Revista de Humanidades Digitales

Complejidad léxica de artículos editoriales de la prensa española: Una selección de cuatro cabeceras

Lexical Complexity of Editorial Articles in the Spanish Press: A Selection of Four Newspapers

Fernando SANZ-LÁZARO
Universität Wien
fernando.sanz-lazaro@univie.ac.at
https://orcid.org/0000-0002-8815-6741

RESUMEN

Este estudio de caso explora las diferencias en la complejidad léxica (LC) de una selección de la prensa de calidad española. Los resultados muestran variabilidad del léxico entre periódicos y falta de correlación entre el número de lectores y una alta LC. Se calculan índices LS1 y CVS1 de sofisticación y HD-D, MAAS y MTLD de diversidad para evaluar 2741 artículos editoriales de *Abc*, *El Mundo*, *El País* y *El Periódico* publicados en línea durante 2019. Los resultados revelan diferencias significativas tanto en diversidad como en sofisticación, siendo *El Mundo* el periódico con los textos más complejos y *El Periódico* con los menos complejos. Adicionalmente, la comparación de HD-D, MAAS y MTLD con variaciones de TTR sugiere una ventaja de los primeros para muestras de tamaño heterogéneo, como las empleadas en el estudio.

PALABRAS CLAVE

Complejidad léxica, diversidad léxica, sofisticación léxica, NLP, prensa española.

ABSTRACT

This case study explores the differences in lexical complexity (LC) in the Spanish quality press. The results show variability of the lexical quality among papers and lack of correlation of number of readers and higher LC. The lexical sophistication indexes LS1 and CVS1 and lexical diversity indexes HD-D, MAAS, and MTLD were calculated for 2741 editorial articles of *Abc*, *El Mundo*, *El País*, and *El Periódico* published online in 2019. The results revealed significant differences in both LD and LS between the newspapers, with *El Mundo* producing the most and *El Periódico* the less complex texts overall. Posthoc analyses showed further differences between publications, being *El Periódico* the most disparate. Additionally, the comparison of HD-D, MAAS, and MTLD with TTR-based measures suggests benefits of the former for samples of heterogeneous sizes.

KEYWORDS

Lexical Complexity, Lexical Diversity, Lexical Sophistication, NLP, Spanish Press.

RHD 6 (2021)

ISSN
2531-1786

Recibido 26/05/2021 – Aceptado 13/07/2021

1. INTRODUCCIÓN[1]

La prensa de calidad[2] (Russ-Mohl, 2008) ha llevado tradicionalmente a gala la buena prosa de sus artículos editoriales. Estos textos, libres de muchas de las constricciones impuestas por el espacio o la inmediatez de piezas periodísticas de otros géneros, han desarrollado un estilo minuciosamente elaborado, con aspiraciones que trascienden la mera transmisión del parecer de la cabecera, con vocación de convertirse en la insignia del periódico y contribuir a su buena reputación, de la cual depende en buena medida la fuerza para crear opinión. En efecto, los periódicos de calidad no confían sus editoriales solo a la argumentación de posturas más o menos bien fundadas, sino que las envuelven en un cuidado lenguaje, capaz incluso de conferir mérito literario al texto. En consecuencia, este género de artículos suele diferir del lenguaje coloquial, pero también del lenguaje empleado en otro tipo de texto, incluso de aquellos que se hallan página con página con el propio editorial. Aparte de estas propiedades discursivas, estos artículos presentan otra característica que hace de ellos un caso singular: la atribución de su autoría. No es una persona individual quien firma el editorial sino que es el periódico de forma colectiva –y su director en último término –quien asume la responsabilidad del texto. Así, pues, mientras que otros tipos de texto son trabajo de un autor y, por lo tanto, tomados individualmente, sus cualidades solo son atribuibles a la pluma que los firma, cada editorial, por el contrario, es por sí mismo un indicador cualitativo de la tónica de la publicación.

Una medida que ofrece una descripción razonable de estos artículos es su complejidad léxica. Para ella, la lingüística computacional proporciona los medios para tomar mediciones objetivas de acuerdo a diferentes criterios. Laufer y Nation (1995) proponen como indicadores de la complejidad léxica la originalidad, la densidad, la sofisticación y la diversidad. Este estudio se basa en dos de estas magnitudes: la diversidad, que alude al ratio de palabras diferentes respecto al número total, y la sofisticación, o la frecuencia relativa de elecciones léxicas inusuales.

Los resultados presentados en este artículo pretenden arrojar luz sobre la situación de la prensa de calidad en España, determinando si los artículos editoriales de las publicaciones estudiadas presentan una complejidad léxica equivalente. Este propósito se alcanza mediante el análisis del estudio de la diversidad y sofisticación léxica de los artículos editoriales publicados en 2019 en cuatro periódicos generalistas de calidad para responder a las siguientes preguntas:

1. ¿Presenta la muestra de artículos editoriales seleccionada una sofisticación léxica análoga?
2. ¿Tienen los artículos editoriales una diversidad léxica equivalente?

[1] Este trabajo forma parte del proyecto Sound and Meaning in Spanish Golden Age Literature (P32563-G) financiado por FWF, Austrian Science Fund. En atención a la legislación sobre propiedad intelectual, no podemos distribuir el corpus de textos. No obstante, todos los programas empleados para recuperar y procesar las muestras están disponibles en línea como software libre (accesible desde: https://github.com/fsanzl/diarios), así como el conjunto de datos en bruto (accesible desde: https://zenodo.org/record/3888971).

[2] También llamada prensa *de referencia* o *de elite*, en contraposición a la prensa *popular* o *sensacionalista* (Hernández-Muñoz, 1997).

3. ¿Existe una relación entre la difusión del periódico y la complejidad léxica de sus editoriales?

