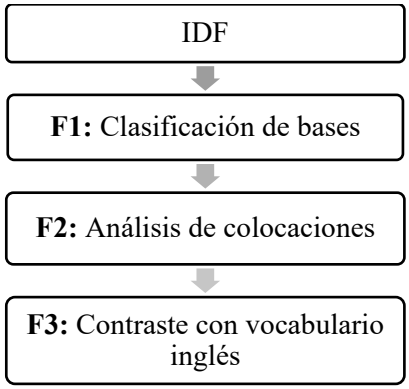

Figura 1. Proceso para filtrar los nombres de la *SpAKWL*

En las siguientes secciones, se explica en detalle el proceso que llevamos a cabo en las tres fases.

3.1. Clasificación de las bases a partir del resultado del IDF

Una vez identificados los 119 nombres que están por encima del punto de corte del IDF, un grupo de anotadores conformado por tres lingüistas los clasificó en dos grupos. En el G1 se incluyeron los nombres que, tras observar los ejemplos en contexto en el corpus HE y con la ayuda de los diccionarios para analizar los sentidos y las marcas de especialidad, resultaron ser terminológicos o más asociados a un número reducido de disciplinas. También se incluyeron aquellos nombres que presentan una frecuencia elevada en los corpus de lengua general, utilizando herramientas de corpus para consultar su frecuencia, como el corpus de *esTenTen* en

Sketch Engine (Kilgariff y Renau, 2013). En el G2 se incluyeron los nombres que, en cambio, requieren un análisis posterior por no mostrar indicios claros sobre su interdisciplinariedad. En función del descarte o mantenimiento, se les asignaron puntuaciones a los nombres. Por ejemplo, encontramos casos como el de la palabra *fracción,* que se descarta (=0), e *indicación,* que pasa a una siguiente fase (=análisis) (Tabla 1):

	Fase 1
fracción	0
indicación	análisis

Tabla 1. Ejemplo de puntuación asignada a nombres a descartar o analizar en la F1.

3.2.Análisis de colocaciones

En esta fase analizamos las bases clasificadas en el G2 (n=54), con el fin de llevar a cabo una valoración acerca del número de colocativos, que se asocia a la riqueza del vocabulario, así como de la distribución de las colocaciones que conforman, relacionada con la interdisciplinariedad. En este sentido, una base que se combina con varios colocativos y que forma colocaciones que están bien distribuidas en los textos académicos debería incluirse en la lista de nombres académicos.

Para analizar la distribución de las colocaciones, se representaron las frecuencias de las colocaciones en cada dominio (AH, CS, CF, BM) en forma de porcentajes, se calculó la desviación estándar (DE), y se indicó el número de subdominios en los que aparece cada colocación. Los valores de la DE oscilaron entre 0,00 y 0,50: cuanto más bajo es el valor, más homogénea es la distribución de la colocación en los textos de los cuatros dominios. En cuanto a los criterios de análisis de las colocaciones que contienen los nombres académicos, consideramos los tres parámetros: una colocación es interdisciplinar si presenta una DE entre 0,00 y 0,24, si se aproxima a un porcentaje de ≥20% en al menos tres dominios o bien en dos dominios, uno perteneciente a "ciencias duras" (CF y BM) y otro a "ciencias blandas" (CS y AH) y si aparece en ≥3 subdominios. Por lo tanto, si las colocaciones analizadas con un nombre del G2 están bien distribuidas, como, por ejemplo, la colocación *alcanzar difusión* (Tabla 2), el nombre, en este caso *difusión,* recibe una puntuación *=1* y se mantiene:

	DE	**Dom.**	**Sub.**	**=**
alcanzar difusión	0,27	AH 50%	5	Distri. alta
		CS 12,5%		
		CF 12,5%		
		BM 25%		

Tabla 2. Análisis de una colocación con distribución alta (homogénea).

Por el contrario, si las colocaciones que se forman a partir de un nombre presentan una DE entre 0,36-0,50, aproximadamente, una frecuencia de 0% en tres dominios o ≥90% en un dominio, y aparece solamente en uno o dos subdominios (Tabla 3), la base, en este caso, *abundancia*, con una puntuación *=0*, se descarta:

	DE	**Dom.**	**Sub.**	**=**
abundancia mayor	0,47	BM 97%	2	Distri. baja
		CF 3%		

Tabla 3. Análisis de una colocación con distribución baja (heterogénea).

