Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação
ISSN: 1518-2924
Universidade Federal de Santa Catarina

Restrepo-Arango, Cristina; Cárdenas-Rozo, Andrés L.

ANÁLISIS TEXTUAL DE ARTÍCULOS CIENTÍFICOS PUBLICADOS SOBRE FÓSILES COLOMBIANOS

Encontros Bibli: revista eletrônica de biblioteconomia
e ciência da informação, vol. 27, 2022, pp. 1-25
Universidade Federal de Santa Catarina

DOI: https://doi.org/10.5007/1518-2924.2022.e83470

Disponible en: https://www.redalyc.org/articulo.oa?id=14775278015

Artigo Original

Encontros Bibli

ANÁLISIS TEXTUAL DE ARTÍCULOS CIENTÍFICOS PUBLICADOS SOBRE FÓSILES COLOMBIANOS

Textual analysis of scientific articles published on Colombian fossils

Cristina Restrepo-Arango
Doctora en Bibliotecología y Estudios de la Información
Universidad de Córdoba, División de Bibliotecas y Recursos Educativos, Montería, córdoba, Colombia
crestrepoarango@gmail.com
https://orcid.org/0000-0003-4275-4102

Andrés L. Cárdenas-Rozo
Grupo GITEC, Departamento de Ciencias de la Tierra, Escuela de Ciencias, Universidad Eafit, Medellín, Antioquia, Colombia
acarde17@eafit.edu.co
https://orcid.org/0000-0003-3849-1514

A lista completa com informações dos autores está no final do artigo

RESUMO

Objetivo: Identificar as proximidades lexicais em um corpus de textos de artigos científicos publicados em revistas acadêmicas indexadas na base de dados Scopus sobre fósseis colombianos.
Método: Este trabalho aplica a análise textual a cinco artigos paleontológicos sobre fósseis colombianos, a fim de identificar a proximidade lexical em um corpus de textos. Este trabalho permitiu determinar: as categorias gramaticais, a proximidade entre categorias de palavras e variáveis com a análise de especificidades (AE), o agrupamento das palavras com a análise da classificação hierárquica descendente (CJD) e a apresentação gráfica das palavras.
Resultado: Verificamos que o corpus documental é composto por 31.319 ocorrências de palavras, 1.450 formas ativas ou palavras específicas e 303 formas complementares ou palavras comuns. A categoria gramatical de substantivo predomina (24%) e palavras não reconhecidas no dicionário (17%). As palavras comuns com maior número de frequências são artigos, conjugações, proposições e pronomes.
Conclusões: Constatou-se que existe uma proximidade lexical entre o artigo 1 e as formas ativas de "Colômbia" e o artigo 2 e as formas ativas de "fóssil". As palavras foram agrupadas em cinco classes e a nuvem de palavras foi criada com 1271 palavras.
PALAVRAS-CHAVE: Colômbia. Iramuteq. Lexicon. Paleontologia.

ABSTRACT

Objective: Identify the lexical proximities in a corpus of texts of scientific articles published in academic journals indexed in the Scopus database on Colombian fossils.
Methodology: This work applies textual analysis to five paleontological articles on Colombian fossils to identify lexical proximity in a corpus of texts. This work allowed us to determine: the grammatical categories, the proximity between categories of words and variables with the analysis of specificities (AE), the grouping of the words with the study of the descending hierarchical classification (CJD) and the graphic presentation of the words.
Results: The documentary corpus comprises 31,319-word occurrences, 1,450 active forms or specific words and 303 complimentary forms or common words. The grammatical category of nouns predominates (24%) and words not recognized in the dictionary (17%). The familiar words with the highest frequencies are articles, conjugations, propositions, and pronouns.
Conclusions: It was found that there is linguistic proximity between article 1 and the active forms of "Colombia" and article 2 and the active forms of "fossil". The words were grouped into five classes, and the word cloud was created with 1271 words.
KEYWORDS: Colombia. Iramuteq. Lexicon. Paleontology.

RESUMEN

Objetivo: Identificar las proximidades léxicas en un corpus de textos de artículos científicos publicados en revistas académicas indexadas en la base de datos Scopus sobre fósiles colombianos
Metodología: Este trabajo aplica el análisis textual a cinco artículos paleontológicos sobre fósiles colombianos, con el propósito de identificar las proximidades léxicas en un corpus de textos. Este trabajo, permitió determinar: las categorías gramaticales, la proximidad entre categorías de palabras y variables con el análisis de especificidades (AE), el

agrupamiento de las palabras con el análisis de la clasificación jerárquica descendiente (CJD) y la presentación gráfica de las palabras.
Resultados: Encontramos, que el corpus documental está conformado por 31.319 ocurrencias de palabras, 1.450 formas activas o palabras específicas y 303 formas complementarios o palabras comunes. Predomina la categoría gramatical de sustantivo (24%) y las palabras no reconocidas en el diccionario (17%). Las palabras comunes con el mayor número de frecuencias son los artículos, las conjugaciones, las proposiciones y los pronombres.
Conclusiones: Se halló que hay proximidad léxica entre el artículo 1 y las formas activas de "Colombia" y el artículo 2 y las formas activas de "fossil". Las palabras se agruparon en cinco clases y la nube de palabras se creó con 1271 palabras.
Palabras-clave: Colombia. Iramuteq. Léxico. Paleontología.

1 INTRODUCCIÓN

El análisis de textos se originó a finales del siglo XIX cuando en 1888 Benjamin Bourdon analizó el libro del Éxodo de la Biblia y calculó frecuencias, reorganizó, clasificó y eliminó palabras vacías. A partir de este estudio se publicaron otros trabajos a principios del siglo XX entre los cuales apareció la ley de Zipf[1] (IEZZI; CELARDO, 2018). También la informática a mediados del siglo XX permitió que el análisis de textos evolucionara e impulsara la implementación del tratamiento automático de textos, es decir, no se requiere la lectura por parte de un individuo para analizar un texto. En la década del 40 apareció la calculadora mecánica y se generalizó su uso en todas las áreas del conocimiento, esta innovación nació de la fusión entre la informática, la lingüística, la estadística y la matemática. En la década de los 50 la perspectiva del análisis cambió y nacen medidas e índices específicos del vocabulario. Es así como en la década del 60 nació el corpus textual, además el matemático y estadístico francés J. P. Benzécri introdujo el análisis de formas gráficas, segmentos repetidos, análisis de correspondencias de tablas léxicas, etc. al estudio de un *corpus* textual, en otras palabras, la estadística al análisis de textos o textometría (IEZZI; CELARDO, 2018).

A partir de la introducción de la estadística al análisis de textos, en la década de los 80 aparecen varios *softwares* para realizar análisis de textos; por ejemplo, SPAD software (*Système Portable pour l'Analyse des Donneés*) (IEZZI; CELARDO, 2018) y Alceste (Análisis de lexemas concurrentes en los enunciados simples de un texto), entre otros. El *software* de Alceste fue desarrollado por Reinert (1983, 1986, 1995, 1998, 2008), quien incorporó las ideas de Benzécri para realizar el análisis textual. El análisis que introdujo Reinert con el software Alceste se basa en el concepto de "mundos lexicales" que es el conjunto de palabras que forman un discurso y están presentes en las "unidades de

[1] "se basa en contar el número de veces que se usa cada palabra en un texto más o menos extenso y ordenar las palabras de las más frecuentes a las menos frecuentes por rangos. Esta tendencia se explica porque siempre es más fácil escribir una palabra conocida que usar una menos conocida" (URBIZÁGASTEGUI ALVARADO; RESTREPO ARANGO, 2011, p. 17).

contexto" (UCI) que son los segmentos de texto, estos segmentos de texto están compuestos por un sujeto y un predicado que normalmente tienen una extensión de varias líneas (CÉSARI, 2017). "Los 'mundos lexicales' [se analizan por medio de] la organización y la distribución de las palabras principales co-ocurrentes en los enunciados simples de un texto" (CÉSARI, 2017, p. 29), estos análisis no tienen en cuenta la sintaxis del discurso, sino la coocurrencia o presencia simultánea de varias palabras funcionales o principales como son los adjetivos, los sustantivos y los verbos en un mismo segmento de texto, elimina del análisis las palabras complementarias o relacionales como artículos, conjunciones, proposiciones, pronombres, etc. En otras palabras, identifica el uso del vocabulario en uno o más textos. El análisis de textos o la textometría[2] se basa en los postulados de Reinert. Esta técnica permite la identificación de vocabularios complementarios y específicos en un área del conocimiento (CÉSARI, 2017).

