TEISEL
Tecnologías para la investigación en segundas lenguas
ISSN 2696-676X
Universitat de Barcelona
Facultat de Filologia i Comunicació

Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español

Eleonora GUZZI (Autor para correspondencia)
Universidade da Coruña (España)
eleonora.guzzi@udc.es
https://orcid.org/0000-0001-5552-9444

Margarita ALONSO RAMOS
Universidade da Coruña (España)
margarita.alonso@udc.es
https://orcid.org/0000-0002-1353-9270

Resumen: Este artículo presenta la herramienta en línea HARTA (http://www.dicesp.com:8083/), que combina diccionario y corpus, acorde con la corriente de los últimos años en lexicografía. HARTA se centra en las combinaciones léxicas académicas (CLA) en español. Las CLA abarcan fenómenos de naturaleza variada: tanto colocaciones (confirmar/refutar una hipótesis) como lo que hemos englobado bajo el término de fórmulas (sin embargo, por otra parte, como ya hemos señalado, etc.). Con el término de CLA, por tanto, nos referimos a segmentos de palabras recurrentes en el discurso académico, que pueden ser o no composicionales y que pueden cumplir una función discursiva ('comparar', 'reformular', 'expresar certeza o posibilidad', etc.), como es el caso de las fórmulas. Para su descripción, además de apoyarnos en la Teoría Sentido-Texto, aportamos datos cuantitativos del corpus académico del que hemos extraído la lista de CLA (frecuencia y distribución en diferentes campos científicos). Una vez presentada la metodología con la que hemos obtenido los datos, describimos la arquitectura de HARTA para mostrar diferentes entradas de CLA y los diversos modos de acceder a la información. Antes de finalizar con las líneas de investigación en curso, ofrecemos un pequeño estudio experimental sobre la usabilidad de la herramienta.

Palabras clave: español con fines académicos; escritura; lexicografía; combinaciones léxicas

Català:

Descripció i usabilitat d'HARTA, una eina d'ajuda per a la redacció de textos acadèmics en espanyol

***Resum**: Aquest article presenta l'eina en línia HARTA (http://www.dicesp.com:8083/), que combina diccionari i corpus, d'acord amb el corrent dels últims anys en lexicografia. HARTA es centra en les combinacions lèxiques acadèmiques (CLA) en espanyol. Les CLA comprenen fenòmens de naturalesa variada: tant col·locacions (confirmar/refutar una hipótesis 'confirmar/refutar una hipòtesi') com el que hem englobat sota el terme de fórmules (sin embargo 'no obstant això', por otra parte 'd'altra banda', como ya hemos señalado 'com ja hem assenyalat', etc.). Amb el terme CLA, doncs, ens referim a segments de paraules recurrents en el discurs acadèmic, que poden ser o no composicionals i que poden complir una funció discursiva ('comparar', 'reformular', 'expressar certesa o possibilitat', etc.), com és el cas de les fórmules. Per a la seva descripció, a més de basar-nos en la Teoria Sentit-Text, aportem dades quantitatives del corpus acadèmic del qual hem extret la llista de CLA (freqüència i distribució en diferents camps científics). Una vegada presentada la metodologia amb la qual hem obtingut les dades, descrivim l'arquitectura d'HARTA per mostrar diferents entrades de CLA i les diverses maneres d'accedir a la informació. Abans de finalitzar amb les línies de recerca en curs, oferim un petit estudi experimental sobre la usabilitat de l'eina.*

***Paraules clau**: espanyol amb fins acadèmics; escriptura; lexicografia; combinacions lèxiques*

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

English:

Description and usability of HARTA, a tool to help writing academic texts in Spanish

***Abstract**: This article presents the building-up of HARTA (http://www.dicesp.com:8083/), an online tool that combines dictionary and corpus, in line with the trend of recent years in lexicography. HARTA focuses on academic lexical combinations (ALCs) in Spanish. ALCs cover phenomena of a varied nature: both collocations (confirmar/refutar una hipótesis 'to confirm/ to refute a hypothesis') and what we have lumped together under the umbrella term formula (sin embargo 'however', por otra parte 'on the other hand', como ya hemos señalado 'as we have already noted', etc.). By the term of ALC we refer to recurrent word segments in academic discourse, which may or may not be compositional and which may fulfil a discursive function ('to compare', 'to reformulate', 'to express certainty or possibility', etc.), as is the case of formulas. For their description, in addition to relying on Meaning-Text Theory, we provide quantitative data from the academic corpus from which we have extracted the list of ALCs (frequency and distribution in different scientific disciplines). Once we have presented the methodology used to obtain the data, we describe the architecture of HARTA to show different entries for ALCs and the different ways of accessing the information. Before concluding with the lines of research in progress, we offer a short experimental study on the usability of the tool.*

***Keywords**: Spanish for academic purposes; writing; lexicography; lexical combinations*

Introducción

Saber escribir es una competencia transversal e interlingüística que no es fácil de adquirir. A pesar de la multimodalidad del discurso actual, el texto escrito sigue siendo imprescindible, independientemente de su soporte físico, por lo que los miembros de la comunidad académica deben saber escribir observando las formas convencionales de los textos académicos. Sin embargo, cuando los estudiantes entran en la universidad no disponen de herramientas que les faciliten la producción de textos. Además, el estudiante universitario en España debe mostrar destreza en escritura académica tanto en las diferentes lenguas del Estado como en inglés, pero, paradójicamente, dispone de muchos más recursos para esta última (McCarthy & O'Dell, 2008; Swales & Feak, 2012; Lea, Bull & Webb, 2014).

En los últimos años, la escritura académica se ha convertido en un objeto de estudio prioritario, especialmente en el ámbito anglosajón, dado el papel de *lingua franca* que esta lengua ejerce en el ámbito académico internacional (entre otros, Hyland & Shaw, 2016; Tusting et al., 2019; Manchón & Matsuda, 2016). También en el ámbito hispano y, en particular, en diferentes países de Sudamérica (Carlino, 2005; Ramírez, 2013; Natale & Stagnaro 2016; Navarro & Aparicio, 2018) se han interesado desde hace tiempo por el aprendizaje de la escritura académica. La recogida de datos en la Encuesta Europea sobre la Escritura Académica (Marín et al., 2015) es una muestra de que, en efecto, en los últimos años está habiendo un proceso de reflexión sobre la escritura académica. Prueba de ello es que existen abundantes trabajos sobre el uso habitual de la escritura como herramienta epistémica (Cassany & Morales, 2009), así como sobre el propio proceso de escritura (Castelló, 2007). Con todo, el principal interés se ha centrado en las diferencias entre géneros académicos (Regueiro Rodríguez & Sáez Rivera, 2013; Sanz Álava, 2007; Perea Siller, 2013; Núñez Cortés, 2015). En esa línea, cabe destacar los estudios sobre tradiciones discursivas de la Escuela de Valparaíso originada en torno a Giovanni Parodi, que cuentan con un enfoque psicosociolingüístico (Parodi, 2010). También son señeros los trabajos coordinados por Montolío (2014) y no faltan estudios sobre las estrategias gramaticales utilizadas en los textos académicos, por ejemplo, sobre la despersonalización (Álvarez, 2013). Sí faltan, en cambio, estudios sobre

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

el léxico y la fraseología característicos del español académico. Así, Figueras (2016) señala la escasez de estudios sobre fraseología académica y reclama la elaboración de repertorios (listas y concordancias) extraídos de corpus que permitan identificar las combinaciones léxicas más productivas del español académico, para posibilitar así su enseñanza no solo a estudiantes extranjeros. El interés en el español académico ha estado especialmente dirigido a estudiantes no nativos, como se muestra ya en el proyecto pionero ADIEU (Vázquez, 2001), así como en los cursos organizados por universidades como la Universidad de Alicante desde 2010 (Pastor Cesteros, 2022). Sin embargo, ya se ha señalado repetidamente que el dominio de la escritura académica está más ligado a la experiencia que al hecho de ser nativo (Römer, 2009). El problema es que los expertos cada vez escriben menos en español debido a lo que podríamos llamar *globalización académica* (Zusman, 2022) y se ven obligados a escribir en inglés (Mur-Dueñas, 2012). Por todo ello, es importante tener recursos de consulta para todos los miembros de la comunidad académica (expertos y noveles, nativos y no nativos).

Aunque la escritura académica es un fenómeno poliédrico, pensamos que un factor determinante para adquirir destreza en la escritura reside en el dominio de la fraseología. Como ya se ha mostrado en los estudios sobre inglés académico (Biber et al., 2004; Hyland, 2008; Paquot, 2012), la naturaleza convencionalizada de ese discurso proviene, en gran medida, de las *combinaciones léxicas académicas* (CLA). Nos referimos a segmentos de palabras recurrentes en el discurso académico, que pueden ser o no composicionales y que pueden cumplir una función discursiva ('comparar', 'reformular', 'expresar certeza o posibilidad', etc.). Las CLA, por tanto, abarcan fenómenos de naturaleza variada: desde colocaciones (*confirmar/refutar una hipótesis*) hasta marcadores discursivos (*por otra parte*, *sin embargo*) y secuencias no registradas habitualmente en los diccionarios españoles pero recurrentes (p.ej. *como ya hemos señalado*), incluidas en los llamados *lexical bundles* (Biber, Conrad & Cortes, 2004). Este es el objeto de la herramienta HARTA (http://www.dicesp.com:8083/ *Herramienta de Ayuda a la Redacción de Textos Académicos*).

