MLS - EDUCATIONAL RESEARCH

http://mlsjournals.com/Educational-Research-Journal

ISSN: 2603-5820

Como citar este artículo:
Umaña Corrales, O.C. (2023). Análisis bibliométrico del enfoque de la lingüística de corpus en estudios de terminología y lexicología en la categoría linguistics de Web of Science. *MLS-Educational Research, 7*(2), 134-151. 10.29314/mlser.v7i2.1598.

ANÁLISIS BIBLIOMÉTRICO DEL ENFOQUE DE LA LINGÜÍSTICA DE CORPUS EN ESTUDIOS DE TERMINOLOGÍA Y LEXICOLOGÍA EN LA CATEGORÍA *LINGUISTICS* DE WEB OF SCIENCE

Olga Clemencia Umaña Corrales
Centro de Automatización Industrial (SENA) / ESAP
olga.umana.c@gmail.com · http://orcid.org/0000-0003-2596-7798

Resumen. La lingüística de corpus permite analizar, describir y develar el funcionamiento de la lengua, así como reorientar su estudio a partir de la exploración de su uso real. El objetivo de este artículo es presentar un estudio bibliométrico descriptivo para identificar las tendencias sobre la implementación de la lingüística de corpus en las publicaciones más relevantes en Terminología y Lexicología en la categoría Linguistics de la base de datos Web of Science (WoS) entre 2012 y 2021. Se utilizan elementos bibliométricos y técnicas de minería de texto para identificar los autores más relevantes, las instituciones con más publicaciones y las revistas más productivas. También se describen los indicadores de productividad, colaboración y liderazgo científico y se grafican mediante la herramienta VOSviewer. Los resultados muestran que se ha presentado un aumento exponencial en la productividad de las investigaciones basadas en lingüística de corpus en esta última década y que España, con la Universidad de Granada, y Bélgica, con la Universidad Ghent, lideran dicha productividad. También se determinó que los autores más relevantes son Hoste, Lefever, Rigouts Terryn, Faber, Rojas y Tercedor-Sánchez y que la revista independiente Terminology encabeza el número de publicaciones en el área. Adicionalmente, al identificar los estudios más actuales mediante la herramienta Tree of Science (ToS), se estableció que la extracción automática de términos, la metodología de corpus, la semántica de marcos y el trabajo de lexicología y terminología en ámbitos de especialidad son algunas de las áreas con mayor perspectiva de investigación.
Palabras clave: Lingüística de corpus, Terminología, Lexicología, Análisis bibliométrico.

BIBLIOMETRIC ANALYSIS OF THE CORPUS LINGUISTICS APPROACH TERMINOLOGY AND LEXICOLOGY STUDIES IN THE WEB OF SCIENCE CATEGORY *LINGUISTICS*

Abstract. Corpus linguistics makes it possible to analyze, describe and unveil the functioning of language, as well as to reorient its study based on the exploration of its actual use. The aim of this article is to present a descriptive bibliometric study to identify trends in the implementation of corpus linguistics in the most relevant publications in Terminology and Lexicology in the Linguistics category of the Web of Science (WoS) database between 2012 and 2021. Bibliometric elements and text-mining techniques are used to account for the most relevant authors, the institutions with the most publications and the most productive journals. Indicators of productivity, collaboration and scientific leadership are also described and plotted using the VOSviewer tool. The results show that there has been an exponential increase in the productivity of research based on corpus linguistics in the last decade and that Spain, with the University of Granada, and Belgium, with Ghent University, lead this productivity. It was also possible to determine that the most relevant authors are Hoste, Lefever, Rigouts Terryn, Faber, Rojas and Tercedor-

Sánchez, and that the independent journal Terminology leads the number of publications in the area. In addition, through the Tree of Science (ToS) tool, it was possible to determine that automatic term extraction, corpus methodology, frame semantics, and lexicology and terminology work in specialized fields are the areas with the greatest research prospects.

Keywords: Corpus linguistics, Terminology, Lexicology, Bibliometric analysis.

Introducción

La lingüística de corpus consiste en una serie de procedimientos y métodos implementados para estudiar el uso real de la lengua a partir de textos compilados y mediante tecnologías informáticas. La importancia del desarrollo de la lingüística de corpus radica en que esta tiene el potencial de reorientar algunas teorías de la lengua, facilitar la elucidación de sus rasgos, y hacer una descripción más detallada de su estructura, sus funciones y sus repertorios léxicos, entre otros. Los estudios basados en corpus utilizan datos derivados de los corpus con el fin de explorar teorías o hipótesis, especialmente aquellas ya establecidas en la literatura actual, con el propósito de validarlas, refutarlas o refinarlas (McEnery and Hardie, 2011).

Disciplinas como la Terminología y la Lexicología han estado muy relacionadas con el incremento del uso del enfoque de la lingüística de corpus. Durante años, los diccionarios tradicionales contenían ejemplos inventados, sin un contexto natural y se compilaban principalmente sobre la base de la intuición y la introspección de los compiladores de diccionarios (Hanks, 2012). Luego, el uso de corpus en la lexicografía cambió esta situación pues el análisis de corpus ofrece a los lexicógrafos la posibilidad de elaborar diccionarios basados en datos empíricos y datos auténticos (Hanks, 2012).

Considerando la importancia y las ventajas que el enfoque basado en corpus tiene para develar las complejidades que representa el estudio de la lengua y que, a la fecha, no se ha elaborado una revisión que dé cuenta de la evolución de la relación entre dicho enfoque y las áreas de la Lexicología y de la Terminología, se desarrolló este análisis bibliométrico con el fin de presentar los documentos más relevantes y sus contribuciones al respecto. Para ello, se hizo una búsqueda en la base de datos Web of Science (WoS) utilizando la categoría *Linguistics* mediante una ecuación que incluyó los conceptos de interés, lo cual asegura la inclusión de las revistas especializadas más importantes y la identificación los artículos que dan cuenta de la temática.

Adicionalmente, se hicieron análisis de redes para determinar la productividad de las investigaciones, su evolución, y visibilidad, así como la actividad científica y el impacto de las fuentes. También se utilizó la herramienta VOSViewer para graficar los datos obtenidos. Aunque el periodo de tiempo está restringido a 10 años, el análisis de la información da cuenta de algunas de las fuentes de datos más importantes, ofrece datos cuantitativos y muestra comportamientos de la temática, por lo que se espera que los resultados aquí presentados contribuyan a la orientación de investigaciones posteriores en las áreas analizadas.

Método

Los documentos que conformaron el corpus para llevar a cabo el análisis bibliométrico se recuperaron utilizando la categoría *Linguistics* de la base de datos Web of Science (WoS), en una ventana temporal entre 2012 y 2021. Se hizo una búsqueda bajo el concepto de ecuación

canónica, es decir, «aquella que combina dos o más conceptos, con al menos uno de ellos con la necesidad de ser representado con dos o más sinónimos» (Codina, 2020, pág. 5).