En la década de los 90 apareció el internet y su uso se generalizó, esto causó una alta circulación de documentos. A partir de esto se desarrollaron técnicas para la extracción de información relevante (IEZZI; CELARDO, 2018). Es así como aparecieron la minería de datos que es una técnica de "descubrimiento de conocimiento" en datos estructurados y bases de datos relacionales y se denomina *Knowledge-Discovery in Databases* (KDD). A partir de la KDD apareció la minería de textos también conocida como *Knowledge-Discovery in Text* (KDT). La KDT es una técnica que permite predecir patrones y tendencias en grandes cantidades de texto no estructurados en lenguaje natural. Tanto la minería de datos como la minería de textos son técnicas de análisis de información (GÁLVEZ, 2008). Para Mariñelarena-Dondena; Errecalde y Castro Solano (2017), la minería de textos permite predecir y describir características extraídas de los documentos. La descripción intenta obtener patrones que explican o sintetizan las relaciones en los datos. En el caso de la literatura científica permite explicar las tendencias temáticas en términos del uso que hacen los investigadores del lenguaje académico y técnico en sus publicaciones. Según Zanjirchi; Abrishami y Jalilian (2019), la minería de textos se usa para identificar los métodos más comunes y clasificar los temas de investigación.

La minería de textos es una técnica que se ha aplicado a diversos campos del conocimiento, sobre todo en la bioinformática (FEBLES RODRÍGUEZ, 2002) para identificar patrones, genes, moléculas, etc. Es una técnica que genera resultados positivos para los investigadores, porque permite el análisis de grandes volúmenes de información. Debido al

[2] Se entiende por textometría la aplicación de "procedimientos de ordenamiento y de cálculos estadísticos para el estudio de un *corpus* de textos digitalizados" (PINCEMIN, 2010, p. 15).

auge de esta técnica se ha aplicado en la investigación de operaciones para clasificar métodos de investigación (ZANJIRCHI; ABRISHAMI; JALILIAN, 2019); en la psicología (MARIÑELARENA-DONDENA; ERRECALDE; CASTRO SOLANO, 2017); en los artículos publicados sobre comercio electrónico para identificar la coocurrencia de las palabras clave (YAN; LEE; LEE, 2015); en la clasificación de material bibliográfica (CONTRERAS BARRERA, 2016); en las redes sociales Facebook y Twitter usadas en la bibliotecología (JARAMILLO VALBUENA, CARDONA; FERNÁNDEZ, 2015); en los estudios sobre internet para identificar las palabras clave más populares (PENG *et al*., 2013); y en la biología molecular y genómica (GÁLVEZ, 2008), entre otras aplicaciones.

La textometría o el análisis textual ha sido aplicado en la historia de la educación matemática en el Brasil (TAISE HOFFMANN; BISSET ÁLVAREZ; MARTÍ-LAHERA, 2020); en el análisis de dominio de los términos Amazonia y Amazon (RAMOS; CARVALHO; SOUZA, 2020); en la temática biblioteca digital en la Ciencia de la Información (FERREIRA; CORRÊA, 2018); en la investigación cualitativa en la enfermería (SOUZA, 2018); en la investigación brasilera en las ciencias de la salud (SALVADOR, 2018); en la filosofía (SPOLAORE; GIARETTA, 2018); en la psicología social en Europa (RIZZOLI, 2018); en el análisis de la colaboración entre investigadores que publican sobre leishmaniasis en Scopus y PubMed (SAMPAIO *et al*. 2017); y en el análisis de entrevistas a miembros de una lista de discusión especializada en ergonomía en Francia (BARCELLINI; DELGOULET; NELSON, 2016), entre otras aplicaciones en diferentes áreas del conocimiento. En la literatura revisada no se encontraron documentos que hayan usado esta técnica para el análisis textual de la literatura publicada sobre fósiles colombianos, la mayoría de los trabajos analizan palabras clave o los títulos de un conjunto de documentos, pero no textos completos.

Por esto el propósito de este artículo es identificar las proximidades léxicas en un corpus de textos de artículos científicos publicados en revistas académicas indexadas en la base de datos *Scopus* sobre fósiles colombianos, con el fin de obtener información de primera mano sobre las relaciones temáticas que se pueden evidenciar con la aplicación de este tipo de análisis. Se intentará dar respuesta a las preguntas: ¿cómo está conformado el corpus documental sobre fósiles colombianos? ¿cuáles son las palabras y categorías gramaticales predominantes? ¿cuáles son los vocabularios comunes y los vocabularios específicos? ¿hay proximidad léxica entre los artículos, las formas activas y las categorías gramaticales que conforman el corpus documental? ¿cómo se agrupan las palabras del corpus documental? ¿cómo se ven las palabras en una nube?

1 METODOLOGÍA

Para construir el corpus documental se usó una muestra de cinco artículos científicos publicados en inglés, los cuales fueron seleccionados en una búsqueda de información que se realizó de enero de 2020 a febrero de 2021 sobre artículos científicos que investigaron los fósiles colombianos, estos artículos científicos fueron publicados en revistas académicas indexadas en la base de datos *Scopus*. Se seleccionaron con base en la importancia que tienen en la paleobiología y paleobiogeografía del neotrópico. Estos cinco artículos se seleccionaron en inglés, porque el diccionario que incluye el software IRAMUTEQ en este idioma es más completo y por ende los resultados serían más completos. Los cinco artículos son:

1. DUQUE-CARO, H. (1990). Neogene stratigraphy, paleoceanography and paleobiogeography in northwest South America and the evolution of the Panama seaway. **Paleogeography, Palaeoclimatology, Palaeoecology**, v. 77, n. 3-4, p. 203-234, 1990. Estudio que desarrolla hipótesis acerca de la temporalidad de la evolución tectónica del Istmo de Panamá mediante el uso de fósiles de foraminíferos y sus relaciones con los ambientes de acumulación de las rocas.

2. VAN DER HAMMEN, T.; WERNER, J. H.; VAN DOMMELEN, H. Palynological record of the upheaval of the Northern Andes: a study of the Pliocene and Lower Quaternary of the Colombian Eastern Cordillera and the early evolution of its high-Andean biota. **Review of Palaeobotany and Palynology**, v. 16, n. 1-2, p. 1-12, 1973. Estudio que desarrolla hipótesis acerca de la temporalidad de la evolución tectónica de la Cordillera Oriental de Colombia y la posible relación de los cambios de altura en el paisaje, en cortos intervalos de tiempo, con los cambios florísticos en los bosques andinos.

3. LUQUE, J.; FELDMANN, R. M.; VERNYGORA, O.; SCHWEITZER, C. E.; CAMERON, C. B.; KERR, K. A; VEGA, F. J.; DUQUE, A.; STRANGE, A.; PALMER, A. R.; JARAMILLO, C. Exceptional preservation of mid-Cretaceous marine arthropods and the evolution of novel forms via heterochrony. **Science Advances**. v. 5, n. 4, p. 1-14, 2019. Estudio que revela el proceso de la heterocronía en cangrejos neotropicales del Cretácico colombiano y aborda hipótesis acerca de la posible incidencia de la heterocronía en la mega-diversidad del grupo de los decápodos.

4. WING, S.L.; HERRERA, F.; JARAMILLO, C.; GÓMEZ NAVARRO, C.; WILF, P.; LABANDEIRA, C. Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest. **Proceedings of the National Academy of Sciences**, v. 106, n. 44, p. 18627–18632, 2009. Estudio que demuestra que el bosque húmedo tropical se generó al menos hace 57 millones de años. Adicionalmente, muestra que la configuración florística de este bioma se ha conservado desde el Paleoceno hasta hoy.

5. JARAMILLO, C.; RUEDA, M.J.; MORA. G. Cenozoic plant diversity in the Neotropics. **Science**, v. 311, n. 5769, p. 1893-1896, 2006. Estudio que determina la relación entre los cambios de clima globales con los cambios de diversidad en el bosque húmedo tropical durante los últimos 57 millones de años. Demuestra que cuando el planeta tiene óptimos climáticos la diversidad en este bioma aumenta y cuando el clima global es frío la diversidad en el bioma disminuye.

Estos cinco artículos se transformaron a texto plano (extensión .txt) con el convertidor en línea disponible en la página web https://www.pdf2go.com/es/pdf-a-text. Cada uno de los cinco archivos se aglutinó en único archivo en el programa Notepad ++. Este archivo fue procesado con el software IRaMuTeQ (Interfaz de R para el Análisis Multidimensional de los Textos y Cuestionarios) (RATINAUD, 2008).