Los nombres que pasan a la siguiente fase de revisión (F3) en la que se contrastan con las posibles equivalencias en las listas de inglés, y que reciben una puntuación *=análisis*, son aquellas bases que en este proceso presentan colocaciones que oscilan entre los límites de una distribución homogénea y heterogénea. Esto es, en este grupo se incluyen las bases cuyas colocaciones presentan una DE entre 0,25 y 0,35, aproximadamente, aparecen en dos dominios de un único grupo ("ciencias "duras" / "ciencias blandas"), pero con porcentajes equilibrados, o en tres subdominios de forma desequilibrada (Tabla 4):

	DE	**Dom.**	**Sub.**	**=**
indicación precisa	0,33	CS 18%	3	Distri. media
		AH 9%		
		BM 73%		

Tabla 4. Análisis de una colocación que requiere un análisis posterior.

Asimismo, pasan a la fase de revisión 3 un número reducido de bases que no presentan ningún colocativo productivo a nivel fraseológico de entre todos los candidatos extraídos, como, por ejemplo, el nombre *tipología*. Ahí decidimos su mantenimiento o descarte para formar parte de una nueva lista de nombres académicos.

En la siguiente Tabla (5) mostramos las puntuaciones de tres bases que, tras el análisis colocacional, reciben puntuaciones distintas:

	Fase 1	**Fase 2**
abundancia	1	0
indicación	1	análisis
difusión	1	1

Tabla 5. Ejemplos de puntuaciones asignadas a bases a descartar, analizar o incluir en la F2.

3.3. Contraste con listas de vocabulario de inglés académico

En esta última fase se analizaron las bases que presentaron dudas tras pasar por el análisis colocacional y aquellos nombres que no han podido analizarse en la fase 2, debido a que no presentaron ningún colocativo.

Se seleccionaron cuatro listas de palabras académicas en inglés para el análisis: *AKL, AVL, AWL* y la lista de palabras académicas de *Collocaid*. A partir de los nombres seleccionados en la fase anterior, se realizó una comparativa con las palabras académicas pertenecientes a las cuatro listas para encontrar su equivalente en inglés. En el proceso de búsqueda de los equivalentes, se consultaron dos diccionarios, el *Oxford English Dictionary* y el *Cambridge Dictionary*, tanto la versión bilingüe español-inglés como la monolingüe, así como el corpus paralelo *Linguee* para observar ejemplos en contexto. A la hora de buscar la equivalencia de cada nombre, se consideraron las diferentes traducciones posibles debido a la presencia de polisemia. Se considera que una palabra coincide con dos o más listas si en cada lista se presenta la traducción asociada a un mismo sentido, por ejemplo, la palabra *cultura* presenta su correspondencia en las cuatro listas con el nombre en inglés *culture.* Sin embargo, se dan otros casos en los que una palabra se traduce de distintas formas dependiendo del sentido que se adopte en cada una de ellas: por ejemplo, la palabra *señal* se traduce en la *AVL* con el sentido asociado a *signal,* mientras que en la lista de *Collocaid* y en la *AKL* solamente se encuentra el equivalente de otro sentido, *indication.* En estos casos, se considera el número de listas coincidentes para cada sentido: *señal* puede coincidir con una lista (la *AVL*) o dos listas (*Collocaid* y *AKL*), en función del sentido.

En relación con el criterio para filtrar la lista de nombres, se estableció un índice de ≥2, es decir, si un nombre aparece en al menos dos de las cuatro listas se mantiene. Por el contrario, las palabras que coinciden únicamente con una lista o que no coinciden con ninguna finalmente se descartan. Fijamos este índice debido a que la *AWL* aplica criterios más restringidos y no incluye palabras que también pertenecen a la lengua general, lo que provoca un porcentaje de correspondencia bajo porque en el vocabulario académico español se recogen palabras compartidas con la lengua general.

En la Tabla 6, podemos observar los ejemplos con las respectivas puntuaciones de nombres que se analizan en la fase 3, que incluyen tanto nombres que presentaron colocativos en la F2 (ej. *indicación* o *especificación*), como aquellos que no presentaron ningún colocativo (ej. *tipología* o *almacenamiento*):

	Fase1	**Fase2**	**Fase3**	**=**
almacenamiento	1	análisis (no coloc.)	0	1
tipología	1	análisis (no coloc.)	1	2
especificación	1	análisis	0	1
indicación	1	análisis	1	2

Tabla 6. Ejemplos de puntuaciones asignadas a nombres a descartar o incluir en la F3.