2. MÉTODOS Y MATERIALES

2.1. Sofisticación léxica

La sofisticación léxica es un concepto que se refiere a las elecciones de vocablos inusuales, que suele redundar positivamente en la precisión del texto. La lingüística computacional provee una solución para calcular este valor mediante la comparación de las palabras del texto con su frecuencia relativa en un corpus. Este trabajo se vale de este método, considerando sofisticados aquellos términos no incluidos entre las 2000 palabras léxicas más frecuentes, un margen convencional propuesto por Laufer y Nation (1995). Las palabras léxicas comprenden adjetivos, sustantivos, adverbios y verbos plenos, por contraposición a palabras funcionales, con menor carga semántica intrínseca, tales como preposiciones, pronombres, conjunciones o verbos auxiliares, empleadas para expresar relaciones gramaticales entre las otras palabras de la oración.

Para incrementar la precisión, se lematizaron los textos con el objetivo de prevenir la clasificación de diferentes formas flexivas del mismo lema como tipos distintos. De esta manera, diferentes formas de un lema son consideradas tokens de un mismo tipo a pesar de variar en sus accidentes morfológicos. Asimismo, se anotó la función gramatical de las palabras del texto (PoS-tagging), de forma que el análisis discrimina entre tokens homógrafos de diferentes tipos. Por último, se aplicó una corrección al algoritmo para tratar los nombres propios.

Las medidas empleadas producen aproximaciones razonables para el rango de tamaños textuales en el que se encuentran los textos analizados. Estas medidas son Lexical Sophistication-I o LS1 (Linnarud, 1986), que mide el ratio de tokens léxicos sofisticados y el número total de tokens léxicos y un indicador robusto en condiciones de heterogeneidad de tamaños textuales, Corrected Verb Sophistication-I o CVS1 (Wolfe-Quintero, Inagaki & Kim, 1998), que representa el ratio de tipos verbales sofisticados y la raíz cuadrada del doble del número total de tokens verbales.

2.2. Diversidad léxica

La diversidad léxica en su forma más elemental se corresponde con el ratio de tokens y tipos, que es lo que representa el marcador Type-Token Ratio o TTR (Templin, 1957). Desafortunadamente, el valor de esta medida decrece a medida que aumenta el tamaño del texto, con independencia de sus características (Lu et al., 2014, p. 101), lo que hace de TTR una medida poco adecuada para comparar textos de tamaños heterogéneos. A la vista de esto, McCarthy y Jarvis (2007) sugieren emplear indicadores robustos como Measure of Textual Lexical Diversity o MTLD (McCarthy, 2005), Hypergeometric Distribution D o HD-D (McCarthy & Jarvis, 2007) y MAAS (Maas, 1972), o incluso combinarlos para obtener una representación más completa (McCarthy &

Jarvis, 2010, p. 391). Tomando en cuenta esta recomendación, este trabajo hace uso de las tres medidas robustas mencionadas.

Estos indicadores son modelos sofisticados para calcular la misma idea subyacente en TTR, la relación entre tokens y tipos. MAAS ajusta el valor de TTR de acuerdo a una curva logarítmica. Este índice presenta una correlación inversa con la diversidad léxica, esto es, su valor decrece si la diversidad se incrementa, lo que ha de tenerse en consideración a la hora de interpretar los resultados. HD-D es una idealización del índice D (MacWhinney, 2000), que toma muestras aleatorias sin restitución de tokens para calcular la probabilidad de nuevo vocabulario en nuevas muestras de tamaño incremental. Esto produce un modelo matemático de la variación de TTR de acuerdo al tamaño de la muestra, y D es el resultado de comparar ese modelo con los datos empíricos. HD-D mide la suma de probabilidades (*P*) de encontrar alguno de los tokens {*A..K*} del texto en una muestra aleatoria de n palabras del propio texto para cada uno de los tipos. Para hacer los cálculos se utilizó un tamaño de muestra *n*=42, que es el valor medio entre 35 y 50 usado para calcular el índice D. MTLD representa la longitud media de series secuenciales de tokens en un texto con un valor TTR por encima de un valor límite dado. Cuando el valor TTR cae por debajo de ese umbral, se incrementa el contador del factor *F* una unidad, y se pasa a la siguiente cadena, que es evaluada de la misma manera. En tanto que el texto pocas veces termina con un factor completo, los tokens del factor remanente se evalúan contando la proporción entre 1,0 y el factor completo (McCarthy & Jarvis, 2010, p. 384).

2.3. Corpus

Se han considerado los artículos editoriales de la edición digital de cuatro periódicos generalistas tomadas de entre las diez publicaciones con más difusión diaria en 2019. Los cuatro casos observados cubren alrededor del 70% de la suma total de lectores de los diez periódicos generalistas más difundidos. Para las estimaciones de audiencia, se consideraron las mediciones del Estudio General de Medios (Asociación para la Investigación de Medios de Comunicación, 2019).

Publicación	Miles de lectores	Porcentaje (información general)	Porcentaje (total)
*Marca**	1672	-	15,854
El País	1013	14,029	9,606
*As**	772	-	7,32
El Mundo	671	9,292	6,363
La Vanguardia	549	7,603	5,206
La voz de Galicia	514	7,118	4,874
Abc	460	6,37	4,362
*Mundo Deportivo**	387	-	3,67
*Sport**	363	-	3,442
El Periódico	361	4,999	3,423
Otros información general	3653	50,886	34,639
Otros	131	-	1,242
Total (información general)	7221	100	68,471
Total	10546	-	100
* Temático			

Tabla 1. Lectores diarios. Fuente: Asociación para la Investigación de Medios de Comunicación.

En atención a estos criterios, y atendiendo también a razones técnicas, se analizaron *Abc*, *El Mundo*, *El País* y la edición global de *El Periódico*, ya que solo estos diarios aseguraban la disponibilidad de las muestras. Dicho de otra manera, no todas las publicaciones ofrecen acceso a su archivo de contenidos, o lo hacen de forma tal que se dificulta la toma de muestras. A falta de una alternativa viable, este trabajo hubo de limitarse a aquellos periódicos que brindan la posibilidad de recuperar artículos editoriales de días pasados en su web. De acuerdo a esto, el corpus está compuesto de los artículos mostrados en tabla 2.

