Procesamiento del Lenguaje Natural, Revista nº 69, septiembre de 2022, pp. 95-103 recibido 31-03-2022 revisado 20-05-2022 aceptado 30-05-2022

Compilación del corpus académico de noveles en euskera HARTAeus y su explotación para el estudio de la fraseología académica

Compilation of the academic corpus of novels in Basque HARTAeus and its exploitation for the study of academic phraseology

María Jesús Aranzabe,[1] Antton Gurrutxaga,[2] Igone Zabala[1]
[1] Centro HiTZ-Ixa, Universidad del País Vasco (UPV/EHU)
[2] Fundación Elhuyar
{maxux.aranzabe,igone.zabala}@ehu.eus, a.gurrutxaga@elhuyar.eus

Resumen: Se ha compilado un corpus académico de noveles para el euskera comparable con el corpus HARTA-noveles para el español. A partir del corpus se ha extraído una lista de vocabulario académico para el euskera, y sendas listas de colocaciones y fórmulas, a las que se les han asignado funciones discursivas. El objetivo último del proyecto HARTAes-vas, en el que se enmarca este trabajo, es diseñar una herramienta de ayuda a la escritura académica para las dos lenguas centrada en las combinaciones léxicas académicas, que integre diccionario y corpus.
Palabras clave: corpus académico, colocaciones, *lexical bundles*, funciones discursivas.

Abstract: An academic corpus of novices was compiled for Basque, comparable to the corpus HARTA-noveles for Spanish. A list of academic Basque vocabulary, collocations and formulas were extracted from the corpus, and then they were assigned discursive functions. The ultimate objective of the HARTAes-vas project, in which this work is framed, is to design a tool to help academic writing for Basque and Spanish focused on academic lexical combinations, integrating lexicographic information and corpora.
Keywords: academic corpus, collocations, lexical bundles, discursive functions.

1 Introducción

La introducción del euskera en ámbitos académicos, incluidos los de la educación superior, que se produjo a principios de la década de 1980, ha sido crucial para su revitalización (Zabala, 2019), ya que ha contribuido de forma muy significativa al aumento del número de hablantes y al desarrollo de los recursos expresivos necesarios para la comunicación especializada. Sin embargo, ¿podemos decir que el euskera ha "conquistado" los dominios académicos en el sentido de Laurén et al. (2002)? Dicho en otras palabras, ¿el euskera ha desarrollado los recursos expresivos necesarios para la comunicación académica en los diferentes ámbitos de especialidad? Laurén et al. (2002) defienden que a esta pregunta se puede responder de forma individual o de forma colectiva.

A nivel individual, los estudiantes universitarios adquieren los registros académicos necesarios para convertirse en miembros de la comunidad de expertos de su área gracias a numerosas tareas en las que el lenguaje resulta crucial (Biber, 2006). Algunos autores defienden que los textos académicos se elaboran siguiendo esquemas discursivos prefabricados que utilizan unidades fraseológicas semiautomáticas (Paquot, 2018), a las que nos referiremos de forma general como combinaciones léxicas académicas (CLA). No es de extrañar, por tanto, que las CLA del inglés hayan sido el objeto de estudio de numerosas investigaciones de lingüística de corpus con fines aplicados. Este auge es fácilmente explicable teniendo en cuenta el rol predominante del inglés como lengua académica internacional y el gran número de hablantes, principalmente, hablantes no-nativos, que necesitan recursos de ayuda para la

ISSN 1135-5948. DOI 10.26342/2022-69-8

redacción de artículos científicos y de trabajos académicos en general. Posteriormente, también se han ido extendiendo los proyectos de compilación de corpus académicos y de estudio de las CLA a otras lenguas como el francés, portugués de Brasil, sueco, noruego, danés y español (Alonso et al., 2017), ya que numerosos trabajos académicos se producen en las lenguas locales.