4 Resultados

La lista resultante se compone de 519 nombres académicos (Anexo 2), en contraste con los 602 nombres iniciales. La clasificación de los nombres en cada fase ha sido el resultado de las puntuaciones asignadas a cada uno de ellos. Una

puntuación final de 2 implicó la reinclusión del nombre a la lista inicial y una puntuación de 0 o 1 conllevó su descarte.

En la primera fase (F1), se descartaron 65 nombres, con una puntuación de 0, entre los cuales encontramos ejemplos como *emisión, fracción, tejido, geometría, prevención, infraestructura,* etc. Por otra parte, se mantuvieron 54 nombres que pasaron a una siguiente revisión, con una puntuación de 1, como *énfasis, concordancia, fiabilidad, puntuación, descenso, almacenamiento, procesamiento,* entre otros.

En cuanto a la fase de revisión de las colocaciones (F2), 6 bases se descartaron, como *especificación, asignación* o *resto*; 29 bases pasaron a la fase 3, como *trayectoria, gráfico, indicación* o *premisa*; y 19 bases se reincluyeron en la lista inicial por cumplir con el criterio de número y distribución de las colocaciones, como *sesgo, productividad, predominio* o *estándar*.

Por último, en la fase 3, de los nombres que se contrastaron con las listas de palabras académicas en inglés, se descartaron l2 nombres, entre los cuales encontramos ejemplos como *corrección, concordancia* y *fiabilidad*, ya que sus posibles equivalentes en inglés no aparecían en ninguna lista de inglés y *almacenamiento*, *formulación*, *asignación, acumulación*, *regulación*, *procesamiento, trayectoria, afirmación* y *barrera* porque aparecían únicamente en una lista. Por ejemplo, de la palabra *regulación* únicamente encontramos un posible equivalente en la *AVL* como *adjustment.*

Sin embargo, se reincluyeron 17 nombres: *gráfico, indicación, reproducción* y *experto,* que aparecían en dos listas; *heterogeneidad, premisa, variante, diferenciación, bibliografía, tipología, desempeño, instancia*, *supuesto* y *unión,* que aparecían en tres listas, y *paradigma, vínculo* y *rol,* cuyos equivalentes aparecieron en las cuatro listas. Por ejemplo, la palabra *instancia* aparece como *instance* en la *AVL*, en la *AKL* y en la lista de *Collocaid*.

A modo de resumen, en la Tabla 7, mostramos el número de nombres descartados, analizados o reincluidos en cada fase:

	F1	F2	F3	TOTAL
Descarte	65	6	12	83
Análisis	**54**	**29**	-	83
Reinclusión	-	19	17	36
TOTAL		**54**	**29**	119

Tabla 7. Nº de nombres clasificados en cada fase.

5 Discusión

La medida IDF ha permitido detectar las palabras clave de determinados documentos del corpus y, en consecuencia, aquellos nombres empleados más específicamente en algunas de las áreas científicas en las que está dividido el corpus HE. Cabe destacar que la decisión del punto de corte en la palabra *potencia* (IDF =0,900) ha implicado que un número reducido de nombres, como *software* o *regresión*, no se descartaran a pesar de presentar un valor alto de IDF (cerca de 0,800) y de ser especializados. Sin embargo, hemos optado por no establecer un punto de corte más bajo debido a que se habría descartado una gran parte de nombres académicos relevantes, como *disciplina, discurso, síntesis,* entre otros.