Desde el inicio, HARTA fue diseñada como una herramienta que combina diccionario y corpus, en línea con la corriente de los últimos años en lexicografía (Asmussen, 2013; Paquot, 2012). Es similar a herramientas como LEAD (https://leaddico.uclouvain.be/; *The Louvain EAP Dictionnary*; Granger & Paquot, 2015) y SciE-Lex (http://www.ub.edu/grelic/eng/scielex2/scielex.html; *Diccionario electrónico de combinaciones léxicas en el inglés científico*; Laso, 2022) para el inglés, y como LST (http://lst.demarre-shs.fr/ *Lexique Scientifique Transdisciplinaire*; Jacques & Tutin, 2018) y Dicorpus (https://dicorpus.aiakide.net/; Tran & Falaise, 2018) para el francés. Las diferencias residen en matices de enfoques teóricos, tamaño de los equipos de trabajo, financiación y diferentes finalidades, unos más enfocados hacia el usuario final y otros más hacia la investigación, pero todos ellos se centran *grosso modo* en lo que hemos llamado CLA. En español, el recurso pionero fue Estilector (http://www.estilector.com/; Nazar & Renau, 2023), especialmente centrado en la detección de problemas textuales y gramaticales, entre los cuales, problemas de léxico (p.ej. de marcadores discursivos) e incorpora, en menor medida, correcciones gramaticales y

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

ortográficas, y ya desde hace unos años destaca arText (http://sistema-artext.com/; Da Cunha, Montane & Hysa, 2017) que atiende más a la estructura de los textos en diferentes secciones, aunque también ofrece recomendaciones léxicas. Fue precisamente para colmar la escasez de recursos para el español académico por lo que se concibió el diseño de HARTA. Durante las diferentes fases de desarrollo, la herramienta ha ido cambiando y aumentando su base de datos y sus funcionalidades. En lo que sigue, describiremos cómo fue construida, su estructura y su uso.

Este artículo se organiza de la siguiente manera. En la sección 1 nos centraremos en el contenido de HARTA, es decir, en los diferentes tipos de CLA y en las funciones discursivas. La sección 2 presentará la metodología (corpus y extracción de las CLA), para pasar en la sección 3 a describir la arquitectura de la herramienta en donde mostramos cómo se definen las CLA. La sección 4 la utilizamos para mostrar brevemente un estudio experimental sobre la usabilidad de HARTA. Por último, presentaremos las conclusiones, así como las líneas en curso de investigación.

1. Tipos de combinaciones léxicas académicas en HARTA

El contenido de HARTA está formado por unidades que pertenecen al *léxico académico*. Sin seguir una definición estricta de qué se considera como tal (Paquot, 2010, p. 28), nos centramos en combinaciones que sirven para referirse a las actividades que caracterizan el trabajo académico y para organizar el discurso científico. En el ámbito francófono, es más usual hablar de *léxico científico trans/interdisciplinar* (Jacques & Tutin, 2018; Drouin, 2007). No se trata de terminología, sino de combinaciones léxicas que pueden ser utilizadas en diferentes campos científicos y que por tanto se consideran interdisciplinares (vid. Guzzi & Alonso-Ramos 2022).

Para la tipología de las CLA, nos apoyamos en la Teoría Sentido-Texto (TST; Mel'čuk, 2015, 2020) en donde rigen dos principios esenciales para considerar una combinación fraseológica: 1) composicionalidad y 2) restricción en la selección y en la combinatoria. Con respecto al primero, un sintagma es semánticamente composicional si su significado se corresponde con la suma del significado de sus componentes. Así, locuciones como *sin embargo* son no composicionales, mientras que una colocación como *extraer una conclusión* o un segmento recurrente como *en el presente estudio* son perfectamente composicionales, en donde cada componente contribuye al significado global. La composicionalidad no debería ser confundida con la transparencia (Mel'čuk, 2015, p. 62). Esta propiedad caracteriza una expresión desde el punto de vista de su comprensión. Por esta razón, la transparencia admite grados y puede diferir de una persona a otra. Así, una expresión totalmente transparente es necesariamente composicional, pero no ocurre lo mismo a la inversa; por ejemplo, si un hablante no sabe lo que significa el verbo *concernir*, no podrá adivinar el significado de la combinación *en lo que concierne a*, aunque sea totalmente composicional. Por lo tanto, una expresión composicional puede no ser transparente. Si una expresión es totalmente composicional, la razón por la que podría seguir considerándose fraseológica reside en la forma en que se eligen y se combinan sus componentes. Y pasamos al segundo principio que regula las combinaciones fraseológicas.

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

En nuestro marco, cuando un sintagma es *libre*, cada uno de sus componentes léxicos se selecciona estrictamente por su significado, independientemente de la identidad léxica de los otros componentes. El término *libre* debe entenderse entonces estrictamente como que permite la selección de una unidad léxica independientemente de los demás componentes léxicos de la misma expresión (Mel'čuk, 2012, p.33). Así, sintagmas como *la probabilidad de que* o *al revisar la selección* son libres puesto que cada uno de sus componentes léxicos se selecciona por su significado y propiedades combinatorias, de conformidad con las reglas correspondientes del español (Mel'čuk, 2015, p.59). En cambio, un sintagma *no libre* (o *frasema lexémico*; Mel'čuk, 2021) no se construye a partir de sus componentes léxicos seleccionando cada uno de ellos individualmente y ordenándolos según las reglas estándar de la lengua. Las restricciones que operan en la producción de un frasema pueden tener lugar en distintos niveles. En el caso de las colocaciones y las locuciones, la restricción tiene lugar entre su significado y su expresión léxica: el significado de una locución, al igual que el de una colocación, se selecciona libremente, aunque no se expresa libremente. La expresión léxica está totalmente restringida para las locuciones. Así, el significado 'señalo que, aceptado lo dicho anteriormente, añado una opinión que parece opuesta a lo que acabo de decir, aunque pienso que no lo es y que las dos son verdad' es expresado como un todo por la locución *ahora bien*; el hablante no escoge ni *ahora* ni *bien* independientemente uno de otro. Sin embargo, para las colocaciones, la expresión léxica solo está parcialmente restringida; por ejemplo, el significado 'concluir' puede expresarse en español mediante el nombre *conclusión*, que se selecciona libremente, y los verbos de apoyo *extraer* y *obtener*, que se seleccionan en función de este nombre concreto. El nombre *apoyo* selecciona también *obtener*, pero no *extraer*.

La restricción puede darse en un nivel más profundo, cuando incluso la construcción del significado no es libre. Desde un contenido informativo como 'deseo saber la edad que tienes', en inglés se selecciona el significado 'cuánto eres de viejo', mientras que en español el significado seleccionado es 'cuántos años tienes'. Tanto *how old are you* como *cuántos años tienes* son completamente transparentes y composicionales, pero no son libres. Es lo que en TST se llaman *formulemas* (Mel'čuk 2015, 2020). Como veremos abajo, los textos académicos están repletos de formulemas que expresan diferentes funciones discursivas.

Los límites entre los tres tipos de CLA no siempre están del todo claros. La composicionalidad traza una frontera entre las locuciones, por un lado, y las colocaciones y formulemas, por otro. Cuando uno de los componentes es una palabra gramatical, la distinción es menos obvia. Por ejemplo, *sin duda* parece composicional porque su significado incluye 'sin' y 'duda'. Sin embargo, su significado incluye también un componente semántico discursivo que enfatiza las afirmaciones del hablante. Para HARTA, no se necesitan finas distinciones fraseológicas porque lo que interesa es agrupar por CLA que expresan o no funciones discursivas. Así, bajo el término de *fórmulas* se incluyen diferentes combinaciones que pueden ser consideradas locuciones, formulemas o simplemente segmentos recurrentes,

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

pero todas sirven para expresar funciones discursivas ('hacer énfasis', 'comparar', 'reformular', etc.)[1]. En cambio, las colocaciones no están asociadas a funciones discursivas, puesto que una misma colocación dependiendo del contexto puede servir para expresar diferentes funciones discursivas o solo una función referencial; por ejemplo, la colocación *extraer una conclusión* no se usa necesariamente para 'concluir', sino que puede ser usada para referirse a trabajos previos como en *las principales conclusiones extraídas por los autores nos llevan a* (…).

En lo que sigue, presentaremos brevemente las colocaciones y, con algo más de detalle, las fórmulas académicas.