Con el fin de facilitar el procesamiento del corpus documental por el software Iramuteq se construyó el *corpus* teniendo en cuenta que es necesario usar cuatro asteriscos al inicio de cada artículo, así como la palabra “Artículo_Número del artículo” precedida por un asterisco como se muestra en el ejemplo que se presenta a continuación:

*****Articulo_2

_*Articulo2

Neogene stratigraphy, paleoceanography and paleobiogeography in northwest South America and the evolution of the Panama Seaway

Introduction comprises the swampy and forested plains of

the Atrato valley area, at the western side of…

Este proceso se realizó con los cinco artículos para facilitar el análisis de los textos, así como la identificación de las proximidades léxicas, es decir, la cercanía entre las palabras que forman el corpus documental. El proceso que se describe en el ejemplo

permitió establecer las diferencias y similitudes entre artículos y facilitó su graficación en el plano cartesiano.

Se eliminaron gráficos, mapas, referencias bibliográficas, citas bibliográficas, acentos en las palabras, tablas, caracteres especiales como guiones, asteriscos dentro del texto, acentos y comillas. Las palabras compuestas como “Sabana de Bogotá”, “Central América”, “Salto del Tequendama”, entre otras se unieron con un guion bajo (_), con el fin de que el sistema las identificara como una única palabra y no por separado. El software Iramuteq se encargó de realizar los procesos de limpieza de pasar el texto a minúsculas, eliminar los caracteres no incluidos en el software, sustituir apostrofe por espacio, sustituir guion por espacio y puntuaciones.

La lematización se realizó de forma automática por este software, es decir, se estandarizó, desambiguó y segmentó cada forma o palabra. Este proceso lo realizó el sistema con base en el diccionario de palabras del idioma inglés que tiene Iramuteq. Básicamente la lematización convirtió los sustantivos en singular, los adjetivos al masculino singular y las conjugaciones verbales a infinitivo.

Según Morales del Río (2019), el software Iramuteq utiliza la reducción de las unidades del corpus textual, denominadas unidades de contexto elemental (UCE) que son los segmentos de texto. Usa la lematización o la reducción de las palabras a su raíz, crea las unidades de contexto (UC), o reagrupamientos de los segmentos de texto que están formados por un mínimo de formas activas o vocabularios específicos que son verbos, sustantivos y adjetivos y por las formas suplementarias o vocabularios complementarios que son artículos y preposiciones. También utiliza la clasificación jerárquica descendiente (CJD) que genera un clúster creado por un análisis factorial de correspondencias (AFC), entre otras funcionalidades.

Para identificar la estructura o conformación del corpus documental se usó la estadística textual que permitió reconocer la cantidad de palabras, la frecuencia media de palabras con una sola aparición y las formas activas (e.g. verbos, sustantivos, adverbios) y las formas suplementarias (e.g. artículos, proposiciones, conjunciones). También estas estadísticas generaron el diagrama de la ley de Zipf, el cual ilustró la distribución de las palabras dentro del corpus (SALVIATI, 2017).

Para determinar las categorías gramaticales se obtuvo el vocabulario común o suplementario que permitió identificar formas suplementarias y el vocabulario específico que permitió identificar formas activas. A partir de estos datos se obtuvieron informaciones

sobre las frecuencias de estas categorías en cada uno de los artículos que conformaron el corpus documental.

Para analizar la proximidad se utilizó el análisis de especificidades (AE) o análisis de correspondencia (AFC) que permitió ver la proximidad entre categorías de palabras y variables, este cálculo se realizó teniendo en cuenta las palabras con frecuencias iguales o mayores a 40 y la distancia del chi-cuadrado que proporcionó un índice de distancia entre las variables o categorías. Estos datos se representaron en el plano cartesiano que cruzó las variables de formas activas y categorías gramáticas con los artículos que forman el corpus documental (SALVIATI, 2017) para determinar la proximidad. El AE usó el análisis de correspondencias para reducir gran cantidad de información y proyectó las formas activas en un plano cartesiano para ver similitudes y relaciones entre formas de palabras.

Para analizar el agrupamiento de las palabras se utilizó el análisis de la clasificación jerárquica descendiente (CJD) usando el método de Reinert (1983). CJD tiene como propósito obtener clases de segmentos de texto y presentar el vocabulario similar entre cada clase. Este método usa el análisis de correspondencias que compara los perfiles columna y los perfiles fila. En este caso cada fila está representada por cada uno de las clases y las columnas son las categorías gramaticales, la relación entre estas filas y columnas se presentó en un plano cartesiano.

Este análisis se basó en la proximidad léxica y en que las palabras utilizadas en contextos similares están asociadas con el mismo “mundo léxico” y son parte de mundos mentales específicos o sistemas de representación. En este análisis, los segmentos de texto se clasificaron de acuerdo con su vocabulario respectivo y el conjunto de términos se dividió de acuerdo con la frecuencia de las raíces de la palabra. El sistema buscó obtener clases formadas por palabras que están próximas, es decir, usó la distancia del chi-cuadrado (SALVIATI, 2017).

Según Césari (2017, p. 59), CJD permite “la determinación de grupos de palabras que suelen ser empleadas por los mismos individuos y que delimitan, por tanto, campos semánticos o temáticas conectadas entre sí”. Finalmente, para determinar la presentación gráfica de las palabras se usó la nube de palabras, esta representó las palabras en un gráfico y destacó en el centro del gráfico las palabras con mayores frecuencias (SALVIATI, 2017), es decir, las más usadas en el corpus documental.

2 RESULTADOS Y DISCUSIÓN

El corpus textual, compuesto por cinco textos, tiene 870 segmentos de texto que están formados por la ocurrencia de 40 unidades lingüísticas (e.g. letras, sonidos, sílabas, fonemas), 31.319 palabras, 4.233 formas de palabras, 1.450 formas de palabras activas o vocabularios específicos (verbos, adjetivos, adverbios, sustantivos y palabras no encontradas en el diccionario), 303 formas de palabras complementarias o comunes (artículos, proposiciones, conjunciones) y 1.913 hápax o palabras con una única aparición (45.19% de las formas o palabras, 6.11% de las ocurrencias o número total de palabras) Las formas activas que tienen una frecuencia igual o mayor a 60 y con frecuencias de 197 hasta 61 se presentan en la Tabla 1.

Tabla 1 - Formas activas con mayor frecuencia

Palabra	Frecuencia	Categoría gramatical
formation	197	Sustantivo
zone	140	Sustantivo
area	131	Sustantivo
miocene	127	Palabra no reconocida
group	110	Sustantivo
middle	106	Sustantivo
pollen	102	Sustantivo
type	96	Sustantivo
part	84	Sustantivo
section	78	Sustantivo
basin	73	Sustantivo
benthic	74	Palabra no reconocida
sample	72	Sustantivo
forest	72	Sustantivo
crab	68	Sustantivo
upper	66	Sustantivo
clay	63	Sustantivo
nw	62	Palabra no reconocida
time	61	Sustantivo

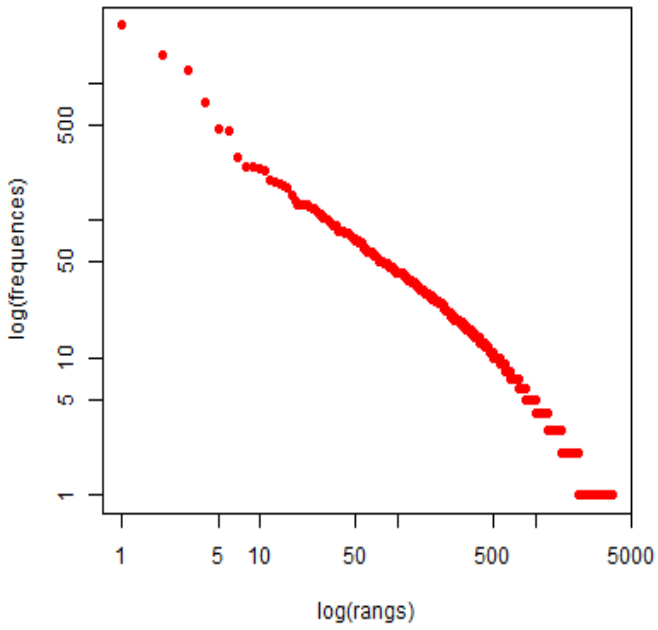
Fuente: autoría propia

Se encontraron 1.450 formas activas, de las cuales 522 palabras tienen una única frecuencia (36%), 228 palabras con dos frecuencias (16%), 145 palabras con tres frecuencias (10%), 92 palabras con cuatro frecuencias (6%) y 85 palabras con cinco frecuencias (5%), y así sucesivamente hasta la frecuencia más alta de la forma activa con 197. También se encontraron un total de 303 palabras complementarias, de las cuales 10 palabras con la frecuencia más alta son artículos (“the” con 2641 y “a” con 472),

conjunciones (“and” con 1258), proposiciones (“of” con 1583, “to” con 445, “in” con 729 y “with” con 250) y pronombres (“is” con 292, “are” con 250 y “this” con 238), representan 1% del total de palabras. Estas palabras forman parte de los segmentos de palabras, es decir, conectan las distintas categorías gramaticales para dar sentido a un texto. Estas palabras por sí solas no tienen sentido, porque simplemente tienen la función de conectar las palabras que forman una oración. También en las palabras activas se encontraron formas con menores frecuencias, aunque se esperaría que en este corpus textual palabras como: “tropical” (59), “sediment” (58), “specie” (58), “diversity” (55), “fossil” (51), “vegetatation” (49), “deposit” (45), “pliocene” (42), “South America” (42), “water” (42), “Colombia” (41), “Cerrejón” (41), “Subachoque” (41), tuviesen mayores frecuencias de aparición en el corpus. También llama la atención que palabras como “tropical”, “pliocene”, “South America”, “Colombia”, “Cerrejón” y “Subachoque” no son palabras que estén en el diccionario de Iramuteq, estas palabras son sustantivos al igual que la mayoría de las palabras listadas.