De esta manera, se estableció la siguiente ecuación de búsqueda[1] para la extracción de documentos: (ling??stica de corpus OR corpus OR metodolog?a? corpus OR corpus linguistics OR corpora OR corpus method*) AND (terminolog?a OR lexicolog?a especializada OR lexicograf?a especializada OR terminograf?a OR terminology OR specialized lexicology OR specialized lexicography OR terminography) en el campo TOPIC (*title, abstract, author keywords, and Keywords Plus*) y las tipologías *Article, Review, Early access, Proceedings papers*.

La Tabla 1 muestra la descripción de los indicadores analizados sobre la implementación de la lingüística de corpus en estudios de las áreas de la Terminología y la Lexicología.

Tabla 1
Descripción de los indicadores analizados

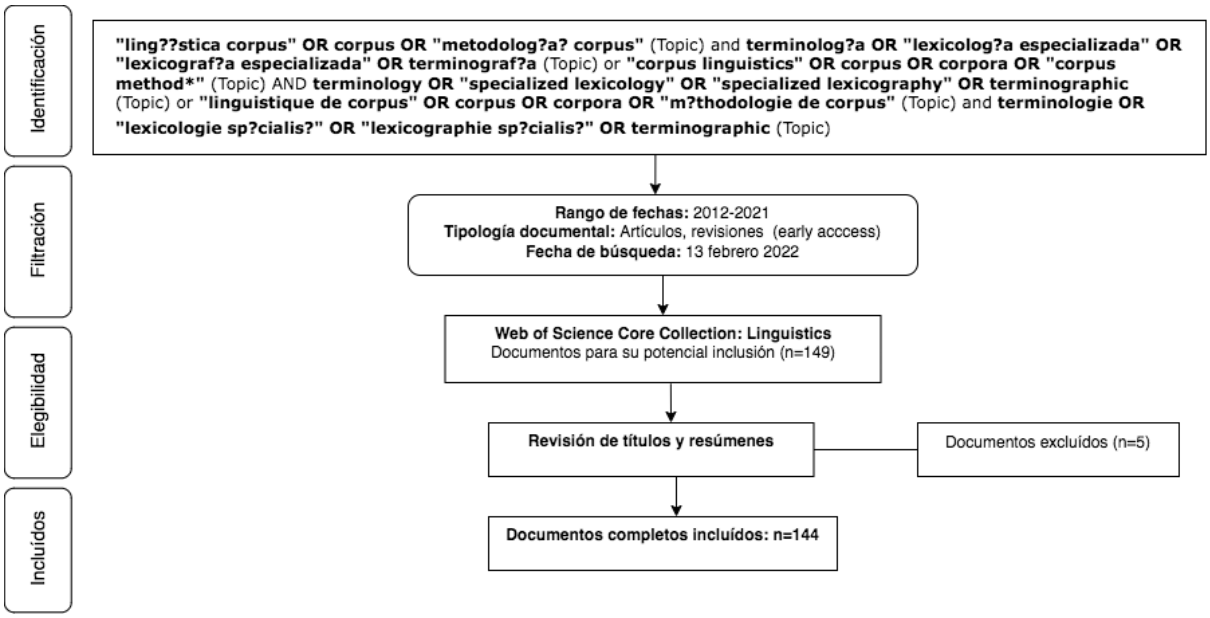
Indicador	Descripción
Comportamiento de la producción científica	Revela las regularidades y tendencias. Se utilizó el modelo de Price que permite evaluar el ritmo de crecimiento de la información científica en el área de interés.
Productividad de los autores y los países	Muestra si una menor cantidad de autores reúne una mayor cantidad de la producción científica. Para ello, se utilizó la ley de Lotka que evidencia que existe una distribución desigual pues la mayoría de los artículos se concentran en un pequeño número de autores altamente productivos, y una relación negativa respecto de su productividad de más o menos igual a dos.
Producción por revistas	Establece las revistas fuente de la producción científica y sus indicadores de visibilidad e impacto. Se utiliza el modelo de Bradford que establece que existe una distribución altamente desigual en la producción de artículos en las revistas porque la mayoría de los artículos están concentrados en una pequeña cantidad de revistas.
Red de coautoría	Consiste en una representación del sistema que surge de las relaciones colaborativas entre autores que investigan en determinada área del conocimiento.
Patrones de colaboración	Indica cómo se relacionan los autores en el proceso de escritura y el grado de apertura de la investigación.
Liderazgo científico	Marca los autores, los países y las instituciones que encabezan la participación en las investigaciones y, por ende, la producción de documentos.
Red de palabras clave	Muestra los nombres de los principales descriptores en los documentos revisados para facilitar el análisis del enfoque temático y las áreas de investigación partir de la creación de los clústeres.
Perspectivas investigativas	Revela aquellos documentos publicados de manera más reciente en el área que permiten determinar el futuro de los estudios en el área.

[1] Se usaron truncadores o máscaras para ampliar la búsqueda en caso de plurales y acentos (ejemplo: terminolog?a y el truncador *(asterisco) para ampliar la búsqueda de la raíz de una palabra (ejemplo: method*)).

La Figura 1 muestra el proceso de selección que dio como resultado 144 documentos. Las variables autor, institución, país y palabras clave fueron normalizadas porque son la base para generar indicadores bibliométricos. Esta estrategia de búsqueda permitió recuperar 149 documentos a los que se les hizo un proceso de normalización de metadatos; se eliminaron 5 documentos por no cumplir los criterios de inclusión del presente estudio, especialmente en cuanto a la temática (*corpus christi, corpus callosum*, por ejemplo).

Figura 1

Planificación del proceso de búsqueda y selección de documentos



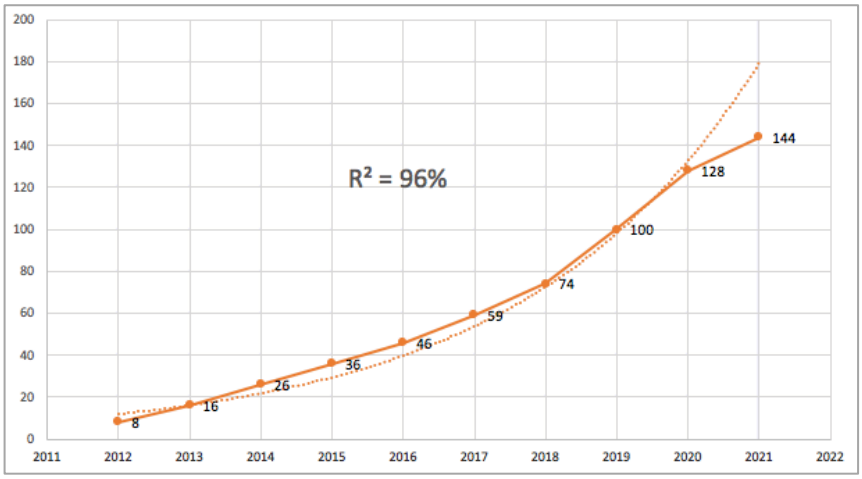
Resultados

Comportamiento anual de la producción

La producción científica acumulada se determinó utilizado el modelo exponencial de Price con un promedio de tasa de variación interanual, es decir, la variación relativa en comparación con el valor inicial de la variable, de 11 % y un índice de bondad de ajuste, es decir, la discrepancia entre los valores observados y los valores esperados en el modelo de estudio, de 96 %.