1.1 Colocaciones académicas

Como ya hemos señalado, las colocaciones son composicionales y están formadas por dos unidades léxicas: la *base*, que se selecciona por su significado y el *colocativo*, que se escoge en función de la base. En diferentes publicaciones se puede encontrar la concepción de las colocaciones en el marco teórico de la TST (entre otras, Alonso-Ramos 2010, 2017). Aquí nos limitaremos a describir los tipos que hemos seleccionado para HARTA (para más detalle, vid. Guzzi, en preparación).

Nos basamos solo en colocaciones de base nominal que entren en los siguientes cuatro patrones sintácticos:

- N+Adj (*amod*) = *análisis detallado*, *hipótesis controvertida*, *estudio profundo*, etc.
- V+N (*obj*) = *llevar a cabo un análisis*, *plantear una hipótesis*, *abordar un estudio*, etc.
- N (*nsubj*) +V = *el análisis revela*, *la hipótesis plantea*, *el estudio demuestra*, etc.
- N (*nmod*[2]) de N= *proceso de análisis*, *contraste de hipótesis*, *objeto de estudio*, etc.

Las colocaciones se incluyen en las entradas de sus respectivas bases organizadas en estos cuatro grupos. Cada colocación recibe una glosa o mínima indicación del significado.

1.2 Fórmulas académicas y funciones discursivas

Bajo el término de *fórmulas*, incluimos toda combinación léxica recurrente que sirva para expresar una función discursiva en los textos académicos. Dado que hemos extraído las fórmulas utilizando métodos estadísticos, no se ha distinguido entre locuciones (no composicionales) y formulemas (composicionales), según la clasificación TST. Dentro de las fórmulas incluimos también segmentos recurrentes con un estatus léxico menos claro. Nos referimos a CLA como *en lo que se refiere a*, *como se ha señalado*, *el objetivo de este trabajo*, etc., que no suelen estar registradas en los diccionarios, pero que se sienten como prefabricadas y parte imprescindible de un texto académico. Lo mismo ocurre con las

[1] No utilizamos el término *marcador discursivo* (Martín Zorraquino & Portolés, 1999) o *partícula discursiva* (Briz et al. 2008) para evitar posibles desajustes con la interpretación de estos términos. Pensamos que la secuencia *como ya se ha dicho anteriormente* no encaja bajo ninguna posible definición de marcador discursivo, pero, sin embargo, nosotros podemos recogerla con otros marcadores bajo el término de *fórmula* interpretado como unidad pluriverbal con función discursiva.

[2] La nomenclatura empleada (*amod, obj, nsubj* y *nmod*) proviene de las relaciones sintácticas de dependencias universales (Nivre et al., 2016). Con todo, queremos aclarar aquí que en el patrón sintáctico N *de* N, *nmod* hace referencia a un dependiente nominal de otro nombre y funcionalmente corresponde a un atributo o a un complemento. El dependiente sintáctico es la base de la colocación.

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

fórmulas en inglés como las que aparecen en la lista de Simpson-Vlach & Ellis (2010), *on the other hand, it should be noted*, *as can be seen*, *it is clear that*, etc. Con la excepción de la primera, ninguna de las demás aparecen en diccionarios ingleses. Con respecto a los diccionarios españoles, también es difícil encontrar muchas de las fórmulas académicas; por ejemplo, *en otras palabras* o *por otra parte* no aparecen en los diccionarios de español usuales (Alonso-Ramos, 2022).

Las funciones discursivas han sido clave en el proceso de selección de fórmulas, ya que los candidatos de fórmulas extraídos que no podían recibir una función discursiva eran descartados. La tipología de funciones discursivas es el resultado de un proceso ascendente y descendente. Partiendo del corpus y siguiendo tipologías previas para el inglés (Biber, Conrad & Cortes, 2004; Hyland, 2008) y el francés (Kübler & Pecman, 2012), fuimos estableciendo la tipología de funciones discursivas. Por ejemplo, veíamos que la función de finalidad puede ser expresada por diferentes fórmulas que habían pasado los filtros estadísticos como *con el fin de*, *con el objeto de*, *con la finalidad de*, etc., y reteníamos la función de 'indicar finalidad'. Una vez establecida la tipología, volvimos al corpus para establecer la asociación de las fórmulas con las funciones. Así, por ejemplo, una vez que se ha establecido que la fórmula *como muestran los resultados* es una fórmula seleccionada, volvemos al corpus para asignarle su función discursiva correspondiente (García-Salido et al., 2019).

Inspirándonos en tipologías previas para otras lenguas, las funciones discursivas están clasificadas en tres grandes grupos: 1) Estructurar el texto; 2) Referirse al contenido de la investigación y 3) Posicionarse y dirigirse al lector. El grupo más numeroso es el primero en donde entran muchos marcadores discursivos, mientras que el segundo abarca más expresiones referenciales y el último proporciona fórmulas que sirven principalmente para expresar la modalidad. El total actual de funciones discursivas es de 34. En este momento, la lista contiene 699 fórmulas con funciones discursivas asociadas.

Durante el proceso de asignación de funciones discursivas, detectamos muchas fórmulas ambiguas en el sentido de que podían expresar funciones discursivas diferentes según el contexto. Hemos optado por desdoblar las fórmulas asignando entradas diferentes según su función discursiva. Uno de los casos más frecuentes se da con fórmulas que sirven para 'delimitar el tema del que se habla' (*framing*) y también para 'introducir un tema'. Es lo que ocurre con *en lo que se refiere a* 1 (http://www.dicesp.com:8083/search/formulas/5520) y *en lo que se refiere a* 2 (http://www.dicesp.com:8083/search/formulas/5713). En la Figura (1), mostramos fusionada parte de la información de cada entrada para poder contrastar las dos fórmulas con diferentes funciones.

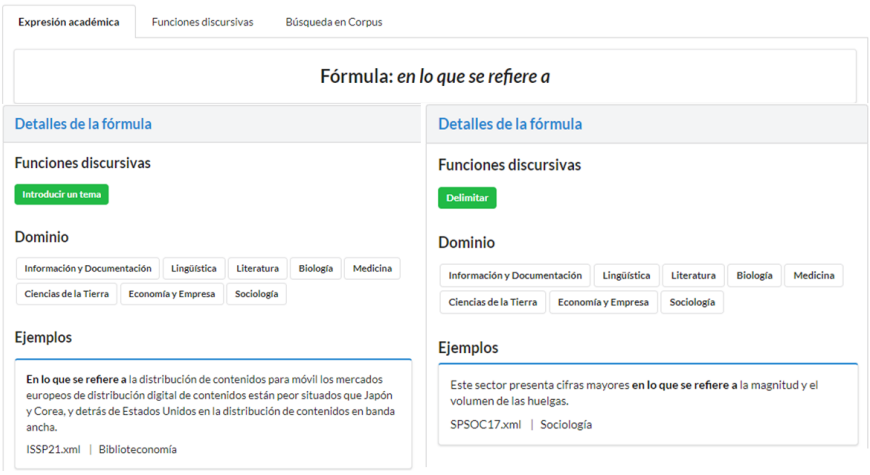
Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Figura 1

Entradas de la fórmula "en lo que se refiere a" con distinta función discursiva



Una situación diferente ocurre cuando una misma fórmula expresa simultáneamente dos funciones; por ejemplo, *alrededor de*, *cerca de*, *del orden de*, *algo más*, todas son fórmulas que expresan cantidad de un modo atenuado, por lo que les asignamos las dos funciones discursivas: 'expresar cantidad' y 'atenuar'. Un caso especialmente complejo lo aporta *por otro lado*. La hemos desdoblado en *por otro lado* 1 (http://www.dicesp.com:8083/search/formulas/5910) que sirve para 'añadir información' y no es correlativo de una fórmula previa; y tenemos *por otro lado* 2 (http://www.dicesp.com:8083/search/formulas/3713) que sirve simultáneamente para 'expresar oposición' y para 'ordenar', en correlación con *por un lado* (http://www.dicesp.com:8083/search/formulas/3722). Por el momento, los datos cuantitativos que figuran en HARTA no dan cuenta de estas distinciones, pero estamos en vías de poder extraer esta información (vid. Conclusiones).