El diagrama de Zipf presenta el comportamiento de las frecuencias de palabras en el corpus textual mostrando que las palabras que tienen una frecuencia mayor a 500 son: the (frecuencia 2641), of (frecuencia 1593), and (frecuencia 1258) e in (frecuencia 729) (Figura 1). Adicionalmente, se encontraron 33.507 ocurrencias, 3.768 formas lematizadas y 1.580 hápax lematizadas o palabras que aparecen una única vez (4.72% de ocurrencias y 41.93% de formas) y una media de ocurrencia por texto de 6.701 palabras (es el resultado de la división entre el número de ocurrencias /número de textos). La frecuencia de aparición de las palabras oscila entre 2.641 y una ocurrencia.

Figura 1 - Frecuencia de aparición de palabras para el corpus textual estudiado



Fuente: autoría propia

El análisis de especificidades o análisis factorial de correspondencias (AFC) permitió identificar las categorías gramaticales de las palabras que aparecen en cada uno de los artículos (Tabla 2).

Tabla 2 - Categorías gramaticales según los artículos científicos

Categorías gramaticales	**Articulo 1**	**Articulo 2**	**Articulo 3**	**Articulo 4**	**Articulo 5**
Sustantivo (nom)	2169	2966	1344	822	174
Verbo (ver)	485	738	477	187	37
Segmentos de palabras (sw)	4019	6516	2234	1207	280
Suntantivos complementarios (num)	285	420	192	164	47
Palabras no reconocidas (nr)	2045	1688	1001	408	87
Adjetivo (adj)	420	566	195	124	22

Fuente: autoría propia

Las categorías gramaticales con el mayor porcentaje son los segmentos de palabras con 46% que son considerados los sonidos individuales que conforman una palabra, sustantivos (24%) se usan para nombrar a sujetos u objetos y palabras no reconocidas en el diccionario (17%). Mientras que algunas otras categorías gramaticales como verbo (6%) que expresa acción, proceso, estado o existencia que afecta al sujeto, sustantivo complementario (4%) que se refiere a aspectos que complementan al sustantivo y adjetivo

(4%) que se usa para calificar al sustantivo, es decir, indica cualidad. También los artículos científicos con la mayor extensión son el artículo 2 con 12.894 palabras y el artículo 1 con 9.423 palabras, mientras que el artículo 3 tiene 5.443 palabras, el artículo 4 tiene 2.912 palabras y el artículo 5 tiene apenas 647 palabras (Tabla 2).

Sin duda el número de palabras por categoría gramatical está relacionada con la extensión de los artículos en términos de número de palabras, pero estos valores no tienen ninguna relación con el número de citas que ha recibido cada uno de estos documentos ni tampoco con el año de publicación. Este número de palabras por categoría obedece aspectos de índole normativos de las revistas científicas que seleccionaron los autores para publicar, al igual que elementos cognitivos de cada investigador. Esto significa que en este artículo no se intenta a ahondar sobre este asunto, simplemente mostrar las diferentes categorías de palabras que usa el análisis textual.

Las palabras no reconocidas en el diccionario son palabras que están en el corpus, pero no están incluidas en el diccionario en idioma inglés que tiene el software con el cual ejecuta el análisis del corpus. En general las formas no reconocidas en el caso del corpus de fósiles colombianos son palabras técnicas utilizadas en paleontología (e.g. intervalos de tiempo geológico, nombres y hábitos de vida de los organismos, localidades donde se encuentran los fósiles, palabras compuestas (normalmente unidas con un guion), abreviaturas de medidas o geográficas. Las formas no reconocidas con una frecuencia mayor a 30 se listan en la Tabla 3.

Tabla 3 - Palabras no reconocidas en el diccionario

Palabra	Frecuencia	Categoría gramatical
miocene	127	Sustantivo
benthic	74	Sustantivo
nw	62	Abreviatura (noroeste)
tropical	59	Adjetivo
cerrejon	41	Sustantivo
south_america	42	Sustantivo
pliocene	42	Sustantivo
subachoque	41	Sustantivo
colombia	41	Sustantivo
Subachoque	41	Sustantivo
ft	38	Abreviatura (unidad de medida pie)
caribbean	37	Sustantivo
choconta	36	Sustantivo
coastal	35	Adjetivo
planktic	35	Sustantivo
foraminiferal	32	Adjetivo

Palabra	Frecuencia	Categoría gramatical
pacific	31	Adjetivo
guasca	31	Adjetivo
bogota	31	Sustantivo

Fuente: autoría propia

La Figura 2 muestra el análisis de especificidades o AFC que asoció las formas léxicas con los artículos y las categorías gramaticales de verbos, sustantivos, palabras no reconocidas en el diccionario y segmentos de palabras. Se definió como criterio para establecer la similitud entre formas de palabras la distancia Chi-cuadrado y la frecuencia de ocurrencia de las formas activas igual o mayor a 40. La Tabla 4 muestra las medidas de masa, distancia Chi e inercia que usa el análisis factorial de correspondencias de las categorías (el sistema convirtió los artículos en clases o categorías).

Tabla 4 - Medidas de análisis de correspondencia para artículos

Clase	Masa	Distancia Chi	Inercia
Clase 1	0,30087168	0,13998165	0,00589554
Clase 2	0,41169897	0,11911132	0,00584098
Clase 3	0,17379227	0,13635527	0,00323128
Clase 4	0,0929787	0,16865917	0,00264486
Clase 5	0,02065839	0,2290484	0,0010838

Fuente: autoria propia

El artículo 1 y el artículo 2 son los que tienen el mayor peso, son las variables más importantes en este análisis. El artículo 2 y el artículo 3 son los que tienen la menor distancia, esto significa que las palabras que conforman estos artículos están cercanas. El artículo 1 y el artículo 2 son los que tienen el mayor valor de inercia (Tabla 4).

La Tabla 5 muestra las medidas de masa, Chi-cuadrado e inercia que usó el análisis factorial de correspondencias de las categorías gramaticales.

Tabla 5 - Medidas de análisis de correspondencia para categorías gramaticales

Categoría gramatical	Masa	Distancia chi	Inercia
Sustantivo (nom)	0,23867301	0,06755966	0,00108938
Verbo (ver)	0,06143236	0,20430895	0,00256432
Segmentos de texto (sw)	0,45518695	0,09309982	0,00394537
Sustantivo numeral (num)	0,03537789	0,25381043	0,00227903