De los dos conjuntos de nombres obtenidos con el IDF, contrastamos la dispersión de algunas palabras con un alto valor de IDF con algunas palabras con valores más bajos en los artículos de cada subdominio. En efecto, observamos que la frecuencia y distribución no es proporcionada en el corpus en el primer caso (IDF alto), pero sí en el segundo (IDF bajo). En la Figura 1, podemos observar este comportamiento en una muestra de seis nombres con valores altos de IDF (IDF>1,06), es decir, menos distribuidos en los textos, y en la Figura 2, seis nombres con valores muy bajos (IDF<0.05). Para facilitar su lectura, presentamos la distribución de los nombres por subdominio en lugar de por artículos:

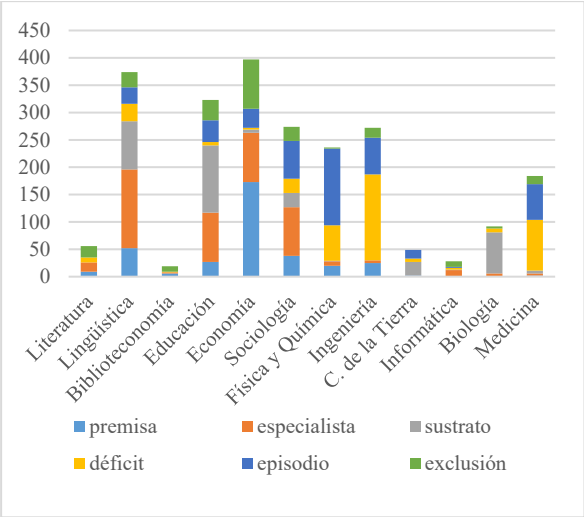

Figura 2. Seis nombres con IDF >1,05.

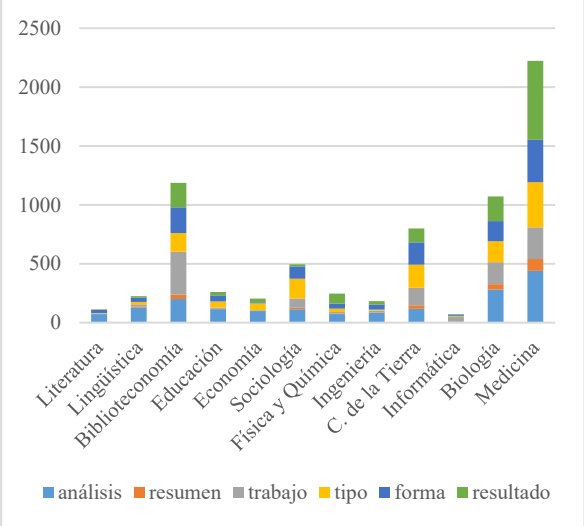

Figura 3. Seis nombres con IDF <0,05.

Como se puede apreciar, los nombres en la Figura 3 presentan una dispersión más homogénea que en la Figura 2. Por ejemplo, el nombre *episodio* presenta un gran predominio en los subdominios de Física y Química y Medicina, y no aparece en 3 de los 12 subdominios, mientras que el nombre *análisis* está bien distribuido, con ocurrencias proporcionales en los 12 subdominios.

En la primera fase de clasificación de los posibles nombres que se descartarían por el IDF, la mayoría se pudo clasificar en función de su especificidad en determinadas disciplinas, gracias al análisis de los contextos en los que aparecen dichas palabras en el corpus, así como a la ayuda de los diccionarios, como, por ejemplo, los nombres *geometría* o *motor*. Sin embargo, en línea con la problemática expuesta en la sección 1, se llevó a cabo una revisión más exhaustiva especialmente de aquellos nombres que se encuentran en un continuum entre la lengua general y la lengua académica, p.ej. *rol*, así como de los nombres polisémicos como *barrera,* que presenta un sentido más metafórico de 'obstáculo', y otro sentido de 'valla […] u otro obstáculo semejante con que cierra el paso' (DLE, f.s., def. 1), con el fin de identificar los sentidos más productivos en el discurso académico.

En relación con el análisis de las colocaciones (F2), un número muy reducido de nombres académicos presentaron únicamente un colocativo. En estos casos, si la colocación no presentó un nivel medio-alto de distribución, la base se eliminó. Por ejemplo, con el nombre *especificación*, únicamente se identificó el colocativo *cumplir*, y la colocación presentó una distribución heterogénea, con un 93% de ocurrencias en Ciencias Físicas (ejemplo 1):

(1) "Se diseñó la estructura de pavimento con agregados de La Calera, por *cumplir con* todas las *especificaciones*".