La Figura 2 muestra que la temática de interés para este estudio tiene una tendencia de crecimiento exponencial en cuanto a publicaciones en el período de tiempo establecido, es decir, 136 entre 2012 y 2021.

Figura 2
Producción científica acumulada 2012-2021

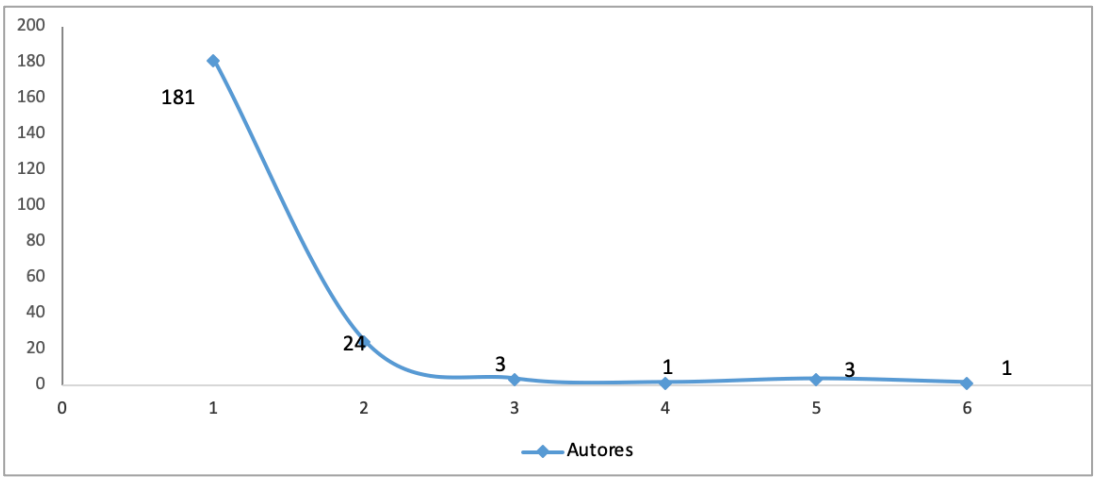


Puede observarse que entre los intervalos de 2012 a 2013 (8 publicaciones en cada año), de 2014 a 2016 (10 publicaciones en cada año) no hubo crecimiento de la producción. El incremento de la productividad empezó a darse en 2017 con 3 publicaciones más, 2018 con 2 más, 2019 con 11 más, y 2020 con 2 más. En contraste, 2021 fue el único año cuyas publicaciones decrecieron en 12. También se estableció que entre 2018 y 2019 se presentó el mayor promedio de crecimiento de la productividad científica (73 %), mientras que entre 2020 y 2021 se presentó el menor (-43 %).

Liderazgo de productividad científica por autor

Este indicador se obtuvo aplicando la Ley de Lotka que describe la relación cuantitativa entre los autores y la frecuencia de sus contribuciones en un campo dado a lo largo de un periodo de tiempo. La Figura 3 muestra que la relación de producción de los 213 autores que participan en los 144 documentos recuperados es la siguiente: 181 autores han aportado en 1 documento, 24 autores han aportado en 2 documentos, 3 autores han aportado en 3 documentos, 4 autores han aportado en 1 documento, 5 autores han aportado en 3 documentos y 1 autor ha aportado en 6 documentos.

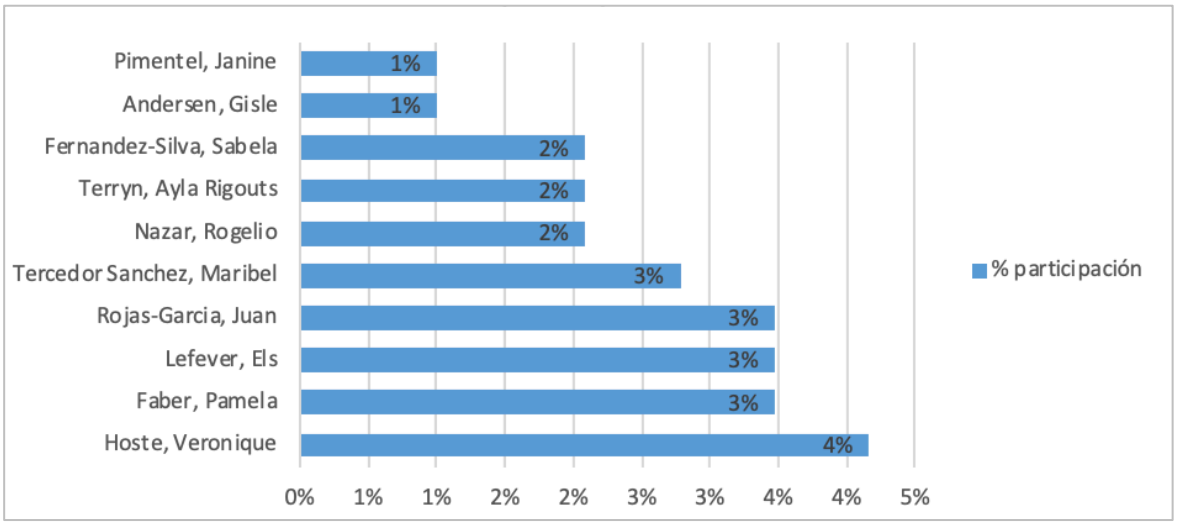
Figura 3
Productividad de los autores



De este modo, con el modelo aplicado de distribución inversa, se identifican los núcleos de producción y se destacan los investigadores élite a través de sus contribuciones. Puede establecerse que los 8 autores más productivos, con un número mayor a 3 publicaciones, participan en 34 documentos y compilan el 23 % de la producción científica para el universo de esta investigación, es decir, entre mayor acercamiento al eje X de los autores, mayor es la productividad en el tema. La Figura 4 muestra el porcentaje de productividad de los autores más especializados en el tema.