El número de observaciones sugiere que el tamaño de la muestra no influye en la distribución de las medidas de centralidad. En cualquier caso, los resultados se han obtenido usando equivalentes robustos a pruebas ANOVA[3] que no se ven afectados por la heterogeneidad de los tamaños de las muestras y no presumen linealidad. Así pues, los datos no se optimizaron para mejorar el ajuste.

[3] ANOVA es la técnica de análisis de varianza también conocida como análisis factorial. Constituye una herramienta básica para el estudio del efecto de uno o más factores (cada uno con dos o más niveles) sobre la media de una variable continua.

Publicación	Artículos	Días publicado	Número de tokens						
			X media	σ	0%	25%	50%	75%	100%
Abc	704	354	394,741	125,319	117	297	419,5	467	2114
El Mundo	758	315	356,16	147,061	114	211	360	472	1239
El País	718	362	497,138	90,066	306	442,25	493,5	548,75	932
El Periódico	665	336	362,129	105,82	218	256	374	415	760
Total	2845	365	402,681	132,391	114	292	426	479	2114

Tabla 2. Artículos editoriales en 2019. Fuente: elaboración propia.

A primera vista, *El País* publicó artículos editoriales de manera más regular, haciéndolo prácticamente a diario. Le siguen *Abc* y *El Periódico*, mientras que *El Mundo* se encuentra en el extremo opuesto pues, a pesar de ser el más prolífico considerando el volumen total de artículos, hizo las publicaciones de forma más esporádica, lo que compensó llegando a sacar hasta cinco textos al día en varias ocasiones. Este fenómeno puede atribuirse en parte a que se fecharon los artículos de acuerdo a los metadatos y no a la ubicación temporal mostrada en la web.

Por consiguiente, los scrapers –esto es, los programas utilizados para inspeccionar el código HTML de las páginas, extraer e interpretar sus elementos– pudieron asignar una fecha diferente a del listado si la página se creó en una fecha anterior a la de la publicación. Del mismo modo, no se recuperaron artículos producidos en 2018 listados en el 1 de enero de 2019 pero sí otros listados en 2020 compuestos en 2019. En cualquier caso, el estudio no considera la fecha salvo para limitar el muestreo, por lo que un margen de unos días es aceptable, lo que corroboran las gráficas de dispersión, que no sugieren tendencia alguna. Por otra parte, esta aproximación permite tomar directamente múltiples artículos diarios sin necesidad de homogeneizar los datos.

Un examen más de cerca de los textos revela detalles interesantes. Un simple conteo de palabras muestra que, en promedio, *El País* produjo los artículos más extensos, mientras que *El Mundo* publicó los más breves. El rango de tokens más amplio corresponde a *Abc*, con 117 y 2144, seguido de *El Mundo*, que también registra el mínimo absoluto, con 114 tokens y la extensión más variable de acuerdo a la desviación estándar. *El Periódico* se ubica dentro de los rangos de los otros tres, con el máximo más bajo y el mínimo más alto tras *El País*, por encima de *Abc* y *El Mundo*. La desviación estándar más baja la tiene *El País*.

2.4. Recolección de datos

Los datos se tomaron de las páginas web de los periódicos empleando scrapers programados específicamente para la tarea. Todos los programas parten de un concepto subyacente común: el scraper escudriña la portada de la sección de Opinión en busca de enlaces a los artículos editoriales y analiza a continuación las páginas a las que apuntan esos enlaces. Esta operación se repite con la siguiente página del sumario de artículos editoriales hasta alcanzar un número dado de repeticiones. Sin embargo, debido a la organización de cada periódico, se necesitó ajustar el código de forma individual con modificaciones más o menos complejas dependiendo del caso.

Recuperar artículos de *El País* resultó relativamente sencillo, pues el código HTML de sus páginas se ciñe a una estructura prefijada regular y consistente. Por el contrario, la página web de *El Mundo* cambió sus plantillas a lo largo de 2019 en numerosas ocasiones, incluso varias veces en un mismo mes, lo que hizo inviable aplicar un solo modelo. En su lugar, se hicieron varias pruebas para identificar los cambios y el momento en los que estos tienen lugar para añadir ajustes individualizados al scraper. *Abc*, a pesar de ser consistente, presentó una dificultad diferente por requerir una cuenta registrada para acceder al contenido completo de algunos artículos, lo que se ve reflejado en el código. Estos cambios no afectan sustancialmente a los algoritmos de selección de páginas y procesamiento del texto pero ralentizaron la recuperación de las páginas de forma sustancial. Por último, las páginas web de *El Periódico* presentan una estructura consistente con un código HTML claro y homogéneo que facilitó las primeras pruebas sobremanera. Sin embargo, su sitio web añadió una función de avance de página automático en el último trimestre de 2019, lo que demandó una reescritura no trivial del código del scraper para incluir la capacidad de recuperación dinámica de las páginas.

2.5. Preparación de los datos

En primer lugar, se inspeccionaron los archivos recuperados en busca de artefactos tales como errores sintácticos en el HTML del código fuente de la página para corregirlos a mano. Solo se registró una instancia de código HTML mal formado (Derrota en dos tiempos, 2019), aunque numerosos editoriales de *El Periódico* mostraban caracteres no imprimibles que hubo que eliminar manualmente para impedir su interpretación como tokens. Segundo, para facilitar la tokenización, se corrigieron los errores de puntuación más habituales y erratas tipográficas, tales como la falta de espacio después de punto mediante búsqueda y sustitución con expresiones regulares. Por ejemplo, las instancias del tipo *palabra1.palabra2* se reescribieron como *palabra1._Palabra2* para que el tokenizador reconociera dos palabras, mientras que acrónimos del tipo *AA.BB* se dejaron sin cambios por constituir una sola entidad.