2. Metodología: extracción de las CLA del corpus

Para alimentar HARTA hemos compilado diferentes corpus. En primer lugar, compilamos HARTA-Expertos (Alonso-Ramos et al., 2017), formado por los textos en español del corpus SERAC (Pérez-Llantada, 2014) y por 180 artículos recogidos de otras revistas científicas para equilibrar los campos científicos. El corpus se divide en cuatro campos principales, Artes y Humanidades (AH), Biología y Ciencias de la Salud (BCS), Ciencias Físicas e Ingeniería (CFI) y Ciencias Sociales y Educación (CS), que, a su vez, están subdivididos en doce disciplinas. Este conjunto de textos suma un total de 2.025.092 de palabras. Con el fin de poder explotar el corpus e identificar el vocabulario académico, se utilizaron distintos programas: LinguaKit (versión estable de marzo de 2018) (https://linguakit.com; Garcia & Gamallo, 2016) para la tokenización y lematización del corpus, FreeLing 4.0 (https://nlp.lsi.upc.edu/freeling/node/1; Padró & Stanilovsky, 2012) para el etiquetado morfológico, y UDPipe (https://bnosac.github.io/udpipe/en/; Straka, Hajic & Strakov, 2016) para el análisis sintáctico de dependencias

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

universales (Nivre et al., 2016), este último paso enfocado solo a la extracción de colocaciones. En cuanto a UDPipe (versión 2.1.), entrenamos un nuevo modelo de análisis sintáctico juntando los tres corpus de español: GSD, AnCora y PUD. Realizamos varias pruebas (únicamente con GSD, con AnCora, combinando dos corpus, etc.) y, finalmente, para el análisis empleamos la combinación de los tres corpus porque fue el que ofreció mejores resultados. A su vez, queremos destacar que empleamos distintas herramientas debido a la efectividad de las mismas en cada tipo de procedimiento: por ejemplo, utilizamos LinguaKit para identificar oraciones, tokenizar y lematizar, debido a que la salida de LinguaKit nos permite una mayor flexibilidad para la posterior conversión al formato *conllu*.

En relación con las colocaciones, en una primera etapa, se extrajo la lista de vocabulario académico del español (García-Salido, 2021) que se ha ido modificando según la productividad colocacional de los ítems (Guzzi & Alonso-Ramos, 2022). A partir de los nombres académicos de dicha lista, se extrajeron automáticamente alrededor de 40 mil candidatos de colocaciones para cuatro tipos de dependencias sintácticas como resultado del análisis con UDPipe (amod, nmod, nsubj, obj), siguiendo un umbral de frecuencia (≥5 ocurrencias en el corpus). Seguidamente, un grupo de lingüistas, basándose en criterios fraseológicos establecidos previamente y en criterios de distribución para cumplir con la interdisciplinariedad, seleccionaron 5.496 colocaciones de base nominal. Dada la baja proporción de colocaciones seleccionadas frente a los candidatos iniciales, es importante resaltar el considerable trabajo de filtrado y selección manual.

A su vez, se escogieron ejemplos para cada colocación seleccionada para su introducción en HARTA (vid. Figura 2):

Figura 2

Interfaz de selección de colocaciones y ejemplos

Por otra parte, para la extracción de fórmulas, hemos seguido la metodología utilizada en la bibliografía sobre inglés académico que gira en torno a los *lexical bundles* como principal unidad de análisis (Biber et al., 1999; Biber,

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Conrad, & Cortes, 2004; Cortes, 2004; Hyland, 2008; Salazar, 2014). A través de un programa informático, se realizó una extracción de n-gramas recurrentes, constituidos entre 2-6 n-gramas, siguiendo un criterio de frecuencia y de distribución (García-Salido et al., 2019). Esta extracción ha resultado ser más sencilla en comparación con las colocaciones, puesto que se trata de secuencias de palabras contiguas y no requieren de un corpus analizado sintácticamente. Sin embargo, como ya señalamos, tras la extracción automática, se llevó a cabo un arduo trabajo de filtrado manual para descartar secuencias únicamente gramaticales o que no cumplen con ninguna función discursiva, puesto que, para nuestros objetivos, no todos los *lexical bundles* son interesantes. Por ejemplo, *la probabilidad de que*, aunque pase los umbrales de frecuencia, será rechazado puesto que no expresa ninguna función discursiva. Por lo tanto, todas las fórmulas de HARTA son *lexical bundles*, pero no a la inversa.

En una segunda etapa, compilamos el corpus HARTA-Noveles, que contiene 2.230.153 palabras y está formado por trabajos de fin de grado y de máster de universidades españolas (Villayandre, 2018). El objetivo aquí era detectar posibles necesidades léxico-discursivas de los estudiantes. Con el fin de analizar contrastivamente las colocaciones de los expertos y de los noveles (Guzzi & Alonso-Ramos, en prensa), necesitamos ampliar el corpus de expertos, lo que llevó al corpus HARTA-Expertos-Plus. Este corpus, por tanto, contiene 21.068.482 palabras que proceden en parte del corpus inicial HARTA-Expertos y en parte de 3.870 artículos de investigación del Corpus Iberia (Ahumada, 2010), que suman un total de 19.043.390 palabras. El resultado actual del proyecto HARTA es el siguiente:

Tabla 1
CLA en el Proyecto HARTA

Corpus	Colocaciones[3]	Fórmulas
HARTA-Expertos	1.746	699
HARTA-Noveles	458	463
HARTA-Expertos-Plus	3.292	----
Total en Corpus	**5.496**	**1.162**
Descritas en HARTA	**3.204**	**699**

Para interpretar correctamente estos datos, es necesario hacer varias puntualizaciones. Incluimos en cada corpus las colocaciones nuevas que fuimos extrayendo de cada procesamiento. Así, al número inicial del corpus de expertos, se fueron añadiendo las nuevas colocaciones encontradas tanto en el corpus de noveles como en el de Iberia. Este último corpus no fue explotado para extraer fórmulas. En la última fila de la Tabla 1, indicamos el número de CLA descritas actualmente en la base datos pública de HARTA (3.204). La descripción de las colocaciones restantes (2.292) se está realizando en la fase actual del proyecto y estará disponible próximamente.

[3] Incluimos los *types*, es decir, las formas consignadas en la herramienta *HARTA*, no todas las ocurrencias o *tokens* en el corpus.

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

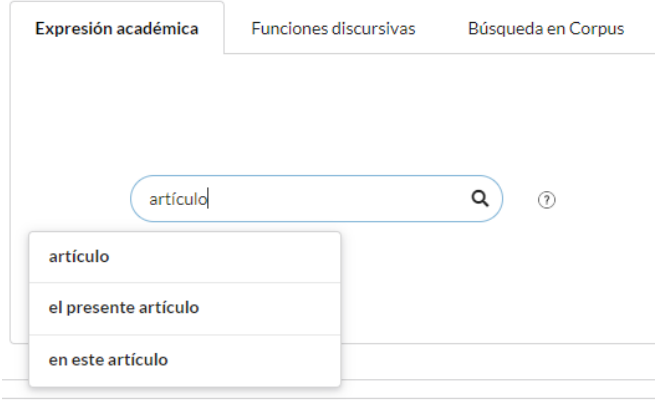
3. Arquitectura de HARTA

HARTA es un modelo híbrido formado por diccionario y corpus donde los límites entre estos dos recursos son cada vez más borrosos (Alonso-Ramos, 2009). En esta sección describiremos cómo se representan las CLA en HARTA: por una parte, se verá que las CLA se pueden consultar tanto por la base de datos léxica como por el corpus; por otra, que HARTA permite un acceso semasiológico por la pestaña "Expresión académica" o un acceso onomasiológico por la pestaña "Funciones discursivas". Con todo, dependiendo de si se trata de colocaciones o de fórmulas, se accede de un modo diferente y se ofrece también información diferente. A continuación, presentamos la información que se incluye en cada una de las pestañas de HARTA.

3.1 Expresión académica

Este es el acceso semasiológico para encontrar bien colocaciones, bien fórmulas. Para las primeras, hay que escribir la base, es decir, el nombre; para las segundas, se puede acceder por cualquier componente de la fórmula. Así, por ejemplo, si el usuario escribe *artículo*, le aparecerá en primer lugar ese nombre y debajo todas las fórmulas que incluyen ese nombre (vid. Figura 3).

Figura 3
Acceso por Expresión académica



Si se quieren consultar las colocaciones de *artículo*, hay que clicar sobre el nombre y aparecen los cuatro tipos sintácticos de colocaciones. Para conocer qué verbos coocurren con el nombre *artículo* como sujeto, se debe clicar en *artículo*+verbo, que proporciona una lista de verbos con la glosa correspondiente. La información específica de cada colocación se encuentra tras clicar en el colocativo. Ahí aparecen dos ejemplos seleccionados manualmente, así como los campos y subcampos con los porcentajes correspondientes en que aparece la colocación (vid. Figura 4). Como vemos, la información sobre las colocaciones es como una muñeca rusa que se puede desplegar más o menos según los intereses del usuario.