Por otra parte, la gran mayoría de nombres académicos presentaron ≥2 colocativos, por lo que fue necesario un análisis más detallado de la distribución de cada colocación para definir una media. Por ejemplo, con la base *productividad,* se identificaron los colocativos *aumento, alta* y *mayor,* que tienen una distribución medio-alta, como *productividad alta,* que presenta un 40% de ocurrencias en Ciencias Físicas, un 40% en Biología y Medicina y un 20% en Artes y Humanidades, una DE de 0,19, y aparece en 3 subdominios. Los casos que conllevaron más dudas se corresponden con aquellas bases que presentan 2-3 colocativos y una distribución media de colocaciones. Por ejemplo, con la base *indicación* se identificaron los colocativos *clara* y *precisa*: la colocación *indicación precisa* presenta una distribución homogénea, con una DE de 0,32, un porcentaje de aparición de un 9% en Artes y Humanidades, un 19% en Ciencias Sociales y un 72% en Biología y Medicina, y una aparición en 4 subdominios; sin embargo, la colocación *indicación clara* únicamente aparece

en los dominios de "ciencias blandas", con un 22% de ocurrencias en Artes y Humanidades y un 78% en Ciencias Sociales, una DE de 0,37 y una aparición en 3 subdominios. Debido a que únicamente se presentan dos colocaciones y la media de su distribución no proporciona indicios definitivos sobre su inclusión o descarte como base académica, en estos casos se contrasta el nombre con las listas de vocabulario académico en inglés. Cabe destacar que, en esta fase, también se identificaron 16 nombres que no presentaron ningún colocativo y, por lo tanto, no pudieron analizarse y pasaron a la fase 3. Como hemos mencionado en la sección 1, a pesar de que el objetivo principal del presente trabajo sea filtrar la lista de colocaciones académicas, los nombres sin colocativos se incluyen en este análisis con el propósito de obtener también una lista completa y más refinada de nombres académicos del español.

El análisis de las colocaciones también ha ofrecido indicios sobre el contraste de uso de las colocaciones en la lengua general y la lengua académica: la colocación *jugar un rol* no presenta una frecuencia alta ni una buena distribución en el discurso académico, pues su uso es más extendido en la lengua general, mientras que *desempeñar* y *ejercer un rol* presentan una distribución más homogénea entre los dominios y una frecuencia ligeramente mayor en el discurso académico.

A su vez, hemos podido observar casos en los que un nombre puede combinarse con colocativos y conformar colocaciones que están distribuidas de forma más homogénea que con otros colocativos: por ejemplo, con la base *puntuación,* identificamos la colocación *otorgar una puntuación,* que presenta una buena distribución, con un 14% de apariciones en AH, un 57% en CS, y un 39% en CF, y con una DE de 0,24; contrariamente, la colocación *puntuación mínima*, presenta un distribución heterogénea, con un 14% de ocurrencias en CS y un 86% en CF, una DE de 0,41 y con presencia únicamente en 2 subdominios. Estos casos indican que la base debe ser incluida en la lista, ya que es un nombre utilizado frecuentemente en distintos textos académicos, pero constituye colocaciones que se utilizan con más frecuencia en algunas disciplinas que en otras, probablemente debido a la variedad disciplinar.

En definitiva, hemos observado que los nombres que han tenido que pasar por distintas fases de análisis se corresponden especialmente con los nombres polisémicos, que poseen al menos dos sentidos distintos (ej. *barrera*) y los nombres que también se utilizan frecuentemente en la lengua general y, por lo tanto, no evidencian su especificidad en el discurso académico, como los nombres *unión* o *rol*.

6 Conclusiones

En este artículo se ha presentado una metodología para proponer una lista de nombres del discurso académico en español a partir de la lista *SpAKWL* (García-Salido 2021), aplicando criterios que identifiquen mejor la interdisciplinariedad. Aunque el objetivo principal es obtener una lista de colocaciones académicas que se integrará en HARTA, como objetivo secundario, hemos obtenido una lista de nombres académicos más prototípicos del discurso académico. Partiendo de la medida de IDF que es comúnmente utilizada para identificar de forma automática los nombres más asociados a la terminología, hemos aplicado diferentes análisis para valorar su efectividad y corroborar que los nombres identificados pueden descartarse.