Si bien podemos pensar que el ser hablante nativo de una lengua es un factor que puede facilitar la producción de textos en dicha lengua, también está generalmente aceptado que no hay hablantes nativos de los registros académicos, y que éstos se adquieren gracias a la experiencia lingüística de lectura y producción de textos académicos. Debido a la internacionalización de la comunicación académica y del uso cada vez más extendido del inglés como lengua de instrucción y de elaboración de trabajos académicos en la educación superior, existe la preocupación de que los estudiantes universitarios, e incluso los expertos tengan cada vez más dificultades para adquirir los recursos expresivos necesarios para la comunicación académica en L1 diferentes del inglés (Swales, 2000; Görlach, 2002; Laurén et al., 2002; Johansson Kokkinakis et al., 2012; Gotti, 2012). Es por esto que se hace necesario elaborar recursos y herramientas de ayuda a la escritura académica también para otras lenguas. En el caso del euskera, la idea generalizada es que no ha habido suficiente tiempo para el desarrollo y estabilización de los registros académicos (Zabala et al., 2011; Zabala et al., 2021), y la preocupación por el impacto de la creciente internacionalización es aún mayor.

Las CLA son segmentos de palabras recurrentes que pueden o no ser semánticamente composicionales y que cubren funciones retóricas como añadir información, presentar ejemplos o expresar posibilidad. Incluyen colocaciones (*ondorioak atera* "extraer conclusiones"), locuciones (*oro har* "en general") y fórmulas, que coinciden en gran medida con las denominadas en la literatura *lexical bundles* (*azpimarratu beharra dago* "hay que remarcar"). La recurrencia es el resultado de su uso frecuente en discursos compartidos por la comunidad académica y, por lo tanto, las CLA constituyen un tipo de unidades privilegiadas para el estudio del nivel de desarrollo de los registros académicos del euskera.

Este trabajo se enmarca en el proyecto HARTAes-vas, proyecto coordinado entre la Universidade da Coruña y la Universidad del País Vasco (UPV/EHU) y financiado por el Ministerio de Ciencia e Investigación. El equipo de la Coruña cuenta con un corpus de expertos y otro de noveles en español, compilados en un proyecto anterior, ha explotado dichos corpus para la extracción y clasificación de CLA y ha desarrollado una herramienta de consulta de la fraseología académica en español (Herramienta de Ayuda a la Escritura Académica: HARTA)[1] (García-Salido et al., 2018). La colaboración con el grupo de la Coruña es fundamental para poder contar con elementos de comparación entre dos lenguas que se diferencian por su tipología y por su situación sociolingüística. En este trabajo describimos el corpus académico de noveles para el euskera compilado dentro del proyecto HARTAvas y su explotación para el estudio de la fraseología académica de cara a crear una herramienta de consulta coordinada con HARTAes, que ayude a la escritura académica en euskera, y que contribuya al desarrollo y estabilización de los registros académicos en dicha lengua. Hemos elaborado un corpus comparable con el corpus HARTA de noveles para el español (Villayandre, 2018; García-Salido et al., 2018) con el fin de poder contrastar los resultados obtenidos para el euskera, que es una lengua aglutinante en proceso de normalización, con los obtenidos para el español, lengua flexiva y bien desarrollada.

En el apartado 2 describimos la constitución del corpus de noveles HARTAeus. El apartado 3 lo dedicamos a la extracción, validación y análisis de las fórmulas y colocaciones académicas a partir del corpus. Finalmente, los apartados 4 y 5 recogen los resultados y conclusiones obtenidos hasta el momento.

2 Constitución del corpus HARTAeus

El corpus HARTAeus de noveles para el vasco está constituido por Trabajos Fin de Grado (TFG) y Trabajos Fin de Máster (TFM), por lo que se puede considerar una muestra de la escritura académica de los estudiantes universitarios. Al ser un corpus comparable con el corpus HARTA-noveles del español, su diseño ha seguido los criterios definidos en la

[1] http://www.dicesp.com:8083/search

creación de éste último (Villayandre, 2018). De este modo, los textos del corpus HARTAeus están divididos en cuatro secciones (Arte y Humanidades, Biología y Ciencias de la Salud, Ciencias Físicas y Ciencias Sociales), que se dividen a su vez en distintos dominios temáticos como, por ejemplo, Biología y Medicina en la sección de Biología y Ciencias de la Salud (ver Anexo 1 para la división del corpus en secciones y dominios).