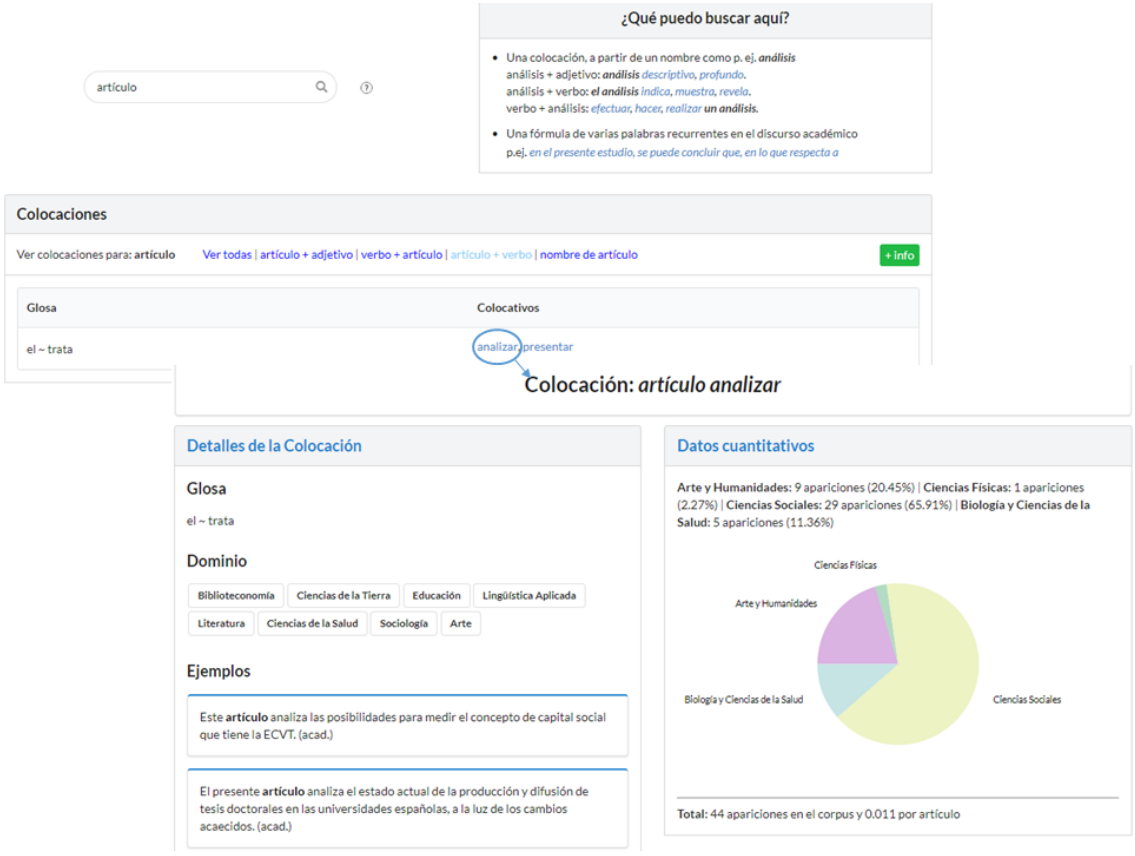
Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Figura 4

Colocaciones del nombre "artículo" en HARTA



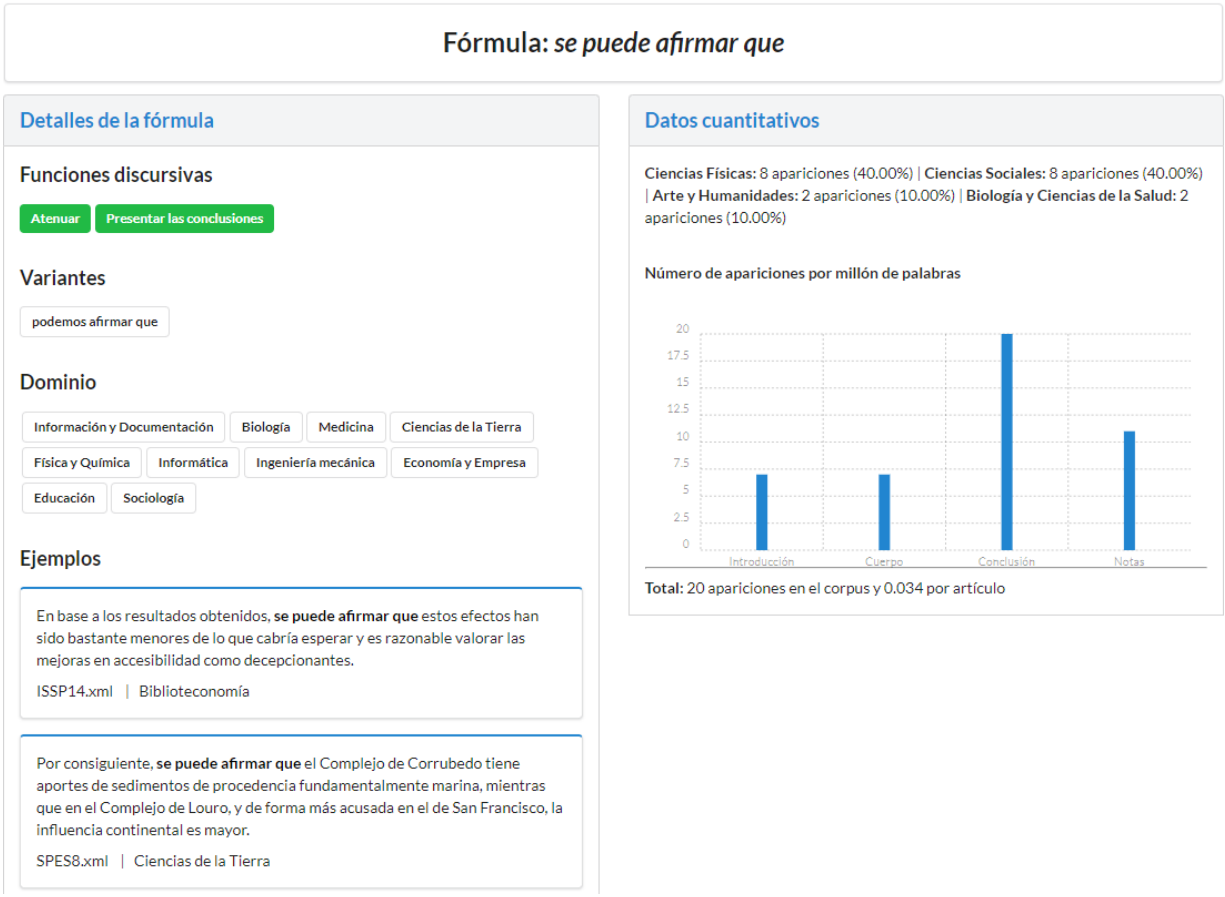
Si para encontrar las colocaciones el usuario parte de la base para ir avanzando por pantallas, el acceso a las fórmulas es más directo puesto que aparecen listadas en el desplegable inicial. Al igual que en el caso de las colocaciones, para las fórmulas el usuario encontrará dos ejemplos seleccionados, así como los campos y subcampos en los que se utiliza. También consignamos las partes textuales en las que aparecen, porque es un dato especialmente útil para la coherencia discursiva. Mientras que en las colocaciones incluíamos la glosa, en las fórmulas mostramos la función discursiva (Figura 5). Ahí se puede encontrar más de una función como en el caso de *se puede afirmar que*. Otra puntualización importante es que, dado que tenemos fórmulas ambiguas, la entrada de cada fórmula contendrá funciones diferentes (el caso ya mencionado de *por otra parte 1* y *por otra parte 2*). La entrada de algunas fórmulas incluye lo que hemos llamado *variantes* (Salazar, 2014) cuando se trata de variación gramatical especialmente en el caso de fórmulas que incluyen verbos. En general optamos por las formas en 3ª persona singular como forma canónica y tratamos la forma en 1ª persona plural como variante: *se puede afirmar que* frente a *podemos afirmar que*. Desde la propia entrada de la fórmula se puede clicar en la función discursiva y se accede al modo onomasiológico que describimos a continuación.

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Figura 5
Entrada de la fórmula "se puede afirmar que"



3.2 Funciones discursivas

Desde esta pestaña se accede a las 34 funciones discursivas organizadas en tres grupos. Al clicar en cada una de las funciones se listan todas las fórmulas que sirven para expresar esa función. Si el usuario clica sobre la fórmula, consulta la entrada, como hemos visto en la sección anterior. El primer grupo "Estructurar el texto" es el que posee un mayor número de funciones y de fórmulas, como podemos observar en la Tabla 2. Las funciones 'añadir información' y 'expresar consecuencia', que se incluyen en este grupo, son las que contienen un mayor número de fórmulas, 43 y 41 fórmulas respectivamente. También cabe destacar las funciones de 'expresar cantidad' y 'presentar la metodología', dentro del segundo grupo, por su elevado número de fórmulas (63 y 60, respectivamente). Contrariamente, algunas funciones son menos ricas y habrá que valorar si es necesario su mantenimiento, como es el caso de 'introducir una alternativa', con solo 2 fórmulas. De las 699 fórmulas descritas actualmente en HARTA, 32 tienen una doble función (como *se puede afirmar que*) y 27 son fórmulas ambiguas (como *por otro lado1*, *por otro lado2*). El recuento de fórmulas por funciones se muestra en la Tabla 2.