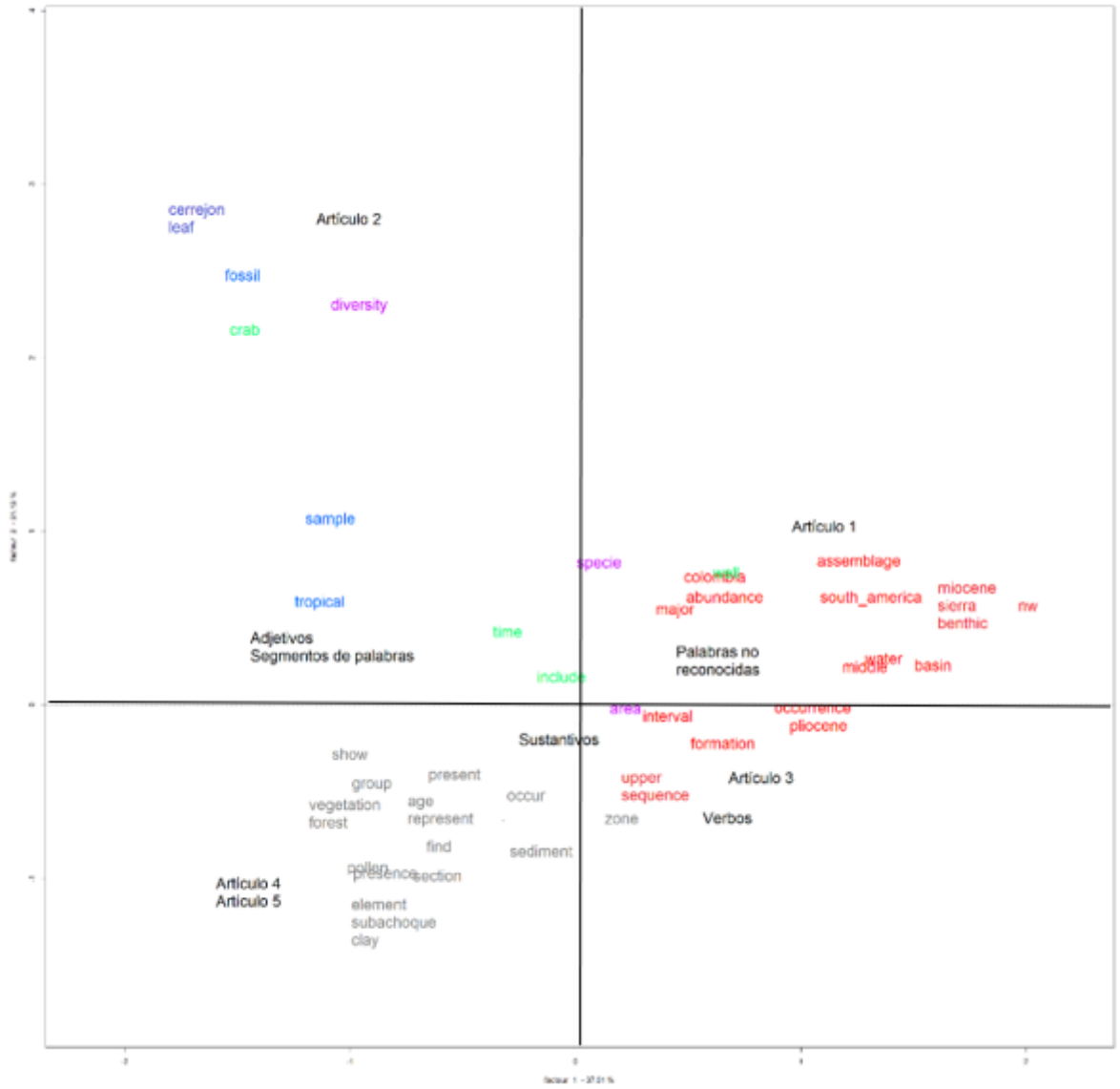
Palabras no reconocidas (nr)	0,16695935	0,22631513	0,00855141
Adjetivos (adj)	0,04237045	0,0793758	0,00026696

Fuente: autoria propia

Las categorías gramaticales con el mayor peso son segmentos de texto (sw), sustantivo (nom) y palabras no reconocidas (nr), esto significa que son las categorías más importantes en este análisis. El sustantivo (nom), los adjetivos (adj) y los segmentos de texto son las categorías gramaticales que tienen la menor distancia, es decir, las palabras que conforman estas categorías son las que están más próximas. Las palabras no reconocidas (nr), segmentos de texto (sw) y sustantivo numeral (num) son los que tienen los mayores valores de inercia (Tabla 5).

En cuanto a la proximidad entre artículos y formas activas se puede distinguir que el artículo 1 (Neogene stratigraphy, pale oceanography and paleobiogeography in northwest South America and the evolution of the Panama seaway) se localiza en la parte superior derecha y está asociado con las formas activas de “Colombia”, “South America”, “miocene”, “benthic” y “nw”, etc., estas palabras tienen en común que no están en el diccionario de Iramuteq. El artículo 2 (Palynological record of the upheaval of the Northern Andes: a study of the Pliocene and Lower Quaternary of the Colombian Eastern Cordillera and the early evolution of its high-Andean biota) se localiza en la parte superior izquierda y está asociado con las formas activas de “fossil”, “Cerrejón”, “leaf”, “crab”, “diversity”, etc., estas palabras tienen en común que son sustantivos, excepto “Cerrejón” que no está en el diccionario (Figura 2).

Figura 2 - Análisis factorial de correspondencias mostrando proximidades entre palabras y categorías gramáticas con los artículos



Fuente: autoría propia

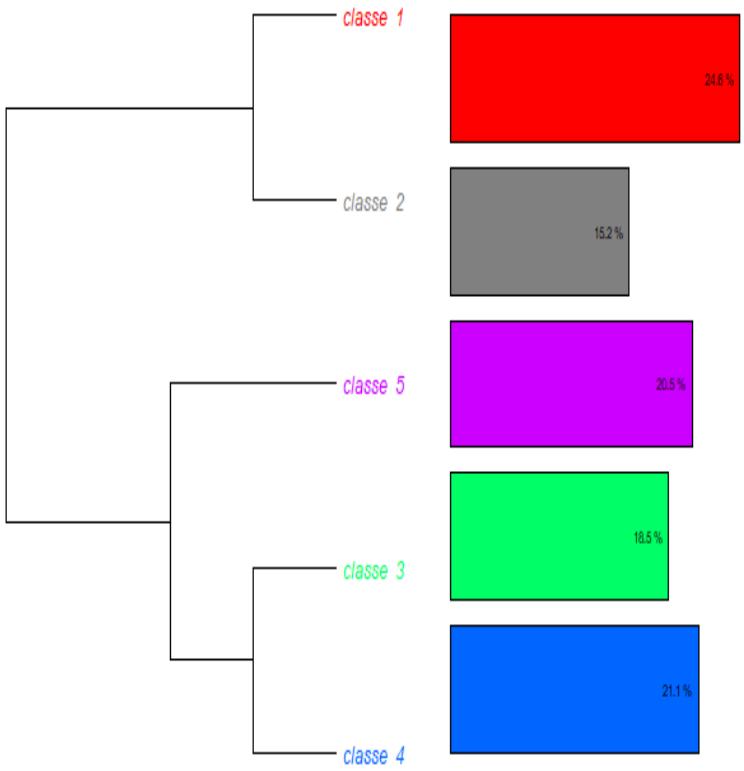
El artículo 3 (Exceptional preservation of mid-Cretaceous marine arthropods and the evolution of novel forms via heterochrony) se localiza en la parte inferior derecha y está asociado con las formas activas de “sequence”, “area”, “interval”, “upper”, “pliocene” etc., estas palabras tienen en común que son sustantivos, excepto “pliocene” que no está en el diccionario, aunque es sustantivo. El artículo 4 (Late Paleocene fossils from the Cerrejón Formation, Colombia, are the earliest record of Neotropical rainforest) y el artículo 5 (Cenozoic plant diversity in the Neotropics) se asocian con las formas activas de “vegetation”, “Subachoeque”, “forest”, “clay”, “pollen” y “presence”, etc., estas palabras tienen en común que son sustantivos, excepto “Subachoque” que no está en el diccionario (Figura 2).

En cuanto a la asociación entre artículos y categorías gramáticas, el análisis factorial de correspondencias muestra que el artículo 1 se asocia con palabras no reconocidas en el diccionario, el artículo 2 con adjetivos y segmentos de palabras y finalmente el artículo con verbos. Por otro lado, los artículos 4 y 5, el artículo 4 y el artículo 5 se asocian con sustantivos y sustantivos numerales. Adicionalmente, este análisis, también muestra que hay dispersión entre las palabras de los cinco artículos, solamente se encontró proximidad léxica entre el artículo 2 y el artículo 3 (Figura 2). En general, entre los cinco artículos no se encontró proximidad léxica o "mundos lexicales" comunes, a pesar de que tienen un tema en común que son los fósiles colombianos, estos artículos investigan sobre este tema desde diferentes subtemas o intereses de los investigadores.

La mayoría de las palabras que están representadas en la Figura 2 son sustantivos y algunas representan vocabularios especializados sobre fósiles colombianos; por ejemplo, "Colombia", "Cerrejón", "fossil", "miocene", "pliocene", "vegetation", "pollen", "crab" y "forest", entre otras. Estas palabras representan el vocabulario especializado usado por los paleontólogos sobre la investigación de los fósiles en Colombia, aunque se encontraron palabras especializadas en fósiles colombianos que solamente se usaron una única vez y no están representadas en la Figura 2; por ejemplo "Chinquinquirá", "etimilogy", "glauconitic", "helicon", "megadiverse", "megaflora", "ontogenetic", "quaternary", "river_Bogotá", "Sogamoso", "topographic", "Turbo", "Ubaté", "zoogeographic", "zeolitic", etc.