El proceso de compilación ha comprendido seis fases: i) recolección de documentos para el corpus: los documentos en formato PDF proceden en su mayoría del repositorio ADDI (Archivo Digital para la Docencia e Investigación) de la Universidad del País Vasco/Euskal Herriko Unibertsitatea (UPV/EHU); ii) normalización: se ha realizado la conversión al formato DOCX de los documentos originales con el fin de realizar la limpieza y ordenación de las secciones y párrafos de los textos, y eliminar las marcas sobrantes; iii) codificación: se han introducido las etiquetas que marcan el inicio y final de las distintas secciones de los textos (título, resumen, presentación, introducción, cuerpo, metodología, resultados y discusión, conclusiones, agradecimientos, notas al pie de página y anexos); iv) se han incorporado de manera automática los textos en el entorno de trabajo Garaterm (Zabala et al, 2013) para su posterior procesamiento; v) almacenamiento: se han anotado los metadatos y las secciones de los textos del corpus de referencia con etiquetas XML. Asimismo, se ha adaptado el conversor de TEI existente en la plataforma Garaterm con el fin de mantener la estructura y las etiquetas de XML utilizadas para marcar los apartados de los documentos originales, y vi) procesamiento: el corpus ha sido tokenizado, lematizado y analizado morfológicamente por medio de Eustagger (Alegría et al., 2002), analizador morfológico y etiquetador de partes del discurso para el euskera. Por medio de este proceso se obtiene la información del lema y los rasgos morfosintácticos necesarios para poder extraer las combinaciones de palabras candidatas a colocaciones: N+N, N+V, N+Adj.

El resultado de este proceso ha sido la creación de un corpus académico monolingüe integrado por 398 textos (71 % TFG y 29 % TFM) y 3.285.098 palabras distribuidas en cuatro áreas de conocimiento (Tabla 1). La distribución en los distintos dominios temáticos puede verse en el Anexo 1.

Secciones del corpus	TFG nº de palabras (nº de documentos)	TFM nº de palabras (nº de documentos)	Total de palabras y documentos
Arte y Humanidades	450.859 (62)	203.831 (12)	654.690 (74)
Biología y Ciencias de la Salud	271.932 (65)	153.533 (26)	425.465 (91)
Ciencias Físicas	1.035.599 (121)	378.685 (36)	1.414.284 (157)
Ciencias Sociales	559.791 (46)	230.868 (30)	790.659 (76)
Totales	2.318.181 (294)	966.917 (104)	3.285.098 (398)

Tabla 1: Distribución de palabras y documentos por secciones y por tipología de textos (TGF y TFM).

Como se puede observar en la Tabla 1, el número de TFM es menor que el de TFG. Esto se debe a que el número de TFM que se elaboran en euskera es muy pequeño, a que bastantes trabajos están protegidos por cláusulas de confidencialidad y a que muchos de ellos no se publican en la plataforma ADDI.

3 Extracción y validación de CLA

Para la extracción y validación de las CLA hemos añadido tres módulos al extractor de terminología para el euskera Erauzterm (Alegría et al., 2004): un módulo para la identificación del vocabulario académico, un segundo módulo para la identificación de colocaciones académicas y un tercer módulo para la identificación de fórmulas académicas. Debido a las deficiencias del desarrollo de los registros académicos en euskera, encontramos CLA que superan los umbrales de frecuencia y dispersión establecidos pero que pueden ser consideradas como incorrectas o no óptimas. Como quiera que el último objetivo del proyecto es desarrollar una herramienta de ayuda a la escritura, en el proceso de validación hemos ido identificando las CLA incorrectas y elaborando una tipología de estas.

3.1 Extracción y validación del vocabulario académico

El módulo para la elaboración de la lista de vocabulario académico utiliza como contraste el corpus Dabilena,[2] obtenido de la web

[2] https://dabilena.elhuyar.eus/

(300.217.903 palabras): es el corpus mayor que tenemos para el euskera y ha sido elaborado por Elhuyar. Los candidatos se pueden filtrar según su categoría gramatical (N, V, Adj., Adv…) y según varias medidas de frecuencia y dispersión: nº de dominios y partes del texto, porcentaje de textos del corpus en los que aparecen, así como *log-likelihood*, frecuencia y umbrales de frecuencia esperada.