Figura 4

Porcentaje de productividad de los autores



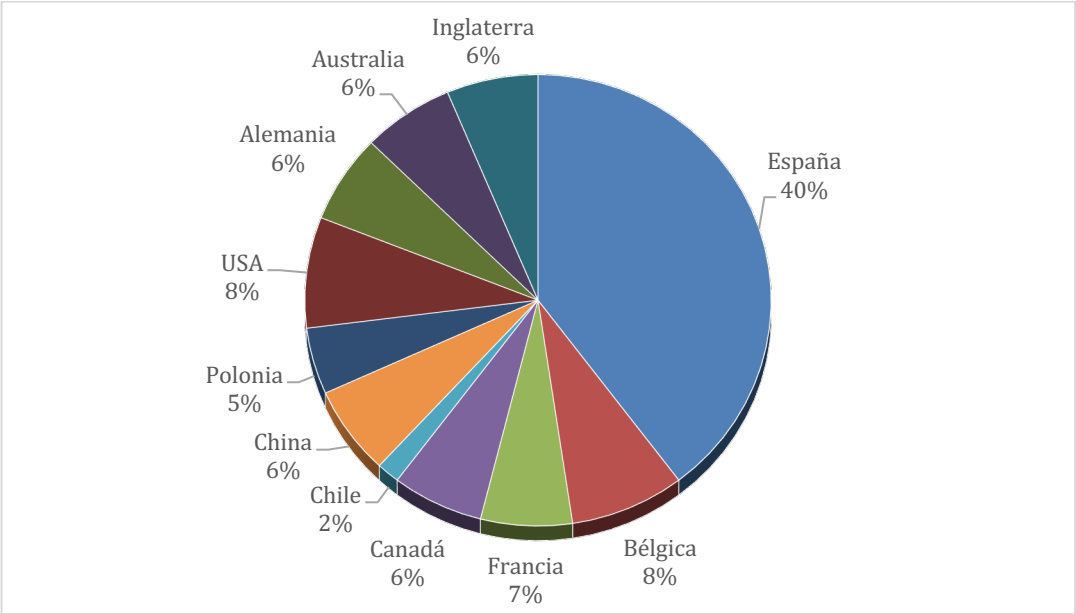
En este caso, Veronique Hoste, con aportes en 6 documentos y 40 citaciones, Pamela Faber, con participación en 5 documentos y 37 citaciones, Els Lefever, con participación en 5 documentos y 13 citaciones, Juan Rojas-García con participación en 5 documentos y 4 citaciones y Maribel Tercedor-Sánchez con participación en 4 documentos y 13 citaciones son los autores que lideran la productividad con aportes por encima de los 3 documentos. Por su parte, Rogelio Nazar, Ayla Rigouts Terryn y Sabela Fernández-Silva tienen participación en 3 documentos cada uno, y 9, 2 y 1 citación respectivamente.

Liderazgo de productividad científica por país

El liderazgo científico por país se determina a través del autor de correspondencia que ejerce la función de contacto principal y determina la afiliación institucional y, por ende, la nación a la que pertenece. Este factor permite además establecer las capacidades científicas de un país en el contexto investigativo. En este estudio, 25 instituciones españolas y 5 belgas reúnen un total de 47 % de la productividad en la temática de este análisis.

Adicionalmente, se identificó el porcentaje de productividad de los países de acuerdo con el número de instituciones. También se evidenció la prevalencia de la lengua inglesa con 78.5 % de las publicaciones, seguida por la lengua española con 15 %. La Figura 5 muestra los 12 países que aportan más de 3 documentos en el área estudiada.

Figura 5
Porcentaje de productividad por país según instituciones



Con respecto a los 144 documentos base de este estudio, España se consolida como país líder de productividad con 57 documentos publicados por 25 instituciones (40 %), seguida por Bélgica con 10 documentos publicados por 5 instituciones (8 %). Después se encuentran Francia con 6 documentos publicados por 4 instituciones (7 %), Canadá con 5 documentos publicados por 4 instituciones (6 %), Chile con 5 documentos publicados por 1 institución (2 %), China con 5 documentos publicados por 4 instituciones (6 %), Polonia con 5 documentos publicados por 3 instituciones (5 %), los EE. UU. con 5 documentos publicados por 5 instituciones (8 %), Alemania con 4 documentos publicados por 4 instituciones (6 %), Australia con 4 documentos publicados por 4 instituciones (6 %) e Inglaterra con 4 documentos publicados por 4 instituciones (6 %).

Liderazgo de productividad científica por institución

Considerando que las universidades contemporáneas tienen tres funciones inherentes: docencia, investigación y extensión, es importante hacer visible su liderazgo en la productividad de investigaciones científicas. Este indicador también se determina a través de la afiliación institucional del autor de correspondencia y permite establecer las capacidades científicas de una institución para hacer aportes en diversas áreas del contexto investigativo. En este estudio se encontraron 89 instituciones que han publicado documentos relacionados con la lingüística de corpus. La Tabla 2 presenta las instituciones con un número de publicaciones mayor a 3, el país al que pertenecen y el número de documentos que aportan.

Tabla 2
Liderazgo científico por institución

Universidad	**País**	**Nº Doc.**	**%**
Universidad Granada	España	15	10 %
Universidad Ghent	Bélgica	6	4 %
Pontificia Universidad Católica de Valparaíso	Chile	5	3 %
Universidad Córdoba	Argentina	4	3 %
Universidad Politécnica de Valencia	España	4	3 %
Universidad Valladolid	España	4	3 %
Universidad Vigo	España	4	3 %
Universidad Hong Kong	China	3	2 %
Universidad París	Francia	3	2 %

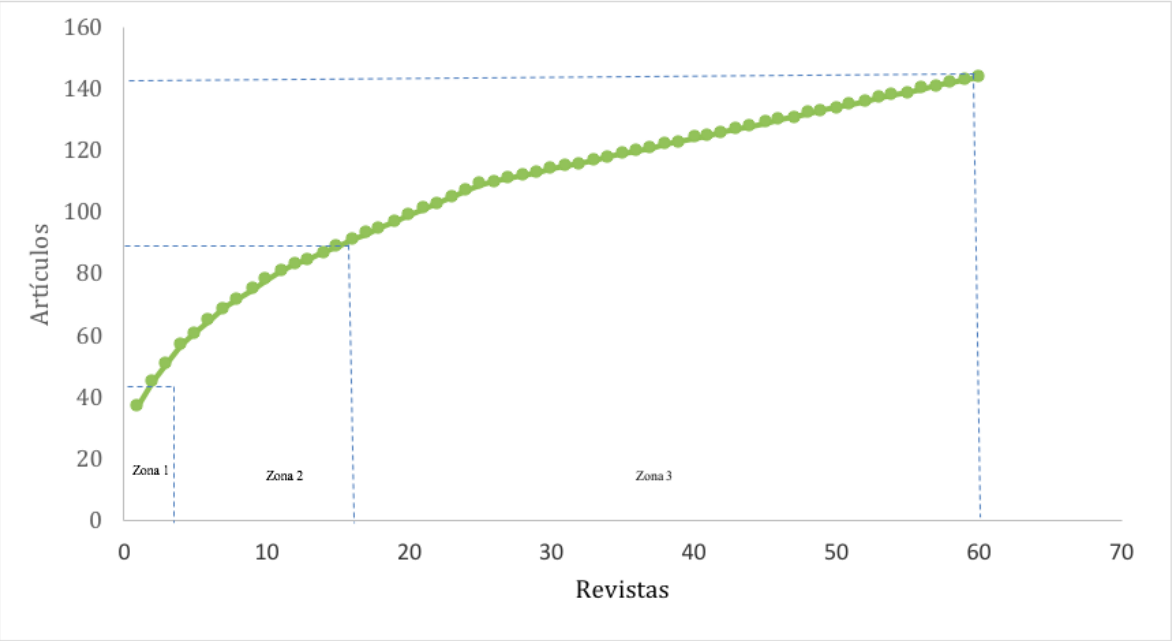
La Universidad de Granada (España) lleva el liderazgo con 15 documentos publicados (10 %), seguida por la Universidad de Ghent (Bélgica) con 6 documentos publicados (4 %). Los autores que aportan más de 5 documentos pertenecen a estas dos instituciones.