La clasificación jerárquica descendiente (CJD) sigue el método propuesto por Reinert (1983). La CJD clasificó las palabras en grupos y generó el dendograma de clase. Este dendograma muestra las clases lexicales, cada clase representa una temática y el vocabulario que las define. Cada clase está representada por un color diferente, y las unidades de contexto elemental (palabras) tienen el mismo color que la clase (Figura 3). El corpus está compuesto por cinco textos, 870 segmentos de texto, 4.233 formas, 31.319 ocurrencias, 3.620 lemas, 1.450 formas activas, 303 formas suplementarias, 1.144 formas activas con una frecuencia >=3:700, media de formas por segmento es de 35.998851, número de clases 5 y 658 segmentos clasificados en 870 (75.63%).

Figura 3 - Dendograma de clases

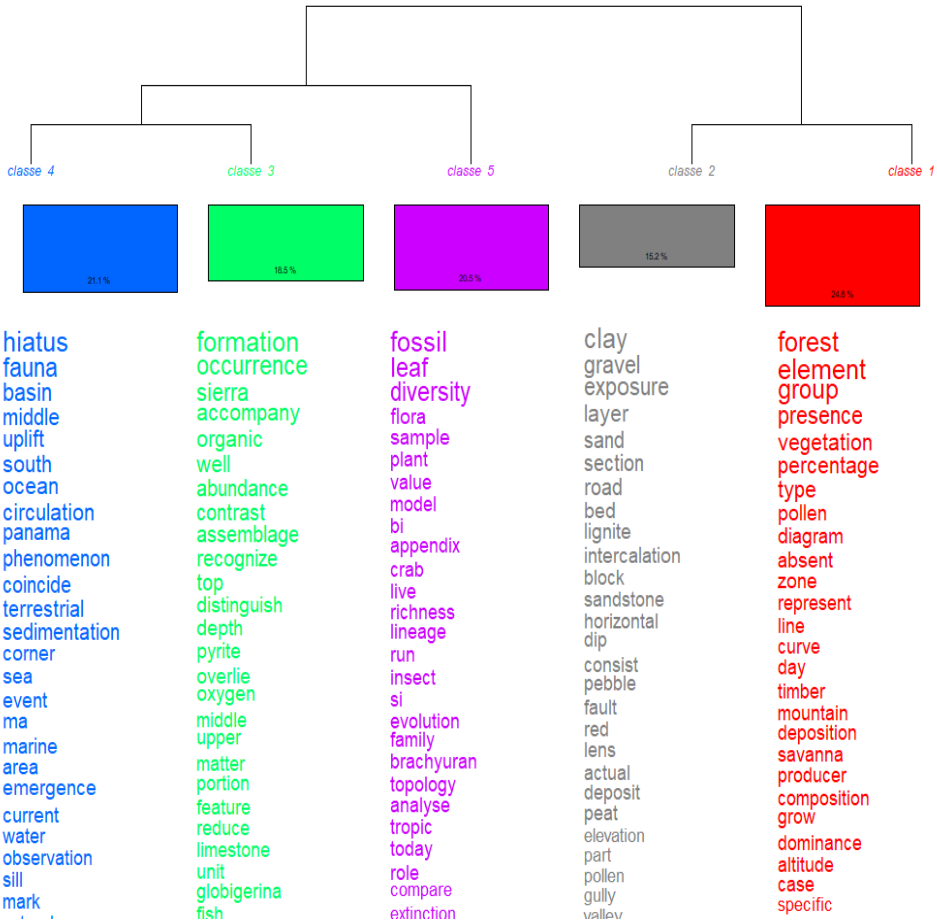


Fuente: autoría propia

El corpus documental incluyó un total de 658 segmentos de texto clasificados en cinco clases. Este dendograma dividió las cinco clases en dos subcorpus. En el primer subcorpus se obtuvo la clase 1 que incluyó 162 segmentos de texto (24.6%) y la clase 2 incluyó 100 segmentos de texto (15.2%). El segundo subcorpus se subdividió en la clase 5 que incluyó 135 segmentos de texto (20.6%), la clase 3 que incluyó 122 segmentos de texto (18.5%) y la clase 4 que incluyó 139 segmentos de texto (21.1%) (Figura 3).

La Figura 4 muestra los segmentos de texto de cada una de las clases, de acuerdo con el agrupamiento y la proximidad lexical entre cada palabra. Con este dendograma fue posible visualizar las palabras que obtuvieron una menor distancia del Chi-cuadrado.

Figura 4 - Dendograma del corpus documental



Fuente: autoría propia

Este dendograma es el resultado del análisis factorial de una matriz de coocurrencias que cruza las filas que son las formas activas extraídas de los segmentos de texto (enunciados compuestos por un sujeto y un predicado que están unidos por preposiciones, conjunciones, artículos, etc., es decir, las formas complementarias) con las columnas que son las clases que representan los UCE o segmentos de texto. Este dendograma presenta las 26 formas activas de cada una de las cinco clases del total de 748 formas activas (Figuras 3 y 4).

Según Reinert (1995), la CJD discrimina los “mundos lexicales” y muestra los vínculos entre formas específicas y las clases o agrupamientos de palabras que tienen relaciones entre cada clase, o bien, entre las palabras de una misma clase. Por ejemplo, en la clase 1, las formas “forest”, “vegetation” y “pollen” representan un “mundo lexical” relacionado con los fósiles vegetales. En la clase 2 las formas “sandstone”, “clay”, “sand” y “graval” representan un “mundo lexical” relacionado con las rocas. En la clase 3 las formas

“limestones” y “pyrite” representan los mundos lexicales relacionados con minerales y rocas. En la clase 4 las formas “ocean”, “water”, “sea” y “marine” representan un “mundo lexical” relacionado con fósiles marinos. En la clase 5 las formas “insect” y “crab”, así como “flora”, “leaf” y “plant” representan los “mundos lexicales” relacionados con flora y fauna. Llama la atención las relaciones entre la clase 2 y en la clase 1 con la forma “pollen” que es común en ambas clases, lo que significa que el “mundo lexical” de esas dos clases está relacionado con la fauna fósil. En las clases 3, 4 y 5 no se encontraron formas comunes relacionadas entre las tres clases, pero si palabras relacionadas temáticamente que representan “mundos lexicales” comunes; por ejemplo, “fauna”, “flora”, “globigerina”, etc.

Para representar las clases en el plano cartesiano se tienen en cuenta las medidas de masa que es la ponderación de cada una de las cinco clases y la distancia Chi-cuadrado para medir la cercanía en el plano cartesiano.