La tarea de identificación del vocabulario académico no es trivial, ya que se trata de identificar los lemas característicos del discurso académico, pero descartando los términos específicos de una determinada área de especialidad. Para que la lista obtenida para el euskera sea comparable con la obtenida para el español en el proyecto HARTA, se han probado algunas de las medidas descritas en García-Salido (2021). Se han realizado dos experimentos con 3 condiciones comunes: presencia de los candidatos en las 4 secciones del corpus y en el 20 % de los documentos, y valores de *log-likelihood* positivos. En el segundo experimento se ha añadido la condición de que la frecuencia no sea 3 veces superior a la frecuencia esperada en cada uno de las cuatro secciones del corpus. Los resultados obtenidos se resumen en las Tablas 2 y 3.

Experimento 1			
	candidatos	validados	precisión
N	443	338	76,30 %
V	167	165	98,80 %
ADJ	147	128	87,07 %
ADV	73	53	72,60 %
Total	830	684	82,41 %

Tabla 2: Validación de los candidatos para la lista de vocabulario académico: presencia en las 4 secciones del corpus y en el 20 % de los documentos + *log-likelihood* positiva.

Experimento 2			
	candidatos	validados	precisión
N	160	116	72,50 %
V	81	81	100 %
Adj.	70	62	88,57 %
Adv.	49	34	68,39 %
Total	360	293	81,39 %

Tabla 3: Validación de los candidatos para la lista de vocabulario académico: presencia en las 4 secciones del corpus y en el 20 % de los documentos + *log-likelihood* positiva + F< 3 $F_{esperada}$ en cada sección.

Como se puede ver en las Tablas 2 y 3, la condición añadida en el experimento 2, encaminada a descartar los términos específicos de las diferentes áreas de especialidad, no aumenta la precisión y, además, disminuye la cobertura en un 57 %.

3.2 Extracción y validación de colocaciones académicas

El módulo de extracción y validación de colocaciones académicas está conectado con el vocabulario académico, de tal manera que permite filtrar los candidatos a colocaciones con un solo lema o con los dos lemas incluidos en la lista de vocabulario académico. Las combinaciones candidatas a colocaciones académicas se extraen utilizando las medidas de asociación desarrolladas en Gurrutxaga et. al. (2011, 2018) y atienden a los siguientes patrones sintácticos: Sujeto-Verbo (*emaitzek erakutsi* "resultados mostrar"), Verbo-Objeto (*helburu lortu* "objetivo conseguir = conseguir objetivos"; *emaitzei erreparatu* "resultados-DATIVO atender = atender a los resultados", *lanetik atera* "trabajo-ABLATIVO = obtener a partir del trabajo"), Nombre-Modificador (*helburu nagusi* "objetivo principal", *funtsezko elementu* "fundamental elemento = elemento fundamental"), N-(posposición)-N (*lagin tamaina* "muestra tamaño = tamaño de muestra"; *laginaren tamaina* "muestra de la tamaño = tamaño de la muestra").

Se han excluido los nombres ligeros *mota* "tipo", *kopuru* "número", *falta* "falta", *multzo* "conjunto", *zati* "parte" y *maila* "nivel". También se han descartado los adjetivos *bakar* "único", *berdin* "igual", *ezberdin/desberdin* "diferente" y los modificadores prenominales *goiko* "superior", *beheko* "inferior", *honako*

"este", *horrelako* "similar", *hurrengo* "siguiente". La razón de este descarte es que, a pesar de que dan lugar a combinaciones recurrentes con nombres académicos, en la mayoría de los casos no se trata de colocaciones.

En la Tabla 4 se resumen algunos de los resultados obtenidos. El número total de tokens normalizado es de 9.471 tokens por millón de palabras.

	Types	Tokens
N-Modif.	495	13.221
N(pos.)N	142	4.112
Sujeto-V	11	192
V-Objeto	357	13.589
Total	1.005	31.114

Tabla 4: Colocaciones extraídas del corpus HARTAeus.