Además de la lematización mencionada, las muestras se anotaron con un PoS-tagger. Ambas tareas se llevaron a cabo con el paquete Stanford CoreNLP (Manning et al., 2014), un conjunto de herramientas de procesamiento del lenguaje natural de última generación. Además de los anotadores, el paquete provee otras facilidades auxiliares para integrarlo en otros lenguajes de programación, como la librería StanfordNLP (Qi et al., 2018), que permitió su uso en el código Python usado en este estudio. Se empleó un pipeline con los siguientes anotadores: tokenizador, expansor de tokens multipalabra, PoS-tagger y lematizador. Asimismo, se tradujeron las etiquetas producidas por el anotador de Stanford Core NLP en notación PENN Treebank, organizando la salida en forma lema PoS.

Se usó el último modelo de lengua española preentrenado UD Spanish AnCora 2.0 (Martínez Alonso y Zeman, 2016) facilitado por StanfordNLP. Este modelo tiene un rendimiento de

98,7 en las pruebas CoNLL Shared Task, 2018 (SIGNLL, 2018) para PoS-tagging y 98,08 para lematización de acuerdo al Stanford NLP Group (2018).

Después de la anotación manual, se hizo una inspección visual de los archivos en busca de etiquetas incorrectamente asignadas. Algunas instancias de etiquetas al final del texto se corrigieron, así como otros errores, consecuencia de errores tipográficos del periódico, tales como el uso de un guion donde corresponde una raya en *Abc*, *El Mundo* y *El Periódico*, que produce un token etiquetado del tipo *-palabra_None* en lugar de dos tokens como *—_— palabra_PoS*, un fallo tipográfico que se da en todos los periódicos de la muestra salvo en *El País*.

2.6. Instrumentos

Los valores de HD-D, MTLD y MAAS se calcularon empleando una versión modificada de la librería lexicalrichness (Shen Yan Shun, 2018). No obstante, esta requirió un ajuste para deshabilitar el tokenizador incluido, pues, al contrario que el de Stanford NLP, producía inconsistencias como ignorar guarismos. Las rutinas resultantes se integraron en el código de *Lexical Complexity Analyzer* de Lu (2012). Sin embargo, este requirió también ajustes para adecuarlo al español. El programa de Lu está orientado al inglés, lo que redunda en la manera en que procesa los textos a diferentes niveles. Por un lado, los sufijos adverbiales, verbos auxiliares y puntuación difieren del español. Por otro lado, el diccionario de frecuencias con el que el programa compara el texto no se ajusta a los valores de la lengua castellana.

Para resolver el problema idiomático, se adaptaron las rutinas que afectaban a elementos específicos del lenguaje y se añadieron los signos de puntuación del español tales como comillas angulares o signos de apertura de interrogación y exclamación. En cuanto a las frecuencias, se resolvió acudiendo a CORPES XXI (Real Academia Española, 2018), un corpus anotado y lematizado de libre acceso que comprende más de 286 millones de formas. Sin embargo, el corpus requirió algunos ajustes previos. Por un lado, tuvo que ser tabulado de acuerdo a las columnas esperadas por el programa de Lu y, por otro lado, necesitó ser reetiquetado traduciendo las etiquetas a la nomenclatura PENN Treebank.

Asimismo, se hizo necesaria una última modificación de orden semántico. En tanto que los nombres propios son ítems léxicos y, por lo tanto, tienen efecto en todos los resultados, no es una cuestión trivial. Algunos autores optan por su exclusión para evitar la representación incorrecta de la sofisticación de los textos (David et al., 2009, p. 150), mientras que otros prefieren incluirlos pues, en sentido estricto, tienen contenido léxico completo al contrario que palabras funcionales como pronombres (Malvern et al., 2004, p. 142; David et al., 2009, p. 7).

Los textos se analizaron considerando el término medio en atención a que los nombres propios son parte integral del discurso periodístico, incluso si, por su aparición en los medios de comunicación masivos pasan a ser, por definición, de uso común desde que son noticia. A consecuencia de esto, en lugar de obviar por completo los nombres propios, los tokens correspondientes a estos son considerados como palabras léxicas junto a los nombres comunes, pero excluyéndolos de la cuenta de palabras sofisticadas.

Los valores así obtenidos se organizan en una tabla en formato CSV que es procesada con R para llevar a cabo los cálculos estadísticos. Estos cálculos incluyen análisis de la varianza usando el paquete MANOVA.RM (Friedrich, Konietschke & Pauly, 2019). Este paquete proporciona una alternativa no paramétrica a la prueba MANOVA, modified ANOVA-type statistic (MATS), que no asume términos de error normales, homogeneidad del tamaño muestral ni homocedasticidad (Bathke et al., 2018; Friedrich & Pauly, 2017; Konietschke et al., 2015). En tanto que esta prueba puede manejar tamaños muestrales heterogéneos, no requiere correcciones de la estadística ni de la muestra.

2.7. Limitaciones

En primer lugar, este trabajo se refiere exclusivamente a la situación de 2019 y, en tanto que es un estudio observacional, no puede inferirse causalidad a partir de él.

Segundo, solo se ha analizado una selección de periódicos. Este artículo es el resultado de un estudio casuístico que representa la situación de un subconjunto de todas las cabeceras españolas. Existen otros periódicos de gran difusión cuyos artículos editoriales podrían presentar características diferentes a los estudiados en este trabajo.

Tercero, los resultados incumben solo a los artículos editoriales. Algunos géneros periodísticos, como las noticias, tienden a ser más sucintos, mientras que otros, como las columnas de opinión, pueden estar firmados por prestigiosas plumas. De esta manera, si bien los resultados son atribuibles al periódico, pueden no corresponderse con los valores medios de la publicación.

Cuarto, la anotación de partes del discurso tiene la precisión indicada arriba.