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Tabla 2
Datos cuantitativos de fórmulas y funciones discursivas

	Nº funciones	Nº fórmulas
Estructurar el texto	16	351
Referirse al contenido de la investigación	11	262
Posicionarse/dirigirse al lector	7	128

3.3 Búsqueda en corpus

El acceso por el corpus lo ofrecemos al usuario como un punto de apoyo, pero no constituye la funcionalidad principal de la herramienta. De hecho, como veremos en la siguiente sección, el usuario eficaz de HARTA apenas usa el corpus para sus búsquedas. La función del corpus en HARTA es proporcionar información para lo que no ha sido registrado previamente en la base de datos léxica. Así, puede darse el caso de que la base de datos no incluya una combinación específica, sea porque no pasó los filtros requeridos o simplemente porque no la consideramos académica, pero aparece en el corpus y el usuario puede consultar los ejemplos. La interfaz de búsqueda es muy simple: por forma (ej. *resultados*), por lema (ej. *resultado*) o por combinación de palabras (ej. *como resultado*); esta última solo requiere una secuencia de palabras cualquiera (Figura 6):

Figura 6
Búsqueda en corpus en HARTA

4. Poniendo a prueba la usabilidad de HARTA

Con el fin de evaluar si HARTA puede satisfacer las necesidades de sus usuarios potenciales, diseñamos un pequeño estudio experimental que nos sirvió primero para identificar cuáles eran esas necesidades y después para intentar adaptar HARTA a ellas. Las preguntas de investigación fueron las siguientes:

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

1) ¿Cuáles son las necesidades de los usuarios en cuanto a las CLA? Aquí queríamos hacer ver a los usuarios si son capaces de identificar sus necesidades y buscarlas en una herramienta. Así, por ejemplo, se les pide buscar algún otro verbo que va con *conclusión* o que ofrezcan una expresión similar a *hay que subrayar*. Una vez que se les muestra que tienen una necesidad podemos medir si HARTA responde o no.
2) ¿El corpus es un medio de resolver las necesidades vinculadas a las CLA? Dada la preponderancia que están tomando los corpus como recurso lexicográfico, queríamos medir si nuestra interfaz de corpus era suficiente.
3) ¿Pueden mejorar su actuación en una tarea de producción con HARTA sin formación previa? Ya en Alonso-Ramos y García-Salido (2019), mostramos que es posible diseñar herramientas que no exijan un entrenamiento para su uso.

A continuación, presentamos brevemente en qué consistió este estudio (para más detalles vid. Guzzi, en preparación).

4.1 Diseño del estudio experimental

La tarea consistió en buscar una alternativa, con y sin HARTA, a las colocaciones y fórmulas subrayadas en 10 oraciones. No se decía a los participantes de qué tipo de CLA se trataba en cada caso. Como ejemplo de las 10 oraciones, podemos mostrar una para cada tipo de CLA, en el (1) una colocación, y en el (2) una fórmula:

(1) *Actualmente se sigue trabajando en la metodología para extraer* ***conclusiones*** *más precisas*.
(2) *Con el objetivo de promover la dimensión social de la integración regional* […].

Participaron 12 personas distribuidas en tres grupos: 8 estudiantes de grado, master y doctorado y 4 profesionales (un profesor de instituto, un profesor de universidad, un asistente técnico y un corrector). Incluimos también 3 participantes no nativos, pero con un nivel de C1 o C2 en español (2 estudiantes de máster y 1 de doctorado). Todos recibieron por correo electrónico la tarea con las instrucciones. Debían leerlas, visionar los vídeos tutoriales de HARTA y registrar su pantalla y su voz en cuanto empezaran la tarea que debería durar alrededor de 30 minutos. El protocolo de *pensamiento en voz alta* (*think-aloud*) fue el más apropiado para nuestro propósito, porque nos permitió examinar el proceso del pensamiento de los participantes mientras hacían las búsquedas.

4.2 Resultados y discusión

Sin entrar en presentar el análisis de datos, respondemos aquí las preguntas anteriores. Con respecto a las necesidades de los usuarios relacionadas con las CLA, observamos que varios usuarios desconocen el modo de seleccionar colocativos y no son conscientes de que la elección del colocativo está controlada por la base. Por esa razón, piensan que buscar un sinónimo puede valer para la tarea de buscar una alternativa. Sin embargo, para el ejemplo de *extraer conclusiones*, no se trata de buscar un sinónimo de *extraer* sino de buscar otro verbo en función del nombre *conclusión*. El usuario que intentaba consultar HARTA por el colocativo buscando un sinónimo fracasaba en la tarea. Con respecto a las fórmulas, algunos usuarios se sorprendían cuando veían como lema la fórmula y antes probaban diferentes alternativas en lugar de escribir explícitamente en la pestaña "Expresión académica" la fórmula buscada. El usuario que

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

conseguía mejor puntuación era el que primero identificaba cuál era la función discursiva de la fórmula subrayada e iba directamente a la pestaña "Funciones discursivas" para buscar diferentes fórmulas que expresen la función requerida.

En cuanto a la eficacia de las búsquedas en corpus, comprobamos que el usuario eficaz no llega a la pestaña "Búsqueda en corpus" y, el que llega, es porque no ha sabido cómo encontrar la información antes. Sin embargo, nuestra interfaz de corpus no le sirve para responder a todas las preguntas. Como ya señalamos, nosotros diseñamos el corpus como un instrumento de respaldo para buscar más ejemplos de las fórmulas descritas o incluso posibles fórmulas no incluidas en la base de datos. Sin embargo, no permite búsquedas sofisticadas como "verbos a la izquierda o a la derecha de un nombre dado", que serían necesarias para encontrar colocaciones verbales, por ejemplo. Por lo tanto, el usuario que no había encontrado la información colocacional antes, tampoco la encontraba en el corpus y fracasaba en la tarea.

En relación con la mejora de la producción, hay que señalar que los resultados fueron esperanzadores. El sistema de puntuación se basa en unos parámetros que miden, principalmente, la adecuación en la elección del ítem léxico y la estrategia utilizada para encontrar la información en HARTA, dado que hay diversas maneras de llegar a la misma información, pero unas más directas que otras. Con una puntuación máxima de 30, comprobamos que 11 participantes obtuvieron una puntuación mayor de 15 y hubo tres participantes con puntuaciones respectivas de 27, 28 y 30 puntos. Lo más interesante es que algunos usuarios que sin HARTA no eran capaces de ofrecer alternativas o las que ofrecían no eran muy acertadas mejoraron su producción con la herramienta y lo que es destacable, sin una formación previa de la herramienta. Esto no sería posible con los diferentes diccionarios de partículas o de marcadores discursivos que existen en español, en los que se analizan con detalle estas formas y sus funciones discursivas, pero están más orientados al investigador que al usuario (Briz et al., 2008; Rodríguez, 2009; Santos Río, 2003).

Recapitulando, este experimento ha servido tanto para identificar mejor las necesidades de los usuarios como para mejorar la herramienta. Con respecto a las primeras, detectamos que varios usuarios tienen dificultades para ofrecer alternativas; es decir, es lo que Tarp (2009) llama *function-related needs*, necesidades que surgen en una situación extra o prelexicográfica. Y una vez en la herramienta, comprobamos también las necesidades que surgen durante el proceso de consulta (*usage-related needs*). La que más destacó era que varios usuarios no saben qué estrategias son necesarias para encontrar una colocación. De diferentes problemas que salieron a la luz durante este experimento, surgieron diferentes mejoras de HARTA que hemos ido implementado; por ejemplo, ahora hemos integrado mejor la información colocacional. También hemos implementado un enlace directo a la función discursiva dentro de la entrada de la fórmula, lo que facilita el paso del modo onomasiológico al semasiológico.

Conclusión y líneas futuras

Aunque HARTA, como todos los recursos lexicográficos, está siempre en proceso (de modificar contenidos, de aumentar la nomenclatura de CLA, de mejorar la usabilidad, etc.), pensamos que es un primer intento para poder

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

llenar el vacío que había con respecto al vocabulario académico en español. A partir de nuestra base de datos es posible crear una red de recursos interlingües con posibles vinculaciones al LEAD o al LST, por ejemplo.

Por el momento, hemos empezado por el vasco y contamos ya con una herramienta bilingüe (Alonso-Ramos & Zabala, 2022). Para ello, hemos compilado un corpus de noveles vasco al que hemos aplicado las mismas estrategias de extracción de fórmulas que al español con el fin de obtener fórmulas académicas en vasco. Con todo, dada la naturaleza aglutinante del vasco, en donde muchas de nuestras fórmulas se expresan con una sola palabra (*por consiguiente -ondorioz* 'consecuencia + INSTR'-), tuvimos que recurrir a otras estrategias basadas en semántica distribucional que nos llevaron a anotar el corpus académico. Si las fórmulas inicialmente fueron extraídas automáticamente y fueron filtradas y enriquecidas con la función discursiva para alimentar HARTA, ahora hemos anotado el corpus con esa lista para poder entrenar modelos de lenguaje y verificar si pueden aprender a identificar fórmulas con función discursiva automáticamente. Se trata de investigaciones en curso de las que esperamos rendir cuenta próximamente.