3.3 Extracción y validación de fórmulas académicas

Para la detección de fórmulas se han extraído n-gramas entre 2 y 5 elementos. Se han filtrado únicamente los que estaban presentes en las 4 secciones del corpus y cuya frecuencia era igual o superior a 10 apariciones por millón de palabras, criterio generalmente utilizado para la identificación de *lexical bundles* (Biber et al., 1999). A la hora de validar los candidatos, hemos asignado una o más funciones discursivas a cada n-grama validado, siguiendo la tipología usada en el proyecto HARTA para el español (García-Salido et al., 2019), ya que, con el fin de que los usuarios puedan encontrar las fórmulas fácilmente en la herramienta de consulta, es más detallada que la de Biber et al. (2004) y la de Hyland (2008). Además, una de las tareas principales del proyecto consiste en comparar las fórmulas extraídas de los corpus de noveles vasco y español.

En el proceso de validación y de asignación de función discursiva a los n-gramas, en algunos casos hemos eliminado algún elemento que no aportaba valor semántico a la función discursiva, como es la conjunción *eta* "y". Así por, ejemplo, si un candidato validado era *eta hala ere* "y aun así", lo hemos eliminado y hemos mantenido únicamente la fórmula de dos elementos *hala ere* "aun así". Con este procedimiento, hemos identificado algunas fórmulas monoléxicas plurimorfémicas, que en principio no esperábamos recoger. Por ejemplo, el n-grama *eta ondorioz* "y por consiguiente" lo hemos validado como la fórmula *ondorioz* "por consiguiente" y le hemos asignado la función "expresar consecuencia".

Una vez validados los n-gramas y asignadas las funciones discursivas, hemos identificado las variantes de una misma fórmula. Por ejemplo, *aipatu den moduan* "como se ha mencionado" y *aipatu dugun moduan* "como hemos mencionado" son dos variantes de la misma fórmula, y lo mismo sucede con las fórmulas *horrek esan nahi du* "eso quiere decir" y *horrek ez du esan nahi* "eso no quiere decir". En estos casos, las variantes las hemos considerado como un solo *type*. Se han validado y clasificado 644 fórmulas (*types*), 1.028 variantes y 125.398 tokens (38.171 tokens por millón de palabras). Como puede verse en la Tabla 5, a falta de estrategias complementarias para la extracción de fórmulas monoléxicas, las fórmulas de 2 palabras son las más numerosas.

	Número de palabras	Types
Fórmulas académicas	1 palabra	42
	2 palabras	490
	3 palabras	126
	4 palabras	12
	5 palabras	4
	Totales	644

Tabla 5: Nº de fórmulas académicas validadas una vez analizada la variación.

3.4 Identificación y clasificación de CLA incorrectas

Algunas colocaciones y fórmulas que llegan a los umbrales de frecuencia y dispersión establecidos, pueden considerarse como incorrectas o no óptimas. Estas CLA las hemos recogido y clasificado para poder así tenerlas en cuenta a la hora de diseñar la herramienta de consulta ya que, aunque no son muy numerosas, presentan un importante grado de recurrencia y compiten con formas más correctas o genuinas. En la Tabla 6, ofrecemos una clasificación preliminar y algunos ejemplos.

Ortografia no-estándar	
kontutan hartu behar da "hay que tener en cuenta"	*kontuan hartu behar da*
gutxi gora behera "poco más o menos"	*gutxi gorabehera*
pausuak eman "dar pasos" pausu "descanso"	*pausoak eman* pauso "paso"
Demostrativo de 1er grado como anáfora	
honek ez du esan nahi "esto no quiere decir" Demostrativo de 1er grado (catáfora)	*horrek ez du esan nahi* Demostrativo de 2º grado (anáfora)
Orden de palabras inadecuado	
Lan honen helburua […] *da* "El objetivo de este trabajo […] es"	*Lan honen helburua da* […]
Forma incorrecta de un conector	
Alde batetik […] *eta beste aldetik,* […] "Por un lado […] y por el otro lado […]"	*Alde batetik* […] *eta, bestetik,* […] "Por un lado […] y por el otro […]"
Asignación de una función discursiva incorrecta	
Hau da "esto es" "expresar causa"	*Hau da* "reformular"
Colocación incorrecta desde el punto de vista semántico-sintáctico	
datuek adierazi "datos expresar"	*datuek erakutsi* "datos mostrar"
datu adierazgarri "dato representativo"	*datu esangarri* "dato significativo"
Calcos	
besteen artean "entre otros"	*besteak beste* "entre otros"

Tabla 6: Clasificación y ejemplos de CLA incorrectas.