Finalmente, más una prevención que una limitación, los indicadores empleados miden la complejidad léxica y la analizan mediante cálculos estadísticos de diferentes constructos. Su representatividad de la calidad del texto está sujeta a discusión. Esto es, por lo general, un texto preciso implica diversidad léxica, pero lo opuesto no es necesariamente cierto. Al contrario, figuras retóricas de repetición típicas de otros géneros textuales como las anáforas reducirían el índice de diversidad léxica de acuerdo a estos indicadores, que no considerarían el efecto estético buscado.

3. RESULTADOS

3.1. Sofisticación léxica

De acuerdo a los datos mostrados en tabla 3, *El Mundo* presenta los valores de sofisticación léxica más altos. Sin embargo, *Abc* supera a *El País* en su LS1, mientras que sucede lo contrario con CVS1. *El Periódico* se asemeja a *El País* en lo concerniente a LS1, pero queda detrás de los otros tres en CVS1, donde el primer cuartil queda casi al mismo nivel de la mediana del inmediatamente anterior. Para averiguar si esas diferencias se hizo una prueba RM MANOVA unidireccional con el periódico como variable independiente y los índices de sofisticación léxica LS1 y CVS1.

Publicación	0%		25%		50%		75%		100%		xmedia	
	LS1	CVS1	LS1	CVS1	LS1	CVS1	LS1	CVS1	LS1	CVS1	LS1	CVS1
Abc	0,195	0,289	0,304	0,905	0,338	1,162	0,373	1,443	0,526	2,67	0,34	1,196
El Mundo	0,202	0	0,316	0,935	0,352	1,251	0,384	1,584	0,515	3,064	0,351	1,272
El País	0,193	0,134	0,292	0,982	0,325	1,268	0,355	1,554	0,523	2,611	0,327	1,289
El Periódico	0,183	0,158	0,291	0,746	0,321	1,032	0,353	1,296	0,539	2,462	0,324	1,034
Total	0,183	0	0,3	0,889	0,333	1,18	0,37	1,474	0,539	3,064	0,336	1,202

Tabla 3. Distribución de la sofisticación léxica. Fuente: elaboración propia.

Los datos siguen las distribuciones unimodales que se aprecian en la Figura 1, pero también un número significativo de valores atípicos que producen un sesgo a la derecha y, aparentemente, desviaciones de la normalidad. El test se realizó con 10000 iteraciones usando wild-bootstrapping (Friedrich et al., 2017). A la semilla de aleatorización se le asignó un valor arbitrario (123) para asegurar la reproductibilidad del experimento. El resultado muestra una diferencia significativa en el análisis combinado (MATS=402,167, p<0,001). La realización de pruebas diferenciadas produce resultados equivalentes para LS1 (MATS=123,833, p<0,001) y CVS1 (MATS=172,271, p<0,001), lo que sugiere una diferencia significativa en las distribuciones. Los resultados llevaron a análisis posthoc para encontrar la distribución responsable del rechazo de la hipótesis nula.

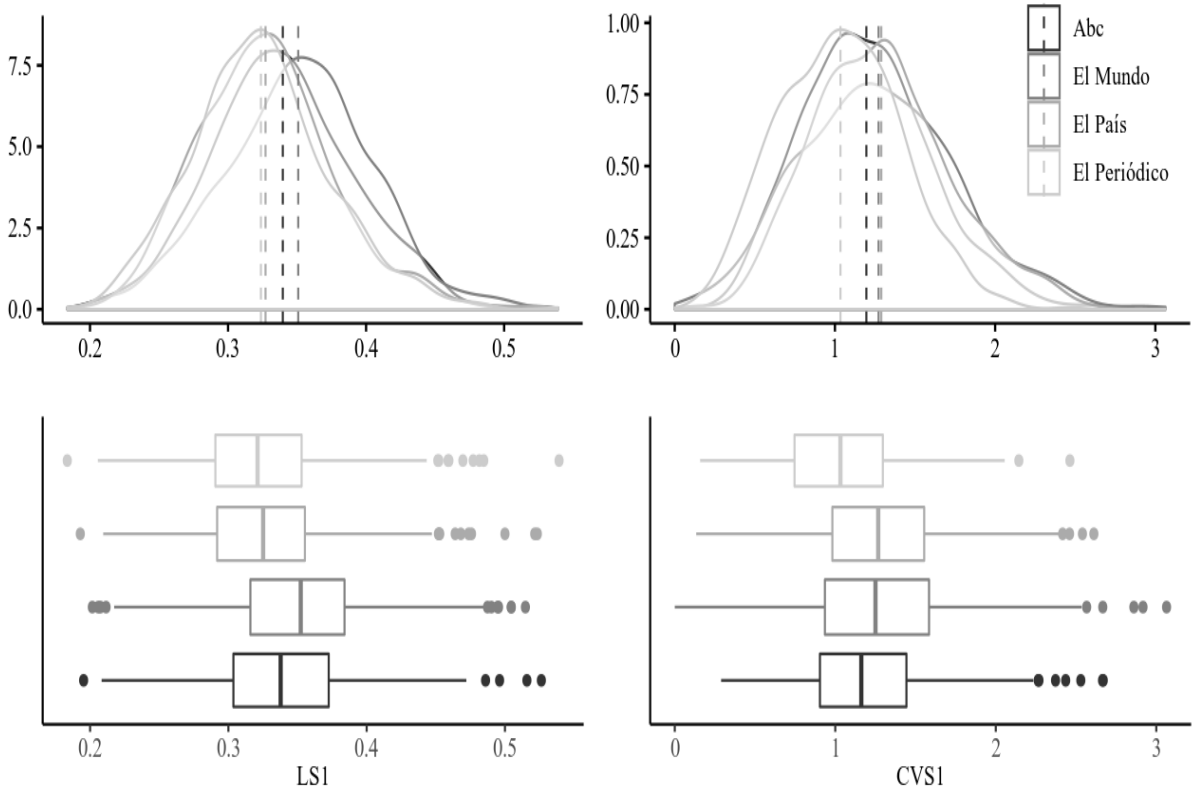

Figura 1. Distribución de la sofisticación léxica. Fuente: elaboración propia.