Los resultados han demostrado que con esta metodología es posible identificar la interdisciplinariedad y establecer la lista de nombres académicos junto con una lista de colocaciones que puedan ser integradas en una herramienta que ayude a redactar textos académicos (HARTA). Específicamente, se ha obtenido un criterio para eliminar la terminología e identificar el vocabulario que puede ser utilizado independientemente de la disciplina y que ayuda a describir actividades y procesos académico-científicos y a estructurar la argumentación. No obstante, los resultados siguen remarcando la necesidad de desambiguar los sentidos de las palabras y la posibilidad de que, aunque las bases sean interdisciplinares en

el discurso académico, las colocaciones pueden utilizarse en unas disciplinas más que en otras.

El presente estudio forma parte de un proyecto de investigación más amplio, en el cual, a partir de este trabajo, planteamos integrar la nueva versión de la *SpAKWL* en HARTA de manera que, a partir de un texto, la herramienta pueda detectar las palabras académicas y sugerir las colocaciones correspondientes, en línea con lo que proponen herramientas como *LEAD* (Paquot 2012) o *Collocaid* (Frankenberg-Garcia et al. 2019) para la escritura académico-científica en inglés.

Agradecimientos

Este estudio ha sido posible gracias a la financiación del Ministerio de Ciencia e Innovación (PID2019-109683GB-C21); del Centro de Investigación de Galicia "CITIC", financiado por la Xunta de Galicia y la Unión Europea (FEDER GALICIA 2014-2020), con la ayuda ED431G 2019/01; y del Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020.

Bibliografía

Ackermann, K. y 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.

Ahumada, I., Zamorano, J. P., García, E. D. R. y Lara, I. A. 2011. Design and development of Iberia: a corpus of scientific Spanish. *Corpora*, 6(2): 145-158.

Alonso-Ramos, M., García-Salido, M. y 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 y V. Baisa (Eds.), *Proceedings of eLex 2017 conference*, páginas 571-586. Leiden, the Netherlands.

Cambridge Dictionary. Consultado el 27 de marzo de 2022 en: https://dictionary.cambridge.org/us/dictionary/.

Cobb, T., y Horst, M. 2004. Is there room for an academic word list in French?. En P. Bogaards y B. Laufer (Eds.), *Vocabulary in a Second Language. Selection, acquisition, and testing*, páginas 13-38, John Benjamins (Amsterdam/Philadelphia).

Coxhead, A. 2000. A new academic word list. *TESOL Quarterly*, 34(2): 213–238.

Drouin, P. 2007. Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliquée*, 7(2): 45-64.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., y Sharma, N. 2019. Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, 31(1): 23-39.

García-Salido, M. 2021. Compiling an Academic Vocabulary List of Spanish. Disponible en: https://doi.org/10.13140/RG.2.2.27681.33123.

Garcia, M. y Gamallo, P. 2016. Yet another suite of multilingual NLP tools. En J. P. Leal J. L. SierraRodríguez et al. (Eds.), *Languages, Applications and Technologies. Communications in Computer and Information Science*, páginas 65–75, Springer (Cham).

Gardner, D., y Davies, M. 2013. A new academic vocabulary list. *Applied Linguistics*, 35(3): 305–327.

Gilquin, G., Granger, S., y Paquot, M. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes*, 6(4): 319-335.

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

Hatier, S., Augustyn, M., Tran, T. T. H., Yan, R., Tutin, A., y Jacques, M. P. 2016. French cross-disciplinary scientific lexicon: extraction and linguistic analysis. En *Proceedings of EURALEX*, páginas 355-366, Ivane Javakhishvili Tbilisi State University (Tbilsi).

Hyland, K. y Tse, P. 2007. Is there an "academic vocabulary"?. *TESOL quarterly*, 41(2): 235-253.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1): 4–21.

Jones, K. S. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1): 11-21.

Kilgarriff, A. y Renau, I. 2013. *esTenTen*, a vast web corpus of Peninsular and American Spanish. *Procedia-Social and Behavioral Sciences*, 95: 12-19.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J. y Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography*, 1(1): 7–36.

Lei, L., y Liu, D. 2018. The academic English collocation list: A corpus-driven study. *International Journal of Corpus Linguistics*, 23(2): 216-243.

LINGUEE. Consultado el 28 de marzo de 2022 en: http://www.linguee.es.

Mel'čuk, I. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of phraseology*, 3(1): 31-56.

Nivre, J., Marneffe, M.-C. D., Ginter, F., Goldberg, Y., Manning, C. D., Mcdonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R. y Zeman, D. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. En *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), páginas 1659–1666, European Language Resources Association (ELRA).