4 Resultados y discusión

No contamos con un corpus de expertos para el euskera comparable con el corpus HARTA de expertos, por lo que, para analizar los datos obtenidos hasta el momento, los contrastaremos con los ofrecidos en García-Salido (2021) y en Alonso y Zabala (2022).

La lista de vocabulario académico compilada hasta el momento para el euskera cuenta con 684 lemas. Esta lista es bastante más reducida que la obtenida tras contrastar los resultados de diferentes técnicas para el español (García-Salido, 2021): 833 lemas. La diferencia puede estar motivada por el descarte de los lemas con valores de *log-likelihood* negativos, ya que entre los lemas descartados puede haber algunos que son muy utilizados en textos generales pero que activan significados específicos en los discursos académicos. Nuestra idea es seguir completando la lista obtenida hasta el momento, probando otras técnicas y medidas.

Sin embargo, la principal aplicación de la lista de lemas académicos en el proyecto HARTAvas es la extracción de colocaciones académicas. Para ello, es fundamental la lista de nombres, ya que son los que constituyen las bases de las colocaciones, el número de N obtenidos para el euskera es menor pero comparable al del español: 338 eus / 358 es.

El número de colocaciones académicas obtenidas es también comparable al obtenido a partir del corpus HARTA de noveles para el español (Alonso y Zabala, 2022): 1.005 types eus / 1.197 es. Aunque hay que tener en cuenta que el corpus del euskera es mayor que el del español, que cuenta con 2 M de palabras. Aun siendo menor el número de colocaciones extraídas para el euskera, el número de tokens por M de palabras es notablemente superior: 9.471 tokens/M eus / 6.897 tokens/M es. Por lo tanto, como primera aproximación se puede decir que las colocaciones detectadas para el euskera son más recurrentes que las detectadas para el español en el corpus de noveles.

El número de fórmulas (types) extraídas para el euskera es notablemente superior al obtenido a partir del corpus de noveles en español: 644 types eus / 472 types es. También existe diferencia en la frecuencia de dichas fórmulas: 38.171 tokens/M eus / 20.474 tokens/M es. El mayor número de fórmulas detectadas en euskera podría estar relacionado con el menor grado de fijación de las fórmulas académicas en esta lengua, y con la variación entre fórmulas más genuinas y fórmulas calcadas del español: *besteen artean* (calco) / *besteak beste*; *orokorrean* (calco) / *oro har.* El número mayor de tokens, podría indicar una menor riqueza expresiva de los noveles vascos, que les haría recurrir más frecuentemente a las mismas secuencias dando lugar a un discurso más repetitivo. De cualquier modo, se requiere un análisis más minucioso de los criterios de validación que hemos utilizado para una y otra lengua, así como de los datos, con el fin de poder hacer una comparación más precisa de las CLA obtenidas en una y otra lengua de cara al diseño de la herramienta de ayuda a la escritura académica para las dos lenguas.

5 Conclusiones

A pesar de que el euskera es una lengua minorizada que se introdujo hace solo unas décadas en la enseñanza superior, hemos logrado compilar un corpus de trabajos académicos en euskera (TFG y TFM) comparable, e incluso algo más extenso, que el corpus de noveles para el español.

A partir del corpus hemos obtenido una lista de vocabulario académico en euskera (644 lemas), que aunque debe de considerarse preliminar, nos ha permitido identificar un gran número de colocaciones académicas (1.005 types).

Hemos extraído también n-gramas de 2, 3, 4 y 5 elementos, que hemos validado y a los que hemos asignado funciones discursivas partiendo de la tipología utilizada para HARTAes. Los n-gramas nos han permitido detectar fórmulas poliléxicas o *lexical bundles* como (*kontuan hartu beharrekoa da* "hay que tener en cuenta"), pero en el proceso de validación, hemos podido detectar también 42 fórmulas monoléxicas o *morphemic bundles* como *laburbilduz* "en resumen". Así hemos obtenido 644 fórmulas y 1.028 variantes, a las que les hemos asignado funciones discursivas.

Por último, hemos desarrollado una tipología de fórmulas incorrectas, de cara a elaborar su tratamiento lexicográfico en la herramienta de consulta.