Una corrección Holm-Bonferroni del valor-p indica que ambos indicadores de sofisticación léxica contribuyen a la diferencia (0, 0). Un análisis de bootstrap univariado por pares de Mair y Wilcox (2020) con 10000 iteraciones muestra que solo *El País-El Periódico* (p=0,228) carece de

diferencias significativas para LS1, mientras que para CVS1 no resulta significativo *El Mundo-El País* (p=0,852), mientras que todos los demás resultaron significativos como muestra la Tabla 4.

Publicación	LS1				CVS1			
	Φ	IC inf	IC sup	p	Φ	IC inf	IC sup	p
Abc *El Mundo*	-0,014	-0,021	-0,006	<0,001	-0,084	-0,151	-0,017	0,001
Abc *El País*	0,013	0,006	0,02	<0,001	-0,089	-0,15	-0,027	<0,001
Abc *El Periódico*	0,016	0,009	0,024	<0,001	0,147	0,085	0,207	<0,001
El Mundo *El País*	0,027	0,019	0,034	<0,001	-0,005	-0,072	0,063	0,852
El Mundo *El Periódico*	0,03	0,023	0,038	<0,001	0,231	0,164	0,298	<0,001
El País *El Periódico*	0,003	-0,004	0,01	0,228	0,236	0,173	0,299	<0,001

Tabla 4. Comparación por pares de la sofisticación léxica. Fuente: elaboración propia.

3.3. Diversidad léxica

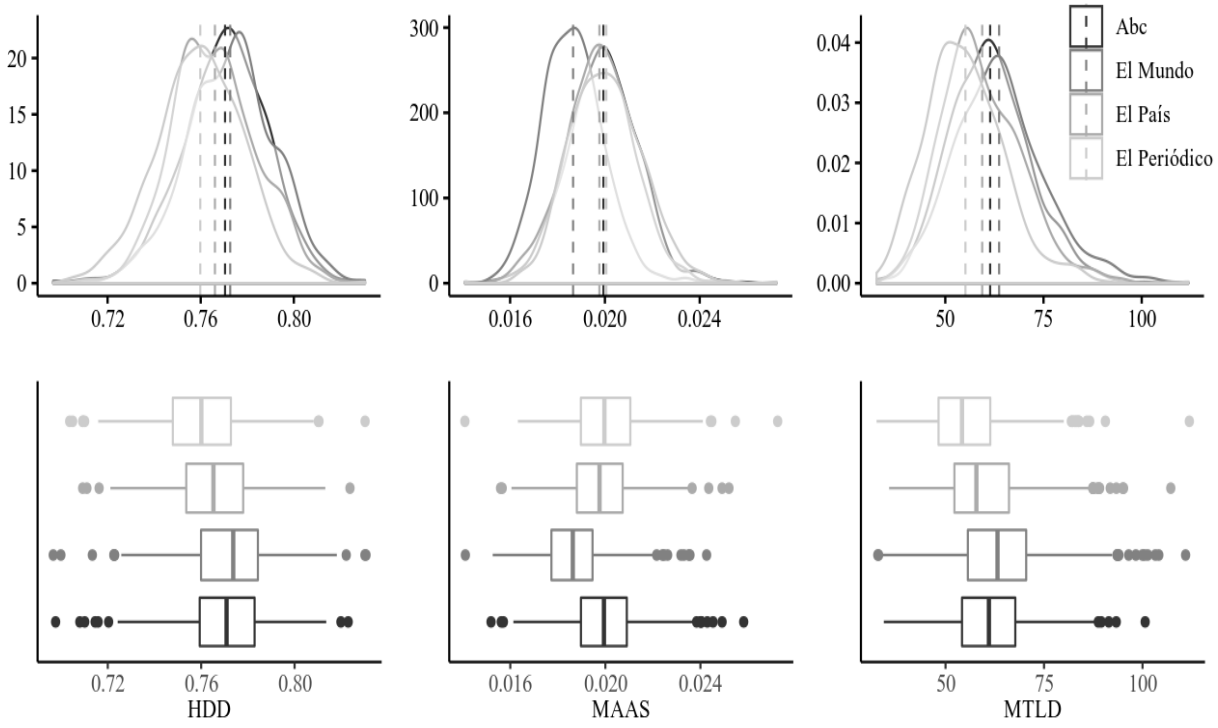

Figura 2. Distribución de la diversidad léxica. Fuente: elaboración propia.

La distribución de la diversidad léxica presenta un patrón parecido a la sofisticación. Los valores de centralidad más altos corresponden a *El Mundo* –más bajos en el caso de MAAS, que, como se menciona arriba, tiene una relación inversa–, como se observa en la figura 2. Después de aplicar una prueba MANOVA no paramétrica, se observa una diferencia significativa en la diversidad léxica (MATS=476,792, p<0,001). El ajuste del valor-p según Holm-Bonferroni sugiere que

los tres indicadores son responsables del rechazo de la hipótesis nula (0, 0, 0).

Publicación	0%			25%			50%			75%			100%		
	HD-D	MAAS	MTLD	HD-D	MAAS	MTLD	HD-D	MAAS	MTLD	HD-D	MAAS	MTLD	HD-D	MAAS	MTLD
Abc	0,698	0,015	34,34	0,759	0,019	54,183	0,771	0,02	61,01	0,783	0,021	67,707	0,823	0,026	100,669
El Mundo	0,697	0,014	32,812	0,76	0,018	55,662	0,774	0,019	63,193	0,784	0,019	70,469	0,831	0,024	110,93
El País	0,709	0,016	35,687	0,754	0,019	52,257	0,765	0,02	57,834	0,778	0,021	66,108	0,824	0,025	107,167
El Periódico	0,704	0,014	32,496	0,748	0,019	48,225	0,76	0,02	54,182	0,773	0,021	61,354	0,83	0,027	111,879
Total	0,697	0,014	32,496	0,755	0,019	52,358	0,768	0,02	59,281	0,78	0,021	66,723	0,831	0,027	111,879

Tabla 5. Distribución de la diversidad léxica. Fuente: elaboración propia.