Oxford English Dictionary. Consultado el 27 de marzo de 2022 en: y https://www.oed.com/.

Padró, L. y Stanilovsky, E. 2012. Freeling 3.0: Towards wider multilinguality. En N. Calzolari et al., (Eds.), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC2012), páginas 2473–2479, European Language Resources Association (ELRA).

Paquot, M. 2007. Towards a productively-oriented academic word list. En J. Walinski, K. Kredens, y S. Gozdz Roszkowski (Eds.), *Practical Applications in Language and Computers 2005*, páginas 127–140. Peter Lang (Frankfurt am main).

Paquot, M., y Bestgen, Y. 2009. Distinctive words in academic writing: A comparison of three statistical tests for keyword extraction. *Language and Computers*, 68(1): 247 269.

Paquot, M. 2012. *The LEAD dictionary-cum-writing aid: An integrated dictionary and corpus tool*. En S. Granger y M. Paquot (Eds.), *Eletronic lexicography,* páginas 161-186, Oxford University Press (Oxford).

Real Academia Española: *Diccionario de la lengua española*, 23.ª ed., (versión 23.5 en línea). Consultado el 25 de marzo de 2022 en: https://dle.rae.es.

Sebastián-Gallés, N., Martí Antonín, M.A., Carreiras Valiña, M. F., y Cuetos Vega, F. 2000. *LEXESP: Léxico informatizado del español*. Barcelona: Edicions de la Universitat de Barcelona.

Straka, M., Hajic, J. y Straková, J. 2016. Udpipe: Trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. En *Proceedings of the 10th International Conference on Language Resources and Evaluation* (LREC 2016), páginas 1659–1666, European Language Resources Association (ELRA).

Tutin, A. 2007a. Autour du lexique et de la phraséologie des écrits scientifiques. *Revue française de linguistique appliquée*, 12(2), 5-14.

A Anexo 1: Nombres descartados a partir del IDF

premisa, trayectoria, especialista, déficit, sustrato, concordancia, trabajador, exclusión, ciudadano, episodio, fiabilidad, ejemplar, prioridad, geometría, infraestructura, señal, madurez, patología, creencia, amplitud, reproducción, paradigma, procedencia, almacenamiento, implantación, complicación, corrección, apertura, desplazamiento, motor, tipología, venta, puntuación, consenso,

ejecución, dispositivo, meta, asignación, autonomía, lector, continuidad, carencia, compuesto, bibliografía, supuesto, costo, normativa, emisión, correspondencia, variante, especificación, instalación, reto, experto, descenso, unión, formulación, expectativa, puesta, mediana, motivación, vínculo, inconveniente, departamento, productividad, desempeño, incertidumbre, plataforma, tejido, tensión, experimento, diferenciación, economía, barrera, satisfacción, requerimiento, dosis, sesgo, acumulación, entrevista, fundamento, regulación, expansión, explotación, transporte, abundancia, promoción, instancia, eliminación, separación, fracción, síntoma, heterogeneidad, efectividad, espectro, preferencia, difusión, presente, predominio, afirmación, transferencia, aceptación, gráfico, distinción, prevención, sugerencia, dispersión, fragmento, énfasis, varianza, canal, indicación, estadio, iniciativa, rol, procesamiento, transición, estándar, potencia.

B Anexo 2: Nombres descartados tras las tres fases

emisión, fracción, tejido, geometría, puesta, presente, prevención, ciudadano, compuesto, sustrato, trabajador, infraestructura, transferencia, carencia, explotación, estadio, transición, transporte, exclusión, varianza, ejemplar, venta, departamento, entrevista, síntoma, *dosis, patología, episodio, creencia, madurez, costo, abundancia, espectro, fragmento, iniciativa, economía, autonomía, potencia, prioridad, dispositivo, expectativa, incertidumbre, dispersión*, *preferencia, tensión*, *inconveniente, déficit, amplitud, desplazamiento, plataforma, requerimiento, expansión, separación, implantación, complicación*, *concordancia, fiabilidad, trayectoria, resto, afirmación, barrera, corrección, almacenamiento*, *formulación*, *acumulación*, *regulación*, *procesamiento, especificación, efectividad, fundamento, distinción, asignación.*