Estamos implementando técnicas de semántica distribucional, con el fin de utilizar los corpus comparables del español y del euskera para la detección de fórmulas, sobre todo fórmulas monoléxicas, y equivalentes de colocaciones y fórmulas entre las dos lenguas.

Además, hemos comenzado la tarea de comparación más minuciosa de las listas de CLA elaboradas para las dos lenguas, con el fin de obtener una clasificación que tenga en cuenta las características tipológicas del español y del euskera. Dicha comparación nos servirá también para decidir el diseño de la herramienta de consulta para ambas lenguas.

Agradecimientos

Este trabajo es parte del proyecto HARTAvas (PID2019-109683GB-C22), financiado por el Ministerio de Ciencia e Innovación.

Bibliografía

Alegria, I., M.J. Aranzabe, A. Ezeiza A., N. Ezeiza, y R. Urizar R. 2002. Robustness and customisation in an analyser/lemmatiser for Basque. *En Third International Conference on Language Resources and Evaluation (LREC): Customizing Knowledge in NLP Applications-Strategies, Issues and Evaluation Workshop*, páginas 1-6, Las Palmas de Gran Canaria (Spain).

Alegría, I, A. Gurrutxaga, P. Lizaso, X. Saralegi, S. Ugartetxea, y R. Urizar. 2004. An Xml-Based Term Extraction Tool for Basque. En *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, páginas 1733-1736, Lisboa (Portugal).

Alonso-Ramos, M., M. García-Salido, y M. Garcia. 2017. Exploiting a Corpus to Compile a Lexical Resource for Academic Writing: Spanish Lexical Combinations. En Kosem, I., J. Kallas, C. Tiberius, S. Krek, M. Jakubíček, V. Baisa (Eds). *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference*, páginas 571-586, Leiden (the Netherlands).

Alonso-Ramos, M. y I. Zabala. 2022. HARTAes-vas: Combinaciones léxicas para una Herramienta de ayuda a la redacción de textos académicos en español y en vasco. En *Proceedings of the Annual Conference of the Spanish Association for Natural Language Processing: Projects and Demonstrations, SEPLN*, September, A Coruña (Spain).

Biber, D. 2006. *University Language. A corpus-based study of spoken and written registers*. John Benjamins, Amsterdam.

Biber, D., E. Finegan, S. Johanson, S. Conrad, y G. Leech. 1999. *Longman Grammar of Spoken and Written Englis*h. Longman, London.

Biber, D., S. Conrad, y C. Viviana. 2004. If you look at…: lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3):371-405.

García-Salido, M., M. García, M., Villayandre, y M. Alonso-Ramos. 2018. A Lexical Tool for Academic Writing in Spanish based on Expert and Novice Corpora. En Calzolari N. et al. (Eds). *Proceedings of the Eleventh International Conference on Language*

Resources and Evaluation (LREC 2018), páginas 260-265, Miyazaki (Japan).

García-Salido, M., M. Garcia, y M. Alonso-Ramos. 2019. Identifying lexical bundles for an academic writing assistant in Spanish. En Corpas Pastor, G. y R. Mitkov (Eds). *Computational and Corpus-Based Phraseology*. Volume 11755 of *Lecture Notes in Artificial Intelligence*, páginas 144-158, Springer, Berlin.

García-Salido, M. 2021. Compiling an Academic Vocabulary List of Spanish. DOI: 10.13140/RG.2.2.27681.33123

Görlach, M. 2002. *Still More Englishes*. John Benjamins, Amsterdam.

Gotti, M. 2012. Variation in Academic Texts. En Gotti, M. (Ed). *Academic Identity Traits. A Corpus Based Investigation*, páginas 21-42, Peter Lang (Switzerland).

Gurrutxaga, A. e I. Alegria. 2011. Automatic extraction of NV expressions in Basque: basic issues on cooccurrence techniques. En *Proceedings of the Workshop on Multiword Expressions: from parsing and generation to the real world*, páginas 2–7, Portland, Oregon (USA).

Gurrutxaga, A., I. Alegria, y X. Artola. 2018. Caracterización computacional de la idiomaticidad: aplicación a la combinación nombre+verbo en euskera. En Ruiz Miyares, L. (Ed). *Estudios de Lexicología y Lexicografía. Homenaje a Eloína Miyares Bermúdez.* Santiago de Cuba (Cuba).