La comparación por pares que se presenta en la tabla 6 muestra valores significativos para todos los valores excepto Abc-El Mundo para HD-D, Abc-El País y Abc-El Periódico para MAAS.

Publicación	HD-D				MAAS				MTLD			
	♆	IC inf	IC sup	p	♆	IC inf	IC sup	p	♆	IC inf	IC sup	p
Abc *El Mundo*	-0,002	-0,004	0,001	0,082	0,001	0,001	0,002	<0,001	-2,08	-3,629	-0,604	0,001
Abc *El país*	0,006	0,003	0,008	<0,001	0	0	0	0,055	2,389	0,844	3,889	<0,001
Abc *El Periódico*	0,011	0,008	0,014	<0,001	0	0	0	0,393	6,502	4,984	8,006	<0,001
El Mundo *El País*	0,008	0,005	0,01	<0,001	-0,001	-0,001	-0,001	<0,001	4,469	2,922	6,049	<0,001
El Mundo *El Periódico*	0,013	0,01	0,016	<0,001	-0,001	-0,002	-0,001	<0,001	8,583	7,069	10,159	<0,001
El País *El Periódico*	0,005	0,003	0,008	<0,001	0	0	0	0,007	4,114	2,602	5,623	<0,001

Tabla 6. Comparación por pares de la diversidad léxica. Fuente: elaboración propia.

3.4. Complejidad léxica y distribución de ejemplares

La prueba de Pearson, realizada para detectar la correlación entre la difusión de la publicación y las medianas de los índices de complejidad léxica de los editoriales, indica que esta o bien no existe o es muy débil, excepto para CVS1. Asimismo, como muestra Tabla 7, da cuenta de una fuerte correlación (p=0,85) que, no obstante, puede ser casual.

	Ejemplares	LS1	CVS1	MAAS	HD-D	MTLD
Ejemplares	1	0.015	0.85	-0.247	0.155	0.201
LS1	0.015	1	0.518	-0.839	0.947	0.947
CVS1	0.85	0.518	1	-0.558	0.653	0.687
MAAS	-0.247	-0.839	-0.558	1	-0.686	-0.709
HD-D	0.155	0.947	0.653	-0.686	1	0.999
MTLD	0.201	0.947	0.687	-0.709	0.999	1

Tabla 7. Correlación entre difusión y complejidad léxica. Fuente: elaboración propia.

4. CONCLUSIONES

De acuerdo a los resultados, parece claro que la complejidad léxica de los artículos editoriales publicados en 2019 por los periódicos estudiados varía de una publicación a otra de forma no trivial. Dicho de manera más precisa, hay diferencias estadísticas significativas en las distribuciones de la variedad léxica y de la sofisticación léxica de los artículos editoriales de los cuatro periódicos analizados.

El análisis de los datos revela que *El Mundo* publicó en 2019 los artículos editoriales con el léxico más elaborado aunque, paradójicamente, también es el único periódico que sacó un artículo todos cuyos verbos pueden considerarse no sofisticados (Por un turismo de mayor calidad, 2019). Asimismo, la distribución de las variables de complejidad presenta una notable variabilidad, lo que hace que buena parte de los artículos de *El Mundo* se encuentren en unos términos semejantes a los del tercer cuartil de *Abc* y *El País*. En la sofisticación de sus formas verbales, *El Mundo* es similar a *El País* en general, lo que sugiere que ambos describen acciones, estados y actitudes con una precisión léxica pareja. Por otra parte, aunque uno de los índices de diversidad léxica sugiere cierta similitud entre *El Mundo* y *Abc*, los otros dos indicadores no muestran signos de esta.

Los resultados de *El Mundo* llaman la atención al considerar que la dirección del periódico cambió cada año entre 2014 y 2017, incluso varias veces, y cada uno de los seis directores de ese periodo pudo influir en los artículos editoriales. Por lo tanto, sería interesante dilucidar si los valores para 2019 obtenidos en este estudio se corresponden con los históricos. La distribución no muestra tendencias que permitan aventurar una predicción, como cabe esperar de la estabilidad del último periodo, por lo que se haría necesario un estudio longitudinal con datos de los años anteriores a 2017 para encontrar una posible variación de la complejidad a lo largo del tiempo.

Los artículos menos complejos los publicó *El Periódico*, aunque también es la cabecera que tiene las distribuciones más homogéneas, con casos extremos más infrecuentes y valores centrales más habituales que los de otros periódicos. A pesar de esto, atendiendo solo a la sofisticación léxica del conjunto de todas sus palabras con valor semántico pleno, este diario está en la misma categoría que *El País*. A raíz de los resultados surge la pregunta acerca de si estos son fortuitos o existe otra causa. El cambio de dirección de 2019 no parece haber influido en los resultados, máxime considerando su regularidad. Por otra parte, se trata de una cabecera editada en una región bilingüe y parte de cuyos textos son traducciones. El hipotético peso de la traducción sería, en todo

caso, también relativo a la vista de la centralidad de la distribución. Sería, tal vez, más interesante comparar periódicos catalanes y otras regiones bilingües para comprobar si la baja complejidad léxica es un fenómeno aislado limitado a este diario. Esto requeriría, claro está, acceso a los artículos que, como se ha mencionado, no es una cuestión sencilla pues requiere que las empresas editoriales tengan voluntad y medios para permitir el acceso a los contenidos.