Después se encuentran la Pontificia Universidad Católica de Valparaíso (Chile) con 5 documentos publicados, seguida por la Universidad de Córdoba (Argentina), la Universidad Politécnica de Valencia (España), la Universidad de Valladolid (España), y la Universidad de Vigo (España), cada una con 4 documentos publicados (3 % cada una). A ellas le siguen la Universidad de Hong Kong (China) y la Universidad de París (Francia), cada una con 3 documentos publicados (2 % cada una). Finalmente, las 80 instituciones restantes aportan 1 o 2 artículos publicados (1 %) al total de los datos.

Liderazgo de productividad científica por revistas

El modelo de dispersión de Bradford permitió identificar las publicaciones periódicas más relevantes y observar la ubicación de las revistas más utilizadas por los investigadores en el tema de interés en la zona nuclear, la zona intermedia o la zona periférica que van mostrando la productividad de las revistas de mayor a menor. La Figura 6 muestra la distribución logarítmica de las 60 revistas que publicaron artículos sobre la temática de interés de este estudio.

Figura 6
Distribución de revistas por zonas según el modelo de Bradford



La aplicación del modelo mencionado en este análisis permitió establecer que en la Zona 1 (núcleo) aparecen 2 revistas con 45 artículos que concentran el 31 % de publicaciones del total de la muestra. Entre tanto, en la Zona 2 (intermedia) hay 15 revistas con 48 artículos publicados (33 %) y en la Zona 3 (periferia) hay 43 revistas con 51 artículos publicados (35 %).

La Tabla 3 muestra las características de las 2 revistas ubicadas en la Zona 1. Se trata de *Terminology*, editada por John Benjamins Publishing Company y *Onomázein*, editada por la Pontificia Universidad Católica de Chile. El nivel académico de los trabajos publicados en ambas revistas está garantizado por la revisión objetiva de jueces externos internacionales, reclutados entre la comunidad internacional de especialistas.

Tabla 3
Descripción de las revistas de la Zona del núcleo

Revista	Número de documentos	% 144	Cuartil categoría: ***Linguistics WoS***	JIF 2020	H-index
Terminology	37	26 %	Q2	0.826	25
Onomázein	8	6 %	Q2	0.419	12

De acuerdo con el sitio web de Scimago, *Terminology* es una revista independiente de ámbito transcultural e interdisciplinar. Se centra en la discusión de soluciones (sistemáticas), no sólo de los problemas lingüísticos encontrados en la Traducción, sino también, por ejemplo, de los problemas (monolingües) de ambigüedad, referencia y desarrollo en la comunicación multidisciplinar. En la categoría *Linguistics and Language*, *Terminology* se ubica actualmente en Q2.

Por su parte, de acuerdo con el sitio web de Scimago, *Onomázein* –Revista de Lingüística, Filología y Traducción– acoge trabajos no publicados anteriormente, originados en la investigación científica en las diferentes ramas de la Lingüística Teórica y Aplicada, en Filología Clásica, Indoeuropea, Románica e Hispánica, así como en Teoría de la Traducción y Terminología, y estudios relevantes en lenguas indígenas. En la categoría *Linguistics and Language*, *Onomázein* se ubica actualmente en Q2.

Redes de colaboración entre autores y entre instituciones

A partir de los 213 autores que se identificaron en los 144 documentos, se seleccionaron aquellos con al menos 3 colaboraciones para la construcción de la red de coautoría; así se llegó a 8 autores que conformaron 4 clústeres. A continuación, se construye una matriz de coautoría en la que se identifican las veces que estos autores top trabajaron conjuntamente.

La Figura 7 muestra que el clúster rojo está conformado por las autoras Veronique Hoste, Ayla Rigouts Terryn y Els Lefever, vinculadas con la Ghent University. El clúster azul está conformado por Sabela Fernández (Universidad Católica de Valparaíso) y Maribel Tercedor-Sánchez (Universidad de Granada). El clúster verde está conformado por Pamela Faber y Juan Rojas-García (Universidad de Granada) y, finalmente, se encuentra Rogelio Nazar de la Universidad Católica de Valparaíso con un aporte en solitario.

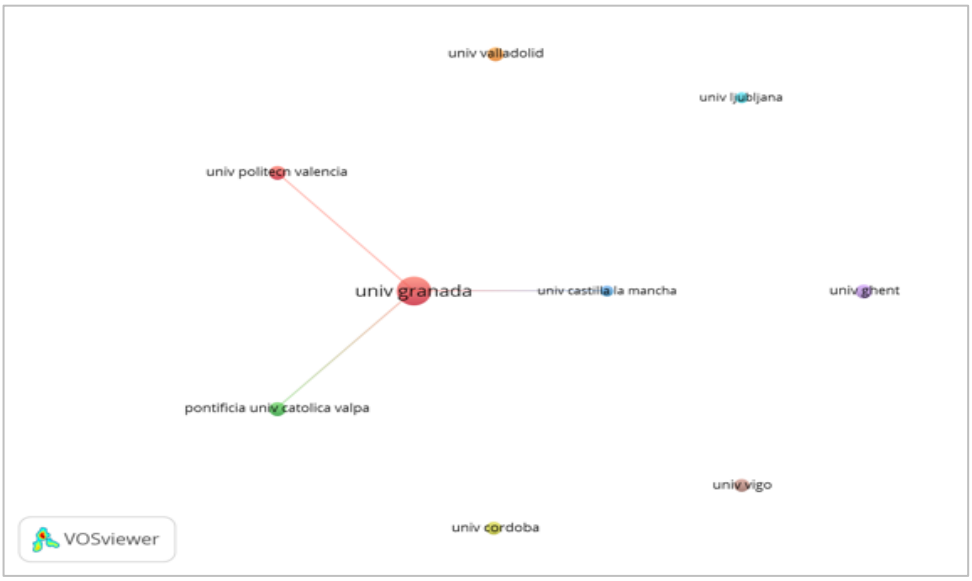
Figura 7
Red de coautoría

Nota. fuerza de enlace mínima de los ítems: 0. De los 213 autores, 8 cumplieron el umbral (3 documentos); método de normalización: *association strength* (fuerza de asociación); atracción: 1; repulsión: -2; resolución de agrupamiento: 1,0.
Resultado: ítems: 8; clústeres: 4; Links: 5; *Total Link strength*: 15.