Tabla 6 - Medidas de masa e inercia

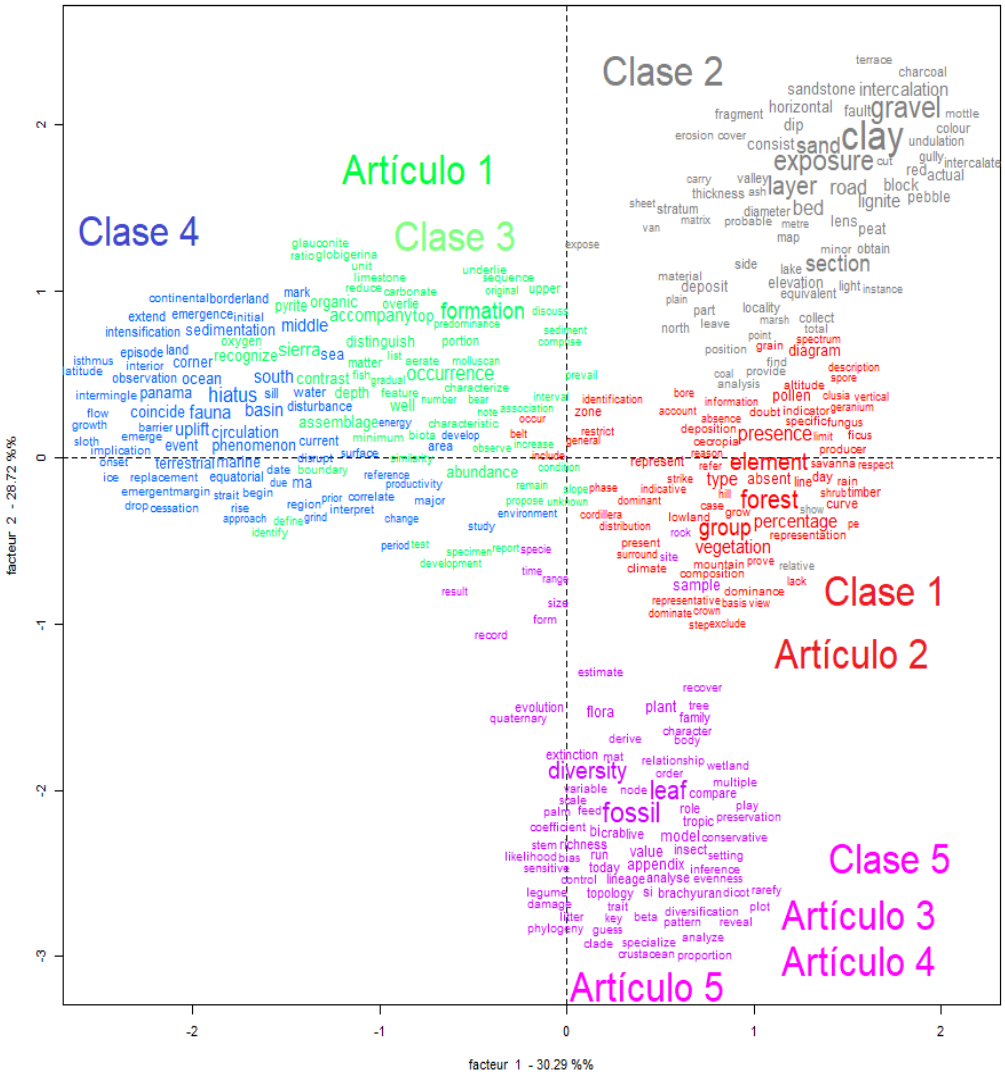
Clases	Masa	Distancia Chi	Inercia
Clase 1	0,22239775	1,24073921	0,34236662
Clase 2	0,1716693	1,52677634	0,40016904
Clase 3	0,21713182	1,25367033	0,34126376
Clase 4	0,17114271	1,31880318	0,29765855
Clase 5	0,21765842	1,32820533	0,38397761

Fuente: autoría propia

Las clases que tiene el mayor peso son la clase 1, la clase 2 y la clase 5, es decir, que son las clases más importantes en este análisis. La clase 1, la clase 3 y la clase 4 son las clases que tienen la menor distancia. La clase 1, la clase 2 y la clase 5 son los que tienen los mayores valores de inercia (Tabla 6).

La proximidad temática entre formas activas, clases y artículos (Figura 5) muestra que las palabras que forman el subcorpus uno son las clases 1 y 2 que identifican la posición del artículo 2 que se publicó en la década del 70. El subcorpus dos lo forman la clase 5 que identifica la posición de los artículos 3, 4 y 5, estos artículos se publicaron entre 2006 y 2019. La clase 3 identifica la posición del artículo 1 que se publicó en la década del 90 y la clase 2 identifica la posición del artículo 2 que se publicó en la década del 70.

Figura 5 - Análisis factorial de correspondencias



Fuente: autoría propia

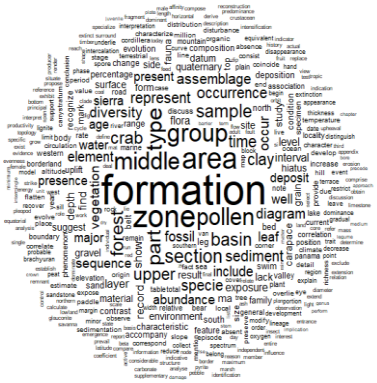
Este AFC muestra las proximidades de los “mundos lexicales” que representan cada clase. En la clase 1 sobresalen las formas activas de “forest”, “element” y “group”. En la clase 2 sobre salen las formas activas “clay”, “gravel” y “exponsure”. En la clase 3 sobre salen las formas activas de “formation” y “occurrence”. En la clase 4 sobresalen las formas activas de “hiatus”, “fauna” y “basin”. En la clase 5 sobre salen las formas activas de “diversity”, “leaf” y “fossil”. En otras palabras, en la Figura 5 se observan las relaciones temáticas que existen entre cada clase y que fueron descritas en el análisis de la Figura 4. Si bien es cierto la Figura 5 muestra dispersión entre el total de las cinco clases, también muestra las proximidades entre clases y artículos; por ejemplo, la clase cinco se asocia con el artículo 3, artículo 4 y artículo 5 que muestran proximidad entre los “mundos lexicales”

de los autores de estos artículos que fueron publicados entre 2006 a 2019. La clase 3 con el artículo 1 y la clase 1 con el artículo 2.

Esos “mundos lexicales” que son “[el] discurso [de un autor, es decir], los mundos que construye para establecer sus puntos de vista y construir su obra” (REINERT, 1993, p. 37), en otras palabras, el lenguaje que utilizó para plasmar los resultados de sus investigaciones. Ese discurso que se traduce en unidades lingüísticas que forman enunciados y que le van dando forma al discurso para conectar al autor con el lector, este último agente analiza e interpreta ese discurso y lo usa para construir sus propios “mundos lexicales”. Entre los “mundos lexicales” de los autores de los documentos que forman el corpus documental no hay una construcción aparentemente similar, o bien, no se observan asociaciones entre el total de las cinco clases y los cinco artículos, a pesar de que el tema de estos documentos son los fósiles colombianos. Esto significa que las construcciones del discurso difieren y que no necesariamente por tratar una misma temática deberían tener en común idénticos “mundos lexicales”.

Finalmente, la nube de palabras (Figura 6) facilita la visualización de los términos más utilizados en el corpus. Se excluyeron los términos de las clases gramaticales: adverbios, artículos, conjunciones, onomatopeyas y preposiciones. Se seleccionaron las formas activas que corresponden a sustantivos y verbos que se muestran en esta Figura. Las formas activas que se usaron son 1271 palabras con una variación en la frecuencia entre 3 a 290. Las palabras que aparecen con mayor tamaño y que se ubican en el centro de la Figura y son las que tienen la mayor frecuencia; por ejemplo, “formation”, “pollen”, “America”, “miocene”, etc. Mientras que las palabras con menor tamaño tienen menores frecuencias.

Figura 6 - Nube de palabras



Fuente: autoría propia

CONCLUSIONES

El análisis textual constituye una herramienta que permite comparar la proximidad léxica entre varios documentos, mediante un software especializado como es el caso de Iramuteq. La comparación textual muestra tres elementos que aportan información a la ciencia de la información, pero también al campo del conocimiento al cual pertenecen los documentos seleccionados para hacer dicho análisis.

El primer elemento caracteriza al corpus documental, en el caso de los fósiles colombianos tiene 31.319 ocurrencias de palabras, 1.450 formas activas o palabras específicas y 303 formas complementarios o palabras comunes. Las palabras que tienen una mayor frecuencia de coocurrencia en el corpus son: “formation” (197), “miocene” (167), “zone” (140), “area” (131), “group” (110), “middle” (106) y “pollen” (102). Predomina la categoría gramatical de sustantivo (24%) y las palabras no reconocidas en el diccionario (17%), esta última categoría corresponde a las palabras especializadas en paleontología, lugares geográficos, palabras compuestas y unidas con el guion, abreviaturas de medidas o geográficas.

El segundo elemento evidencia los vocabularios específicos y comunes. El vocabulario específico son las palabras especializadas en paleontología que no tienen una frecuencia de coocurrencia alta como los vocabularios comunes; por ejemplo, “miocene”. El vocabulario común lo constituyen las palabras que unen las palabras de los “mundos lexicales”, estas palabras son artículo, conjunciones, proposiciones, etc. En el caso de las palabras comunes con la mayor frecuencia son los artículos the (2641) y a (472), las conjunciones and (1258), las proposiciones of (1583), to (445), in (729) y with (250) y los pronombres is (292), are (250) y this (238).

El tercer elemento muestra la proximidad léxica. En el caso del corpus textual de los fósiles colombianos mostró proximidad entre el artículo 1 y las formas activas de “Colombia”, “South America”, “Miocene”, “benthic” y “NW”, etc., entre el artículo 2 y las formas activas de “fossil”, “Cerrejón”, “leaf”, “crab”, “diversity”, etc., entre el artículo 3 y las formas activas de “sequence”, “area”, “interval”, “upper”, “Pliocene” etc., entre el artículo 4 y el artículo 5 “vegetation”, “Subachoeque”, “forest”, “clay”, “pollen” y “presence”, etc.

El análisis textual muestra un conjunto de palabras que están relacionadas, pero no necesariamente muestra que “fósiles” y “Colombia” sean las palabras que predominan en el análisis. Permite comprobar cuáles son esas proximidades entre cinco artículos que

tienen por objeto de investigación en el mismo asunto; sin embargo, no lo tratan de la misma manera ni tampoco encontraron los mismos hallazgos.