Podemos también concluir que un mayor número de lectores no es sinónimo de mayor complejidad léxica, pues no se ve una correspondencia clara. Estos resultados son inesperados, pues *El País*, además de ser el medio de mayor difusión, ha sido considerado históricamente como el diario de referencia de la prensa escrita española (Imbert & Vidal-Beneyto, 1986). En todo caso, no puede dejar de mencionarse que hubo un cambio de dirección en 2018. Desafortunadamente, no es posible determinar si esto es un factor determinante, pues los datos de ese 2019 no sugieren una tendencia que permita extrapolar una predicción, por lo que se haría necesario comprobar si tuvo lugar un cambio visible en 2018.

REFERENCIAS BIBLIOGRÁFICAS

Asociación para la Investigación de Medios de Comunicación. (2019). Ranking de diarios (2019-3.ª Ola). *Estudio General de Medios*. http://reporting.aimc.es/index.html#/main/diarios

Bathke, A. C., Friedrich, S., Pauly, M., Konietschke, F., Staffen, W., Strobl, N., & Höller, Y. (2018). Testing Mean Differences among Groups: Multivariate and Repeated Measures Analysis with Minimal Assumptions. *Multivariate Behavioral Research*, *53*(3), 348-359. https://doi.org/10/gjpksj

David, A., Myles, F., Rogers, V., & Rule, S. (2009). Lexical Development in Instructed L2 Learners of French: Is There a Relationship with Morphosyntactic Development? En B. J. Richards, D. D. Malvern, M. H. Daller, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application* (pp. 147-163). Palgrave Macmillan.

Derrota en dos tiempos [Editorial]. (2019, 13 de agosto). *El País*. https://elpais.com/elpais/2019/08/12/opinion/1565629594_068797.html

Friedrich, S., Konietschke, F., & Pauly, M. (2017). A Wild Bootstrap Approach for Nonparametric Repeated Measurements. *Computational Statistics & Data Analysis*, *113*, 38-52. https://doi.org/gk9k

Friedrich, S., Konietschke, F., & Pauly, M. (2019). *MANOVA.RM* (Version 3.4.0) [Software]. https://cran.r-project.org/web/packages/MANOVA.RM/index.html

Friedrich, S., & Pauly, M. (2017). MATS: Inference for Potentially Singular and Heteroscedastic MANOVA. *Journal of Multivariate Analysis*, 165, 166-179. https://doi.org/gk9m

Hernández-Muñoz, R. (1997). Avances en la búsqueda de un sistema de calidad para las redacciones de los diarios. *Communication & Society*, *10*(1), 169-192. https://revistas.unav.edu/index.php/communication-and-society/article/view/35640/31018

Imbert, G., & Vidal-Beneyto, J. (1986). *El País o la referencia dominante*. Mitre.

Konietschke, F., Bathke, A. C., Harrar, S. W., & Pauly, M. (2015). Parametric and Nonparametric Bootstrap Methods for General MANOVA. *Journal of Multivariate Analysis*, *140*, 291-301. https://doi.org/gk9n

Laufer, B. y Nation, P. (1995). Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/bwf4sp

Linnarud, M. (1986). *Lexis in Composition: A Performance Analysis of Swedish Learners' Written English*. CWK Gleerup.

Lu, X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, *96*(2), 190-208. https://doi.org/f3722w

Lu, X., Gamson, D. A., & Eckert, S. A. (2014). Lexical Difficulty and Diversity of American Elementary School Reading Textbooks. *International Journal of Corpus Linguistics*, *19*(1), 94-117. https://doi.org/gk9p

Maas, H. D. (1972). Zusammenhang zwischen Wortschatzumfang und Länge eines Textes. *Zeitschrift Für Literaturwissenschaft Und Linguistik*, *8*, 73-79.

MacWhinney, B. (2019). *Tools for Analyzing Talk, Part 2: The CLAN Program*. Carnegie Mellon University. https://talkbank.org/manuals/CLAN.pdf

Mair, P., & Wilcox, R. R. (2020). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods, 52*, 464-488. https://doi.org/gf7fm2

Malvern, D. D., Richards, B. J., Chipere, N., & Durán, P. (2004). *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55-60. https://doi.org/gf3xhp

Martínez Alonso, H., & Zeman, D. (2016). Universal Dependencies for the AnCora Treebanks. *Sociedad Española para el Procesamiento del Lenguaje Natural, 57*, 91-98. http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5341

McCarthy, P. M. (2005). *An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)* [Tesis doctoral, University of Memphis]. https://bit.ly/36cy6sY

McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, *24*(4), 459-488. https://doi.org/c3qdpx

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods*, *42*(2), 381-392. https://doi.org/bbvvsj

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*, Vol. 2 (3.ª ed.). Lawrence Erlbaum.

Por un turismo de mayor calidad [Editorial]. (2019, 15 de agosto). *El Mundo*. https://www.elmundo.es/opinion/2019/08/16/5d5590dffdddffa4548b45cf.html

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160-170. https://doi.org/gf6gst

Real Academia Española. (2018). *Banco de datos (CORPES XXI)*. Corpus del Español del Siglo XXI (CORPES) [Corpus]. https://www.rae.es/recursos/banco-de-datos/corpes-xxi

Russ-Mohl, S. (2008). Quality Press. En W. Donsbach (Ed.), *The International Encyclopedia of Communication*. Wiley-Blackwell. https://doi.org/gwpr

Shen Yan Shun, L. (2018). *LexicalRichness: A small module to compute textual lexical richness* (Version 0.1.3) [Software]. https://github.com/LSYS/lexicalrichness

SIGNLL. (2018). *CoNLL 2018 Shared Task*. SIGNLL: ACL's Special Interest Group on Natural Language Learning. https://universaldependencies.org/conll18/

Stanford NLP Group. (2018). *System Performance*. StandfordNLP. https://stanfordnlp.github.io/stanfordnlp/performance.html

Templin, M. C. (1957). *Certain Language Skills in Children*. University of Minnesota Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity*. University of Hawai'i Press.