De las 116 instituciones que participan en las publicaciones de este análisis, 9 se identificaron por medio de un punto de corte mínimo de 3 documentos por cada una (no se incluyó la citación) para construir la red de colaboración institucional. La Figura 9 muestra que la Universidad de Granada lidera la producción de este tema en la categoría *Linguistics* y, por ende, se establece como el punto de colaboración institucional con las Universidades Castilla de la Mancha y la Politécnica de Valencia, ambas de España, y la Pontificia Universidad Católica de Valparaíso de Chile.

Además, se evidenció una participación conjunta en 20 documentos y liderazgo de 15 instituciones.

Figura 8
Red de colaboración entre instituciones



Nota: fuerza de enlace mínima de los ítems: 0. De las 116 instituciones, 9 cumplieron el umbral (3 documentos); método de normalización: *association strength* (fuerza de asociación); atracción: 4; repulsión: -5; resolución de agrupamiento: 1,0
Resultado: ítems: 9; clusters: 8; Links: 3.

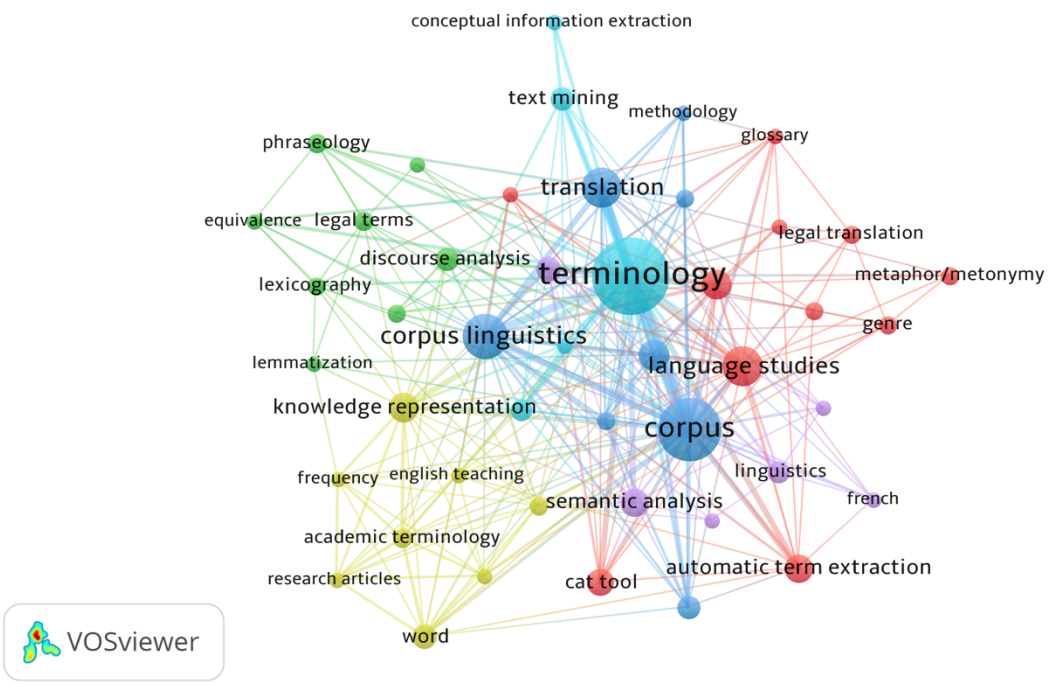
Es de destacar que el artículo "Pragmatic borrowing" de la autora Andersen Gisle de la Norwegian School of Economics tiene el mayor número de citas en todo el conjunto de datos (45 citaciones en total). Este artículo explora la noción de préstamo pragmático, es decir, la incorporación de rasgos pragmáticos y discursivos de una lengua de origen a una lengua receptora. El estudio ilustra cómo las funciones pragmáticas se transfieren de forma interlingüística, a través de nociones como la estabilidad funcional, la adaptación, el estrechamiento, la ampliación y el cambio. También ilustra el grado de préstamo de frases fijas y coloquialismos, centrándose especialmente en los improperios, las interjecciones y los marcadores del discurso en inglés que han aparecido recientemente en noruego.

Redes de coocurrencia

En la muestra de los 144 artículos se obtuvieron 634 palabras clave que fueron normalizadas a 453 luego de crear y aplicar una lista de términos o tesauro. Para simplificar la representación de las estructuras de conocimiento se consideraron solo aquellas palabras clave cuya frecuencia fuera ≥3 (un umbral más bajo habría generado una lista de palabras clave muy larga). Antes de crear la red de co-palabras, se eliminaron manualmente las palabras clave *named river* y *wine testing notes* porque aparecían relacionadas con la palabra corpus, pero en un ámbito diferente a la lingüística.

La Figura 9 muestra los 6 clústeres que se obtuvieron. En la interpretación del mapa se tuvo en cuenta el número de palabras clave dentro de cada grupo temático, el número de ocurrencias de cada palabra clave, su interrelación y su localización espacial.

Figura 9
Clústeres de coocurrencia de palabras clave



Nota: fuerza de enlace mínima de los ítems: 0. De las 634 keywords (author + keyword plus), 453 cumplieron el umbral (3 ocurrencias); método de normalización: association strength (fuerza de asociación); atracción: 1; repulsión: -3; resolución de agrupamiento: 1,0.

Los colores indican agrupaciones de palabras clave con algún tipo de relación entre sí según la asociación obtenida mediante el programa VOSviewer. Se hizo además un análisis del enfoque temático de cada clúster tomando como base los conceptos transmitidos mediante sus palabras clave. La Tabla 4 muestra los clústeres y los enfoques temáticos de cada uno, sus palabras clave, el número de ocurrencias.

Tabla 4
Clústeres y enfoques temáticos

Clúster	Palabra clave	Ocurrencias	Enfoque temático
Cluster 1. Rojo	automatic term extraction	10	
	cat tool	9	
	comparable corpora	4	
	english	11	
	genre	4	
	glossary	3	
	language studies	20	Traducción
	legal translation	4	
	metaphor/metonymy	4	
	natural language processing	3	
	standardization	3	

Cluster 2. Verde	discourse analysis	7	Estudios de traducción
	equivalence	3	
	eu terminology	3	
	legal terms	5	
	lemmatization	3	
	lexicography	4	
	phraseology	5	
	science	4	
Cluster 3. Azul	corpus	49	Traducción especializada
	corpus linguistics	25	
	lexicology	12	
	medical terminology	7	
	medical translation	4	
	methodology	3	
	spanish	4	
	translation	20	
Cluster 4. Amarillo	academic terminology	5	
	collocation	4	Didáctica de las lenguas
	engineering english	3	
	english teaching	3	
	frequency	3	
	knowledge representation	11	
	research articles	3	
	word	8	
Cluster 5. Morado	distributional semantics	3	Terminología
	french	3	
	grammar	3	
	linguistics	6	
	semantic analysis	10	
	term extraction	8	
Cluster 6. Azul claro	conceptual information extraction	3	Terminótica
	FunGramKB	6	
	ontology	7	
	terminology	4	
	text mining	7	

Nota: Elaboración propia.