Otra investigación en curso que contamos con poder implementar en breve es una extensión de HARTA que permita a los usuarios incluir su propio texto y les proporcione un análisis en términos de CLA reconociendo las colocaciones y las fórmulas empleadas y enlazando con la base de datos de HARTA. De ese modo, por ejemplo, el usuario podrá escoger un colocativo diferente del que ha escrito en su texto solo con clicar en la base, y se le desplegará la información de HARTA. Son múltiples las posibles extensiones de HARTA, pero la tendencia ya iniciada hace tiempo es que los diccionarios se fusionen con herramientas de escritura, en la línea de CollocAid (Frankenberg-Garcia et al., 2019). Con todo, dado el gran avance reciente de los sistemas basados en inteligencia artificial como el chatGPT, es difícil poder predecir el futuro de las herramientas lexicográficas.

Agradecimientos

Esta publicación es resultado del proyecto I+D+i PID2019-109683GB-C21, financiado por MCIN/AEI/10.13039/501100011033. Agradecemos también el apoyo de la Xunta de Galicia a través de la ayuda ED431C 2020/11; del Centro de Investigación 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; y del Programa de Axudas á Etapa predoutoral da Xunta de Galicia, FSE Galicia 2014-2020.

Referencias

Ahumada, Ignacio (2010). El corpus Iberia como recurso para la traducción especializada. *Hikna. Revista de Traducción*, *9*, 9-24. https://doi.org/10.21071/hikma.v9i.5266

Alonso-Ramos, Margarita (2009). Hacia un nuevo recurso léxico ¿fusión entre corpus y diccionario?. En Pascual Cantos Gómez; Aquilino Sánchez Pérez (Eds.), *A Survey of Corpus-based Research. Panorama de investigaciones basadas en corpus* (pp. 1191-1207). AELINCO. http://www.dicesp.com/app/webroot/files/file/CILC%2009.pdf

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Alonso-Ramos, Margarita (2010). No importa si la llamas o no colocación, descríbela. En Carmen Mellado; Patricia Buján; Claudia Herrero; Nely Iglesias; Ana Mansilla (Eds.), *La fraseografía del s. XXI: Nuevas perspectivas de la fraseología del S. XXI* (pp. 55-80). Frank & Timme. http://www.dicesp.com/app/webroot/files/file/Alonso%202010(1).pdf

Alonso-Ramos, Margarita (2017). Diccionarios combinatorios. *ELiEs. Estudios de Lingüística del Español*, *38*, 173-201. https://doi.org/10.36950/elies.2017.38.8651

Alonso-Ramos, Margarita (2022). Academic lexical combinations for a writing tool. About their nature. En Leonid Iomdin; Jasmina Milicevic; Alan Polguère (Eds.), *Lifetime linguistic inspirations*, vol. 101 of Wiener Slawistischer Almanach - Sonderbände (pp. 29-44). Peter Lang.

Alonso-Ramos, Margarita; García-Salido, Marcos (2019). Testing the Use of a Collocation Retrieval Tool Without Prior Training by Learners of Spanish. *International Journal of Lexicography*, *32*(4), 480-497. https://doi.org/10.1093/ijl/ecz016

Alonso-Ramos, Margarita; García-Salido, Marcos; García-González, Marcos (2017). Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations, En Iztok Kosem; Jelena Kallas; Carole Tiberius; Simon Krek; Miloš Jakubíček; Vít Baisa (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 571-586), Leiden, the Netherlands. https://elex.link/elex2017/wp-content/uploads/2017/09/paper35.pdf

Alonso-Ramos, Margarita; Zabala, Igone. (2022). HARTAes-vas: Combinaciones léxicas para una Herramienta de ayuda a la redacción de textos académicos en español y en vasco. En Miguel A. Alonso; Margarita Alonso-Ramos; Carlos Gómez-Rodríguez; David Vilares; Jesús Vilares (Eds.), *Pre-conference Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing 2022: Projects and Demonstrations (SEPLN-PD 2022). Co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2022)* (pp. 25-28). https://ceur-ws.org/Vol-3224/paper06.pdf

Álvarez López, María Fátima (2013). *La despersonalización en el discurso académico escrito*. Tesis doctoral, Universidad de Alcalá de Henares. http://hdl.handle.net/10017/20049

Asmussen, Jörg (2013). Combined Products: Dictionary and Corpus. En Rufus H. Gouws; Ulrich Heid; Wolfgang Sheweickard; Herbert Ernst Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 1081–90). De Gruyter Mouton. https://doi.org/10.1515/9783110238136.1081

Biber, Douglas; Conrad, Susan; Cortes, Viviana (2004). *If you look at…*: Lexical bundles in university teaching and textbooks. *Applied Linguistics 25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, Douglas; Finegan, Edward; Johansson, Stig; Conrad, Susan; Leech, Geoffrey (1999). *Longman Grammar of Spoken and Written English* (1st ed.). Longman. https://doi.org/10.1177/0075424202250290

Briz, Antonio; Pons, Salvador; Portolés, José (Coords.) (2008). *Diccionario de partículas discursivas del español* (DPDE). http://www.dpde.es

Carlino, Paula (2005). *Escribir, leer y aprender en la universidad. Una introducción a la alfabetización académica*. Fondo de cultura económica. https://n2t.net/ark:/13683/p1s1/Uaw

Cassany, Daniel; Morales, Oscar Alberto (2009). Leer y escribir en la universidad: los géneros científicos. En Daniel Cassany (Ed.), *Para ser letrados. Voces y miradas sobre la lectura* (pp. 109-128). Paidós.

Castelló, Montserrat (Coord.). (2007). *Escribir y Comunicarse en contextos científicos y académicos. Conocimiento y estrategias*. Graó.

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

Cortes, Viviana (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for specific purposes*, *23*(4), 397-423. https://doi.org/10.1016/j.esp.2003.12.001

Da Cunha, Iria; Montané, M. Amor; Hysa, Luis (2017). The arText prototype: an automatic system for writing specialized texts. En Anselmo Peñas; Andre Martins (Eds.), *Proceedings of the EACL 2017 Software Demonstrations* (pp. 57–60). Association for Computational Linguistics. https://doi.org/10.18653/v1/e17-3015

Drouin, Patrick (2007). Identification automatique du lexique scientifique transdisciplinaire. *Revue française de linguistique appliqué*, *12*(2), 45-64. https://doi.org/10.3917/rfla.122.0045

Figueras Solanilla, Carolina (2016). Teaching multiword sequences in the native language. En Sergi Torner; Elisenda Bernal (Eds.), *Collocations and other lexical combinations in Spanish. Theoretical, lexicographical and applied perspectives* (pp. 287-302). Routledge.

Frankenberg-Garcia, Ana; Rees, Geraint; Lew, Robert; Roberts, Jonathan; Sharma, Nirwan; Butcher, Peter (2019). CollocAid: a tool to help academic English writers find the words they need. En Fanny Meunier; Julie Van de Vyver; Linda Bradley; Sylvie Thouësny (Eds.), *CALL and complexity–short papers from EUROCALL* (pp. 144-150). http://doi.org/10.14705/rpnet.2019.38.1000

Garcia, Marcos; Gamallo, Pablo (2016). Yet another suite of multilingual NLP tools. En José Paulo Leal; José Luís Sierra-Rodríguez; Alberto Simões (Eds.), *Languages, Applications and Technologies. Communications in Computer and Information Science* (pp. 65–75). Springer.

García-Salido, Marcos (2021). Compiling an Academic Vocabulary List of Spanish. ResearchGate Preprint. https://doi.org/10.13140/RG.2.2.27681.33123

García-Salido, Marcos; García-González, Marcos; Alonso-Ramos, Margarita (2019). Identifying lexical bundles for an academic writing assistant in Spanish. En Gloria Corpas Pastor; Ruslan Mitkov (Eds.), *Computational and Corpus-Based Phraseology. Third International Conference, Europhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings*, *volume 11755* of *Lecture Notes in Artificial Intelligence* (pp. 144-158). Springer. https://doi.org/10.1007/978-3-030-30135-4_11

Granger, Sylviane; Paquot, Magali (2015). Electronic lexicography goes local design and structures of a needs-driven online academic writing aid. *Lexicographica. International Annual Lexicography*, *31*(1), 118–141. https://doi.org/10.1515/lexi-2015-0007

Guzzi, Eleonora; Alonso-Ramos, Margarita (2022). Selección de colocaciones académicas en español a través de un filtro de interdisciplinariedad, *Procesamiento del Lenguaje Natural*, *69*,83-94. https://doi.org/10.26342/2022-69-7

Guzzi, Eleonora; Alonso-Ramos, Margarita (en prensa). Sofisticación y diversidad como medidas de complejidad léxica para determinar el perfil colocacional de textos académicos en español. *Revista Signos*, 56(112). http://revistasignos.cl/index.php/signos/issue/view/26

Guzzi, Eleonora (en preparación). *Identificación automática de colocaciones interdisciplinares para una herramienta en línea de ayuda a la redacción de textos académicos en español*. Tesis doctoral, Universidade da Coruña.