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

Johansson Kokkinakis, S., E. Sköldberg, B. Henriksen, K. Kinn, y J. Bondi Johannessen. 2012. Developing Academic Word Lists for Swedish, Norwegian and Danish a Joint Research Project. En Fjeld, R.V. y J.M. Torjusen (Eds). *Proceedings of the 15th EURALEX International Congress,* páginas 563–569, University of Oslo (Norway).

Laurén, Ch., J. Myking, y H. Picht. 2002. Language and domains: a proposal for a domain dynamics taxonomy. *LSP and Professional Communication*, 2(2):23-30.

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

Swales, J. 2000. Language for Specific Purposes. *Annual review of Applied Linguistics,* 20:59-76.

Villayandre, M. 2018. "HARTA" de noveles: un corpus de español académico. *CHIMERA: Revista De Corpus De Lenguas Romances Y Estudios Lingüísticos*, 5(1): 131–140.

Zabala, I., I. San Martin, M. Lersundi, y A. Elordui. 2011. Graduate teaching of specialized registers in a language in the normalization process: Towards a comprehensive and interdisciplinary treatment of academic Basque. En Maruenda-Bataller, S. y B. Clavel-Arroita (Eds). *Multiple voices in academic and professional discourse*, páginas 208-218, Cambridge Scholars (Newcastle upon Tyne, UK).

Zabala, I., M. Lersundi, I. Leturia, I. Manterola, y G. Santander. 2013. GARATERM: euskararen erregistro akademikoen garapenaren ikerketarako lan-ingurunea. En Alberdi, X. y P. Salaburu (Eds). *Ugarteburu terminologia jardunaldiak (V). Terminologia naturala eta terminologia planifikatua euskararen normalizazioari begira*, páginas 98-114, Servicio Editorial de la UPV/EHU (Bilbao).

Zabala, I. 2019. The elaboration of Basque in Academic and Professional Domains. En Grenoble, L., P. Lane, y U. Røyneland (Editor-in-Chief), Igartua, I. y L. Oñederra (Basque Eds). *Linguistic Minorities in Europe Online*, De Gruyter Mouton.

Zabala, I., M.J. Aranzabe, y I. Aldezabal. 2021. Retos actuales del desarrollo y aprendizaje de los registros académicos orales y escritos del euskera. *Círculo de Lingüística Aplicada a la Comunicació*n, 88:31-50.

A Anexo 1: Corpus HARTAeus

Distribución de palabras y documentos (TFG y TFM) por secciones del corpus y dominios temáticos.

Secciones del corpus	Dominios temáticos	TFG nº de palabras (nº de documentos)	TFM nº de palabras (nº de documentos)	Total de palabras y documentos
Arte y Humanidades	Arte	42.956 (6)	0	42.956 (6)
	Lingüística	207.443 (27)	203.831 (12)	411.274 (39)
	Literatura	90.139 (12)	0	90.139 (12)
	Historia y Cultura	110.321 (17)	0	110.321 (17)
	Biblioteconomía y documentación	0	0	0
	Totales	450.859 (62)	203.831 (12)	654.690 (74
Biología y Ciencias de la Salud	Biología	182.906 (44)	153.533 (26)	336.439 (70)
	Medicina	89.026 (21)	0	89.026 (21)
	Totales	271.932 (65)	153.533 (26)	425.465 (91)
Ciencias Físicas	Ciencias de la Tierra	52.263 (7)	0	52.263 (7)
	Física	113.027 (16)	0	113.027 (16)
	Ingeniería	656.156 (69)	20.031 (1)	676.187 (70)
	Informática	75.160 (8)	332.982 (32)	408.142 (40)
	Química	138.993 (21)	25.672 (3)	164.665 (24)
	Totales	1.035.599 (121)	378.685 (36)	1.414.284 (157)
Ciencias Sociales	Economía y Empresa	158.433 (10)	0	158.433 (10)
	Educación	102.335 (16)	230.868 (30)	333.203 (46)
	Sociología	233.632 (16)	0	233.632 (16)
	Derecho	65.391 (4)	0	65.391 (4)
	Totales	559.791 (46)	230.868 (30)	790.659 (76)