Perspectivas investigativas

Utilizando la metodología propuesta por Robledo, Osorio y López (2014), se cargaron los documentos recuperados a la plataforma web *Tree of Science* (ToS) con el fin de clasificar los artículos de acuerdo a su posición en el árbol, analogía usada por los autores mencionados para determinar los siguientes tres grupos:

- La raíz con los referentes teóricos del tema, es decir, los autores clásicos que sentaron las bases en ese campo de estudio y que son citados con mayor frecuencia que el resto de los autores. Estos artículos tienen fechas entre 1991 y 2010.
- El tronco con los artículos estructurales basados en los artículos de la raíz, pero con un marco teórico más elaborado. Estos artículos tienen fechas entre 2012 y 2019.

- Las hojas con aquellos artículos que muestran las diferentes perspectivas actuales del tema, utilizando los hallazgos de los artículos de la raíz y del tronco. Estos artículos tienen fechas entre 2013 y 2021.

La Tabla 5 muestra esta última categoría por ser la de mayor relevancia para el presente estudio. Los artículos de las hojas se caracterizan también por tener como referencia los escritos que conforman las raíces y el tronco.

Tabla 5
Artículos más actuales del tema

Autor	Artículo
Rojas-García, J. (2021).	Extraction of Terms Semantically Related to Colponyms: Evaluation in a Small Specialized Corpus.
Kwong, OY. (2021).	User-driven assessment of commercial term extractors.
Rojas-García, J. (2020).	Application of Topic Modelling for the Construction of Semantic Frames for Named Rivers.
Ortego-Anton, MT. (2021).	e-DriMe A Spanish-English frame-based e-dictionary about dried meats.
Terryn, AR. (2021).	HAMLET Hybrid Adaptable Machine Learning approach to Extract Terminology.
Unzalu, IZ. (2021).	[en] Current challenges in the development and learning of the oral and written academic registers in Basque.
Polyakova, O. (2021).	An integrated approach to the higher education terminology in Spanish-Russian university texts.
Trigo, ES. (2021).	The terms manifestation (fr) and manifestación (es) in biomedical journal articles: a corpus-based research.
Rodríguez, CIL. (2020).	Predicative frames for the concept SIGN AND SYMPTOM in Spanish Medical Texts.
San Martin, A. (2020).	Present and future of the terminological knowledge base EcoLexicon.
Hoste, V. (2019).	The trade-off between quantity and quality. Comparing a large crawled corpus and a small focused corpus for medical terminology extraction.
Rieder-Bunemann, A. (2019).	Capturing technical terms in spoken CLIL A holistic model for identifying subject-specific vocabulary.
Cárdenas, BS. (2019).	Eliciting specialized frames from corpora using argument-structure extraction techniques.
Santos, IG. (2019).	La economía esta enferma- l'economie est malade. The chronology of the crisis through terminology.
Terryn, AR. (2019).	Validating multilingual hybrid automatic term extraction for search engine optimisation: the use case of EBM-GUIDELINES.
Perinan-Pascual, C. (2018).	A framework of analysis for the evaluation of automatic term extractors.
Ghazzawi, N. (2018).	Automatic extraction of specialized verbal units A comparative study on Arabic, English and French.
Costa, LA. (2018).	Explicit term variation in Brazilian lexicography: proposal for its representation in the micro structure of the Brazilian Lexicography Dictionary.
Perinan-Pascual, C. (2018).	DEXTER: A workbench for automatic term extraction with specialized corpora.
Gagne, AM. (2016).	Opposite relationships in terminology.

Nazar, R. (2016).	Distributional analysis applied to terminology extraction First results in the domain of psychiatry in Spanish.
Hanoulle, S. (2015).	The efficacy of terminology-extraction systems for the translation of documentaries.
Lefever, E. (2014).	HypoTerm Detection of hypernym relations between domain-specific terms in Dutch and English.
Silva, SF. (2013).	The influence of the disciplinary field on terminological variation: A corpus-based study in the interdisciplinary domain of fishing.

Nota: Elaboración propia.

Una vez se identificaron los artículos más recientes, se hizo un proceso simple de minería de datos utilizando los títulos de los mismos y se creó una nube de palabras mediante la herramienta Voyant. Esto se hizo con el fin de determinar las temáticas que se están trabajando actualmente y que, a su vez, sientan las bases para investigaciones futuras. La Figura 10 muestra los términos de mayor frecuencia en cada título.

Figura 10

Nube de palabras a partir de los títulos de los artículos "hojas"

Nota. Fuente: Voyant tools.

Puede observarse que los temas más relevantes están relacionados con dinámicas de extracción automática de términos, metodología de corpus, semántica de marcos y trabajos de lexicología y terminología en ámbitos de especialidad.

Discusión y conclusiones

Las técnicas de minería de textos tienen una gran pertinencia a la hora de contribuir con la investigación académica porque facilitan la elaboración de una amplia gama de análisis que permiten explorar corpus textuales a fondo y de manera minuciosa. Además, es importante tener en cuenta que un análisis bibliométrico representa una herramienta fundamental que brinda

confiabilidad al proceso de investigación, pues posibilita la indagación de resultados de estudios previos.

Todos estos procesos exploratorios de trabajos antecedentes, por lo tanto, se han convertido en pieza fundamental a la hora de plantear objetivos, planear metodologías y proponer diseños de investigación, entre otros, con el fin de garantizar la relevancia de los aportes. Hacer una delimitación de la información que se busca extraer, cuál es el procedimiento adecuado y qué tipo de datos se obtendrán permite al investigador optimizar cada paso en aspectos de tiempo y calidad.

Por estas razones, la combinación de las técnicas de minería de datos con elementos bibliométricos y las herramientas utilizadas en este estudio permitieron establecer que la lingüística de corpus se ha consolidado como un enfoque fundamental en los estudios de áreas como la Terminología, la Lexicografía, la Traducción y la enseñanza de lenguas, entre otras. Los lenguajes de especialidad son otra área que merece especial mención puesto que se encontró un volumen importante de investigaciones en ámbitos como la medicina, la ingeniería o la administración, cuyos objetivos giran en torno a la resignificación, uso o normalización de términos. Fue posible observar también que los estudios que incluyen procesos con lingüística de corpus tuvieron un incremento constante entre el 2012 y el 2021, excepto entre 2020 y 2021. Esto podría atribuirse a las medidas de confinamiento adoptadas a nivel mundial ante la pandemia de la COVID-19.