Hyland, Ken (2008). *As can be seen*: lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, Ken; Shaw, Philip (Eds.) (2016). *The Routledge Handbook of English for Academic Purposes*. Routledge.

Jacques, Marie-Paule; Tutin, Agnès (Eds.) (2018). *Lexique transversal et formes discursives des sciences humaines*. ISTE Editions. https://doi.org/10.4000/lidil.7531

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Kübler, Natalie; Pecman, Mojca (2012). The ARTES bilingual LSP dictionary: from collocation to higher order phraseology. En Sylvianne Granger; Magali Paquot (Eds.), *Electronic Lexicography* (pp. 187–210). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.003.0010

Laso, Natalia Judith (2022). SciE-Lex Report: Building up a Collocational Database to Assist the Production of Biomedical Texts in L2 English. *TEISEL, Tecnologías para la investigación en segundas lenguas*, *1*, 1-16. https://doi.org/10.1344/teisel.v1.37444

Lea, Diana; Bull, Victoria; Webb, Suzanne (Eds.) (2014). *OLDAE: Oxford Learner's Dictionary of Academic English*. Oxford University Press.

Manchón, Rosa M.; Matsuda, Paul Kei (2016). *Handbook of Second and Foreign Language Writing.* De Gruyter Mouton. https://doi.org/10.1515/9781614511335

Marín, Javier; López, Sonia; Roca-De-Larios, Julio (2015). El proceso de escritura académica en la universidad española: percepciones de estudiantes y profesores. *Cultura y Educación*, *27*(3), 504-533. http://dx.doi.org/10.1080/11356405.2015.1072360

Martín Zorraquino, María; Portolés Lázaro, José (1999). Los marcadores del discurso. En Ignacio Bosque; Violeta Demonte (Dirs.), *Gramática descriptiva de la lengua española* (capítulo 63). Espasa Calpe.

McCarthy, Michael; O'Dell, Felicity (2008). *Academic Vocabulary in Use: 50 Units of Academic Vocabulary Reference and Practice; Self-study and Classroom Use*. Cambridge University Press.

Mel'čuk, Igor (2012). Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology, 3*(1), 31–56. https://doi.org/10.1515/phras-2012-0003

Mel'čuk, Igor (2020). Clichés and pragmatemes. *Neophilologica, 32,* 9–20. https://journals.us.edu.pl/index.php/NEO/article/view/10841

Mel'čuk, Igor (2021). Morphemic and Syntactic Phrasemes. *Yearbook of Phraseology, 12*(1), 33–74. https://doi.org/10.1515/phras-2021-0004

Mel'čuk, Igor. (2015). Clichés, an understudied subclass of phrasemes. *Yearbook of Phraseology, 6*(1), 55–86. https://doi.org/10.1515/phras-2015-0005

Montolío Durán, Estrella (Dir.) (2014). *Manual de escritura académica y profesional. Vol 1: Estructuras gramaticales*. Ariel.

Montolío Durán, Estrella (Dir.) (2014). *Manual de escritura académica y profesional. Vol 2: Estrategias discursivas.* Ariel.

Mur Dueñas, Pilar (2012). Getting research published internationally in English: An ethnographic account of a team of Finance Spanish scholars' struggles. *Ibérica, Revista de la Asociación Europea de Lenguas para Fines Específicos*, *24*, 139-155. https://www.revistaiberica.org/index.php/iberica/article/view/299

Natale, Lucía; Stagnaro, Daniela (Orgs.). (2016). *Alfabetización académica: un camino para la inclusión en el nivel superior*. Universidad Nacional de General Sarmiento.

Navarro, Federico; Aparicio, Graciela (2018). *Manual de lectura, escritura y oralidad académicas para ingresantes en la universidad*. Universidad Nacional de Quilmes.

Nazar, Rogelio; Renau, Irene (2023). Estilector: un sistema de evaluación automática de la escritura académica en castellano. *Perspectiva Educacional*, *62*(2), 37-59. http://dx.doi.org/10.4151/07189729-Vol.62-Iss.2-Art.1427

Nivre, Joakim; de Marneffe, Marie-Catherine; Ginter, Filip; Goldberg, Yoav; Hajič, Jan; D. Manning, Christopher; McDonald, Ryan; Petrov, Slav; Pyysalo, Sampo; Silveira, Natalia; Tsarfaty, Reut; Zeman, Daniel (2016). Universal dependencies v1: A multilingual treebank collection. En *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 1659-1666). https://aclanthology.org/L16-1262

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Núñez Cortés, Juan Antonio (Coord.). (2015). *Escritura académica: de la teoría a la práctica*. Pirámide.

Padró, Lluis; Stanilovsky, Evgeny (2012). Freeling 3.0: Towards wider multilinguality. En *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (LREC'12) (pp. 2473–2479). European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf

Paquot, Magali (2010). *Academic vocabulary in learner writing: From extraction to analysis*. Bloomsbury Publishing. https://doi.org/10.1080/00437956.2019.1708590

Paquot, Magali (2012). The LEAD dictionary-cum-writing aid: an integrated dictionary and corpus tool. En Sylvianne Granger; Magali Paquot (Eds.), *Electronic Lexicography* (pp. 163–185). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.003.0009

Parodi, Giovanni (Ed.). (2010). *Academic and Professional Discourse Genres in Spanish*. John Benjamins. https://doi.org/10.1075/scl.40

Pastor Cesteros, Susana (2022). *Español académico como LE/L2: destrezas, competencias y movilidad universitaria*. Routledge. https://doi.org/10.4324/9780429455162

Perea Siller, Francisco Javier (Coord.). (2013). *Comunicar en la Universidad. Descripción y metodología de los géneros académicos*. Universidad de Córdoba.

Pérez-Llantada, Carmen (2014). Formulaic language in L1 and L2 expert academic writing: convergent and divergent usage. *Journal of English for Academic Purpose*, *14*, 84–94. https://doi.org/10.1016/j.jeap.2014.01.002

Ramírez Gelbes, Silvia (2013). *Cómo redactar un paper. La escritura de artículos científicos*. Noveduc.

Regueiro Rodríguez, María L.; Saéz, Daniel M. (2013). *El Español Académico. Guía Práctica Para La Elaboración de Textos Académicos*. Arco Libros.

Rodríguez, Catalina (2009). *Diccionario de conectores y operadores del español*. Arco Libros

Römer, Ute (2009). English in academia: Does nativeness matter? *Anglistik: International Journal of English Studies*, *20*, 89-100.

Salazar, Danica (2014). *Lexical Bundles in Native and Non-native scientific writing*. John Benjamins. https://doi.org/10.1075/scl.65

Santos Río, Luís (2003). *Diccionario de partículas*. Luso-Española de Ediciones.

Sanz Álava, Inmaculada (2007). *El Español Profesional y Académico en el aula universitaria. El discurso oral y escrito*. Tirant Lo Blanch.

Simpson-Vlach, Rita; Ellis, Nick C. (2010). An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics, 31*(4), 487–512. https://doi.org/10.1093/applin/amp058

Straka, Milan; Hajic, Jan; Strakov, Jana (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)*. European Language Resources Association, Portoroz, Slovenia. https://aclanthology.org/L16-1680

Swales, John M.; Feak, Christine, B. (2012). *Academic Writing for Graduate Students: Essential Tasks and Skills*. Michigan series in English for academic & professional purposes. University of Michigan Press. https://doi.org/10.3998/mpub.2173936

Tarp, Sven (2009). Reflections on Lexicographical User Research. *Lexikos*, *19*(1), 275–296. https://doi.org/10.5788/19-0-440

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023

ISSN 2696-676X

Tran, Thi Thu Hoai; Falaise, Achille (2018). Un dictionnaire basé sur corpus pour une aide à la rédaction universitaire. *Lidil, 58*, 1-19. https://doi.org/10.4000/lidil.5378

Vázquez, Graciela (Coord). (2001). *Guía didáctica del discurso académico escrito: ¿cómo se escribe una monografía?*. Edinumen.

Villayandre, Milka (2018). "HARTA" de noveles: un corpus de español académico. *CHIMERA: Revista de Corpus de Lenguas Romances y Estudios Lingüísticos*, *5*(1), 131-140. https://doi.org/10.15366/chimera2018.5.1.011

Zusman, Perla (2022). Las publicaciones científicas y la búsqueda por construir otra globalización académica, *GEOUSP*, *26*(2). https://www.revistas.usp.br/geousp/article/view/200517

Guzzi, Eleonora; Alonso Ramos, Margarita (2023). Descripción y usabilidad de HARTA, una herramienta de ayuda para la redacción de textos académicos en español. *TEISEL. Tecnologías para la investigación en segundas lenguas*, *2*, Artículo e-a_hr03, 1-22. http://doi.org/10.1344/teisel.v2.42173

 | Recepción: 20/02/2023 | Revisión: 17/03/2023| Aceptación: 24/02/2023| Publicación: 25/04/2023