REFERENCIAS BIBLIOGRÁFICAS

BARCELLINI, Flore; DELGOULET, Catherine; NELSON, Julien. Are online discussions enough to constitute communities of practice in professional domain? a case study of ergonomics' practice in France. **Cognition, Technology & Work**, v. 18, n. 2, p. 249-266, 2016.

BENZÉCRI, Jean Paul. L' Analyse des Correspondances. En : **L'Analyse des Données**, Tomo II. 2de. Éd. París: Dunod, 1976.

CÉSARI, Matilde Inés. **Protocolo de análisis de datos textuales aplicados a la minería de textos**. 1a edición para el alumno. Ciudad Autónoma de Buenos Aires: Universidad Tecnológica Nacional. Facultad Regional Mendoza, 2017.

CONTRERAS BARRERA, Marcial. Minería de texto en la clasificación de material bibliográfico. **Biblios**, n. 64, p. 33-43, 2016.

FEBLES RODRÍGUEZ, Juan Pedro; GONZÁLEZ PÉREZ, Abel. Aplicación de la minería de datos en la bioinformática. **Acimed**, v. 10, n. 2, p. 69-76, 2002.

FERREIRA, Márcio Henrique Wanderley; CORRÊA, Renato Fernandes. Estudo métrico sobre biblioteca digital: uso do software Iramuteq. En: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 19, 2018, Londirna, **Anais** [...], Londrina: UEL, 2018.

GÁLVEZ, Carmen. Minería de textos: la nueva generación de análisis de literatura científica en biología molecular y genómica. **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, v. 13, n. 25, p. 1-14, 2008.

IEZZI, D. F.; CELARDO, L. Text Analytics: Present, Past and Future. In **INTERNATIONAL CONFERENCE ON THE STATISTICAL ANALYSIS OF TEXTUAL DAT**A (pp. 3-15). Cham: Springer, 2018.

IRAMUTEQ. **Bienvenida**. 2021. Disponible en: http://www.iramuteq.org/ Acceso en: 21 may. 2022.

JARAMILLO VALBUENA, Sonia; CARDONA, Sergio Augusto; FERNÁNDEZ, Alejandro. Minería de datos sobre streams de redes sociales, una herramienta al servicio de la Bibliotecología. **Información, cultura y sociedad**, n. 33, p. 63-74, 2015.

MARIÑELARENA-DONDENA, Luciana; ERRECALDE, Marcelo Luis; CASTRO SOLANO, Alejandro. Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología. **Revista Argentina de Ciencias del Comportamiento**, v. 9, n. 2, p. 65-76, 2017.

MORALES DEL RÍO, Cecilia. M. Uso de software lexical: una revisión comparativa. En: CONGRESO DE LA RED INTERNACIONAL DE INVESTIGADORES EN COMPETITIVIDAD, 13, 2019, **Anais** [...], 2019. Disponible en: https://riico.net/index.php/riico/article/view/1794 Acceso en: 21 may. 2022

PENG, T. Q.; ZHANG, L.; ZHONG, Z. J.; ZHU, J. J. Mapping the landscape of Internet studies: Text mining of social science journal articles 2000–2009. **New Media & Society**, v. 15, n. 5, p. 644-664, 2013.

REINERT, Max. Une méthode de classification descendante hiérarchique: application à l'analyse lexicale par contexte. **Cahiers de l'Analyse des Données**, v. 8, n. 2, p. 187-198, 1983.

REINERT, Max. Un logiciel d'analyse lexicale. **Cahiers de l'Analyse des Données**, v. 11, n. 4, p. 471-481, 1986.

REINERT, Max. Quelques aspects du choix des unités d'analyse et de leur contrôle dans la méthode Alceste. **Journées Internationales d´Analyse Statistique des Données Textuelles (JADT),** v. 1, p. 27-34, 1995.

REINERT, Max. Quel objet pour une analyse statistique du discours? Quelques réflexions à propos de la réponse Alceste. **Journées Internationales d´Analyse Statistique des Données Textuelles (JADT)**. Lexicometrica, p. 557-569, 1998.

REINERT, Max. Mondes lexicaux stabilisés et analyse statistique de discours. Actes de la JADT. **Journées Internationales d´Analyse Statistique des Données Textuelles (JADT)**. p. 981-993, 2008

RIZZOLI, Valentina. Histories of Social Psychology in Europe and North America, as Seen from Research Topics in Two Key Journals. In: **Tracing the Life Cycle of Ideas in the Humanities and Social Sciences** (pp. 65-86). Cham: Springer, 2018.

TAISE HOFFMANN, Yohana; BISSET ALVAREZ, Edgar; MARTÍ-LAHERA, Yohannis. Análise textual com IRaMuTeQ de pesquisas recentes em História da educação matemática no Brasil: um exemplo de Humanidades Digitais. **Investigación bibliotecológica**, v. 34, n. 84, p. 103-133, 2020.

URBIZAGÁSTEGUI ALVARADO, Rubén; RESTREPO ARANGO, Cristina. La ley de Zipf y el punto de transición de Goffman en la indización automática. **Investigación bibliotecológica**, v. 25, n. 54, p. 1-15, 2011.

YAN, B. N.; LEE, T. S.; LEE, T. P. Analysis of research papers on E-commerce (2000–2013): based on a text mining approach. **Scientometrics**, v. 105, n. 1, p. 403-417, 2015.

SALVADOR, Pétala Tuani Candido de Oliveira; GOMES, Andréa Tayse de Lima; RODRIGUES, Cláudia Cristiane Filgueira Martins; CHIAVONE, Flávia Barreto Tavares; ALVES, Kisna Yasmin Andrade; BEZERRIL, Manacés dos Santos; SANTOS, Viviane Euzébia Pereira. Uso do software iramuteq nas pesquisas brasileiras da área da saúde: uma scoping review. **Revista Brasileira em Promoção da Saúde**, n. 31, 2018.

SALVIATI, Maria Elisabeth. **Manual do Aplicativo Iramuteq**: (versão 0.7 Alpha 2 e R Versão 3.2.3). Planaltina: [Sin editor], 2017.

SAMPAIO, Ricardo B.; FONSECA, Bruna P.; BAHULKAR, Ashwin; SZYMANSKI, Boleslaw K. Network analysis to support public health: evolution of collaboration among leishmaniasis researchers. **Scientometrics**, n. 111, v. 3, 2001-2021, 2017.

SPOLAORE, Giuseppe; GIARETTA, Pierdaniele. Tracing the Words of the Analytic Turn in the Journal of Philosophy. In: **Tracing the Life Cycle of Ideas in the Humanities and Social Sciences** (pp. 25-44). Cham: Springer, 2018.

SOUZA, Marli Aparecida Rocha de; WALL, Marilene Loewen; THULER, Andrea Cristina de Morais Chave; LOWEN, Ingrid Margareth Voth; PERES, Aida Maris. O uso do software IRAMUTEQ na análise de dados em pesquisas qualitativas. **Revista da Escola de Enfermagem da USP**, v. 52, 2018.

ZANJIRCHI, Seyed Mahmoud; ABRISHAMI, Mina; JALILIAN, Negar. Four decades of fuzzy sets theory in operations management: application of life-cycle, bibliometrics and content analysis. **Scientometrics**, v. 119, n. 3, 1289-1309, 2019.

NOTAS

AGRADECIMENTOS
Se agradece a Daniel Alejandro Díaz Restrepo por el apoyo técnico en la instalación de Iramuteq.

CONTRIBUIÇÃO DE AUTORIA
Concepção e elaboração do manuscrito: C. Restrepo-Arango, A. L. Cárdenas-Rozo
Coleta de dados: C. Restrepo-Arango
Análise de dados: C. Restrepo-Arango, A. L. Cárdenas-Rozo
Discussão dos resultados: C. Restrepo-Arango, A. L. Cárdenas-Rozo
Revisão e aprovação: C. Restrepo-Arango, A. L. Cárdenas-Rozo

Caso necessário veja outros papéis em: https://casrai.org/credit/

PUBLISHER
Universidade Federal de Santa Catarina. Programa de Pós-graduação em Ciência da Informação. Publicação no Portal de Periódicos UFSC. As ideias expressadas neste artigo são de responsabilidade de seus autores, não representando, necessariamente, a opinião dos editores ou da universidade.

EDITORES
Edgar Bisset Alvarez, Ana Clara Cândido, Patrícia Neubert e Genilson Geraldo.

HISTÓRICO
Recebido em: 20-08-2021 – Aprovado em: 15-06-2022 - Publicado em: 08-07-2022.