En cuanto a los resultados del liderazgo de productividad científica por autor estuvieron en consonancia con el modelo aplicado de distribución de la Ley de Lotka que establece una relación inversa en la que unos pocos autores se especializan en un campo del conocimiento y, por ende, concentran el mayor volumen de publicaciones, mientras que muchos autores harán muy pocas publicaciones.

Es notable que los resultados de productividad científica reflejan los esfuerzos coordinados por parte de instituciones y académicos en la búsqueda de descripciones desde la interdisciplinariedad y cada vez más detalladas de los fenómenos lingüísticos. Se encontró que el total de la producción está representado en 60 revistas científicas con una participación de 213 autores. Aunque para los propósitos de este estudio se mencionen los datos cuantitativamente más destacados, es importante destacar que, aparte de España y Bélgica, otros 13 países aportan tres documentos o más a las publicaciones por país.

En lo que respecta a productividad por país, España justifica su liderazgo, ya que la Universidad de Granada se encuentra entre las diez primeras; además, hace varios años que ocupa el primer lugar en estudios de traducción e interpretación, y cuenta con programas de enseñanza de lenguas tales como portugués, italiano, danés, neerlandés, checo, polaco, rumano, búlgaro, ruso, griego moderno, hebreo, árabe o turco. También alberga el único Centro Ruso que la Fundación Russkiy Mir mantiene en España. La Universidad de Granada es una de las mejores universidades públicas en España y ocupa el puesto 494 en el QS *Academic Ranking of World Universities* 2023. Por su parte, Bélgica está en el segundo lugar de este liderazgo, con la Universidad de Ghent a la cabeza. Esta institución ocupa el puesto 74 en esta lista de más de 2500 instituciones de investigación de todo el mundo en el año 2022 y es la universidad belga mejor clasificada en el *Academic Ranking of World Universities*.

En relación con el liderazgo de productividad científica por revistas, el presente análisis tiene correspondencia con la hipótesis de Bradford (1934), quien postuló que la mayoría de los artículos sobre un asunto especializado podrían ser publicados por unas cuantas revistas especialmente dedicadas a dicho asunto, en conjunto con ciertas revistas de frontera y otras generales o de dispersión (Urbizagástegui Alvarado, 2015). En este caso, *Terminology* y *Onomázein* van a la cabeza de las publicaciones en el área de interés de este análisis.

De acuerdo con su sitio web, *Terminology* presta especial atención a las áreas temáticas nuevas y en desarrollo, como la representación y transferencia de conocimientos, las herramientas informáticas, los sistemas expertos y las bases de datos terminológicas. *Terminology* abarca lo general (teoría y práctica) y los campos especializados (LSP), como la Física, las Ciencias Biomédicas, la Tecnología, la Ingeniería, las Humanidades, la Administración, el Derecho, las Artes, la Administración de Empresas, el Comercio, la Identidad Corporativa, la Economía, la Metodología y cualquier otra área en la que la Terminología sea esencial para mejorar la comunicación. Por su parte, *Onomázein* está orientada principalmente a especialistas y pretende servir de vehículo eficaz de intercambio científico entre los investigadores de las ciencias lingüísticas.

Uno de los trabajos ya publicados que tiene hallazgos similares fue el de Liao y Lei, quienes en 2017 desarrollaron un análisis bibliométrico del SSCI (*Social Science Citacion Index*) de WoS utilizando la categoría *Linguistics* OR *Language Linguistics*, entre los años 2000 y 2015. Su propósito era conocer la cantidad de documentos que implementaran metodología de corpus. Los resultados mostraron que la producción de publicaciones relacionadas con los corpus había aumentado considerablemente durante esos 15 años. Además, fue notorio que, si bien las potencias científicas tradicionales, como los Estados Unidos, desempeñan un papel destacado en este ámbito, países como China también ejercen su impacto en el área. El resultado más importante se relacionó con el hecho de que los corpus han impregnado una amplia gama de áreas de investigación en lingüística y han cambiado, al menos en términos de metodología, estas áreas.

Incluso cuando este análisis solo incluye un intervalo de diez años, los resultados obtenidos a través de las herramientas de minería de textos y de las técnicas de análisis biométrico fueron coherentes. Podrían desarrollarse nuevas investigaciones aumentando el volumen de información y utilizando herramientas adicionales para lograr ampliar los resultados y revelar otras tendencias asociadas al tema estudiado.

Referencias

Ardanuy, J. (2012). *Breve introducción a la bibliometría.* http://diposit.ub.edu/dspace/bitstream/2445/30962/1/breve %20introduccion %20bibliometria.pdf

Cobo, M. J. (2011). Science Mapping Software Tools: Review, Analysis, and Cooperative Study Among Tools. *Journal of the American Society for Information Science and Technology*, 1382–1402. https://doi.org/10.1002/asi.21525

Codina, L. (2020). *Interfaces de búsqueda en bases de datos académicas: Análisis comparativo de Scopus, WoS, Google Scholar y Microsoft Academic.* https://www.lluiscodina.com/interfaces-de-busqueda/

Galves, C. (2018). El campo de investigación del Análisis de Redes Sociales en el área de las Ciencias de la Documentación: un análisis de co-citación y co-palabras. *Revista General de Información y Documentación, 28*(2), 455-475. http://dx.doi.org/10.5209/RGID.60805

Hanks, P. (2012). *Corpus evidence and electronic lexicography*.

Liao, S., & Lei, L. (2017). What we talk about when we talk about corpus: A bibliometric analysis of corpus-related research in linguistics (2000-2015). *Glottometrics*, *38*, 1-20.

Maltrás Barba, B. (2003). *Los indicadores bibliométricos. Fundamentos y aplicación al análisis de la ciencia.* Trea S.L.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Merton, R. K. (1968). The Matthew Effect in Science: The reward and communication systems of science are considered. *Science, 159*(3810), 56-63.

Pulgarín, A., Carapeto, C., & Cobos, J. M. (2004). Análisis bibliométrico de la literatura científica publicada en Ciencia. Revista hispano-americana de ciencias puras y aplicadas (1940-1974). *Inf. Res.*, *9*(4).

Robledo, S.; Osorio, G.; Lopez, C. (2014). Networking en pequeña empresa: una revisión bibliográfica utilizando la teoría de grafos. *Revista Vínculos, 11*(2), 6-16. https://revistas.udistrital.edu.co/index.php/vinculos/article/view/9664

Urbizagástegui Alvarado, R. (2016). El crecimiento de la literatura sobre la ley de Bradford. Investigación Bibliotecológica: Archivonomía, Bibliotecología e Información, 30(68), 51-72. https://doi.org/10.1016/j.ibbai.2016.02.003

Fecha de recepción: 29/08/2022
Fecha de revisión: 22/09/2022
Fecha de aceptación: 29/11/2022