http://dx.doi.org/10.1590/1984-6398201914113

Descripción de la diversidad y densidad léxicas en noticias escritas por estudiantes de periodismo[1]

Lexical Diversity and Lexical Density Description of News Written by Journalism Student

Descrição da diversidade e densidade léxicas em notícias escritas por estudantes de jornalismo

Karina Fuentes Riffo*[2]
*Universidad de Concepción, Concepción, Bío Bío / Chile
kafuente@udec.cl
https://orcid.org/0000-0002-3644-3327

Sergio Hernández Osuna**
**Universidad de Concepción, Concepción, Bío Bío / Chile
sergihernandez@udec.cl
https://orcid.org/0000-0002-1366-0409

Pedro Salcedo Lagos***
***Universidad de Concepción, Concepción, Bío Bío / Chile
psalcedo@udec.cl
https://orcid.org/0000-0002-1741-714X

RESUMEN: El objetivo de este trabajo es describir cuantitativamente la riqueza léxica de noticias informativas, escritas por estudiantes de periodismo, mediante los indicadores de diversidad y densidad léxicas. Se realizó un análisis comparativo de corpus de noticias elaboradas por estudiantes en dos ciclos universitarios (primer año y cuarto año) y por un grupo de profesionales. Se

[1] Este artículo forma parte de los resultados de la Tesis Doctoral Competencia léxica de estudiantes de periodismo en el discurso especializado escrito, del año 2018, por el Programa Doctorado en Lingüística, Universidad de Concepción, Concepción, Chile.

[2] Candidata a Doctora en Lingüística por la Universidad de Concepción.

utilizaron las métricas de densidad léxica y de diversidad léxica para comparar los corpus y, mediante pruebas estadísticas, se analizaron las diferencias entre los grupos. Si bien los resultados muestran diferencias, es la densidad léxica el indicador en que se encuentran las diferencias estadísticamente significativas, evidenciando que aunque los estudiantes manejan un léxico amplio y variado, el nivel informativo y la capacidad comunicativa de sus textos son insuficientes comparados con el que se utiliza a nivel profesional.

PALABRAS CLAVE: riqueza léxica; diversidad léxica; densidad léxica; discurso profesional.

ABSTRACT: The aim of this study is to quantitatively describe the lexical richness in informative news articles written by journalism students through indicators of lexical diversity and lexical density. A comparative analysis of a corpus of news articles written by two groups of undergraduate students (first and fourth year) and a group of graduate journalists was performed. In order to compare the corpora, measures of lexical density and diversity were used while statistical tests were applied to analyze differences between the groups. Though results show differences, it is density the indicator where statistically significant differences were found. This would indicate that despite students having a command of a wide and varied lexis, the informative level and communicative capacity of their texts are insufficient compared to the requirements at a professional level.

KEYWORDS: lexical richness; lexical diversity; lexical density; professional discourse.

RESUMO: O objetivo deste trabalho é descrever quantitativamente a riqueza lexical de notícias informativas, escritas por estudantes de jornalismo, através dos indicadores de diversidade e densidade lexical. Foi realizada uma análise comparativa do *corpus* de notícias elaborado por alunos em dois períodos (no primeiro e no quarto ano) e por um grupo de profissionais. As métricas densidade lexical e diversidade lexical foram utilizadas para comparar o *corpus* e, por meio de testes estatísticos, foram analisadas as diferenças entre os grupos. Embora os resultados mostrem diferenças, é a densidade lexical o indicador no qual os dados estatisticamente significativos são encontrados, evidenciando que, ainda que os alunos manejem um léxico amplo e variado, o nível informativo e a capacidade comunicativa de seus textos são insuficientes se comparados ao exigido a nível profissional.

PALAVRAS-CHAVE: riqueza lexical; diversidade lexical; densidade lexical; discurso.

1 Introducción

Densidad y diversidad léxicas (DALLER; VAN HOUT; TREFFERS-DALLER, 2003; GREGORY-SIGNES; CLAVEL-ARROITÍA, 2015; JOHANSSON, 2008; TORRUELLA; CAPSADA, 2013) son medidas

cuantitativas utilizadas para evaluar la riqueza léxica, pero también para medir el progreso de los estudiantes en determinado ámbito lingüístico.

Este estudio se focaliza en el discurso profesional escrito, específicamente en noticias escritas, elaboradas por estudiantes de periodismo y por un grupo de periodistas profesionales, con el objetivo de evaluar la riqueza léxica a nivel de formación universitaria y compararla con la riqueza léxica profesional.

Para llevar adelante esta investigación, se utilizaron dos medidas cuantitativas indicadoras de riqueza léxica de un texto: la **diversidad léxica** y la **densidad léxica**. La diversidad léxica se refiere al número de palabras diferentes utilizadas en un texto, un rango mayor indica una diversidad mayor (JOHANSSON, 2008; LÓPEZ MORALES, 2002; MCCARTHY; JARVIS, 2007, 2010). La densidad léxica (GREGORY-SIGNES; CLAVEL-ARROITÍA, 2015; JOHANSSON, 2008, READ, 2010), en tanto, se entiende como la relación entre el total de palabras léxicas –o de contenido semántico– (verbos, nombres, adjetivos y algunos adverbios) comparado con las llamadas palabras gramaticales –o funcionales– (artículos, preposiciones, conjunciones, entre otros).

Para llevar adelante el objetivo, se propone realizar una comparación entre los índices de riqueza léxica de noticias informativas escritas por estudiantes de periodismo en dos ciclos universitarios (primer año y cuarto año), y la riqueza léxica en el mismo género discursivo, pero en textos escritos por periodistas profesionales, en tres temáticas: ciencia y tecnología, política y deporte. La idea de lo anterior radica en analizar la riqueza léxica que tienen los estudiantes de esta carrera al comenzar sus estudios universitarios y el posible avance que presentan al avanzar en su malla curricular; lo anterior relacionado con lo que necesitan llegar para desenvolverse en el nivel profesional.

Los mayores avances en evaluación de riqueza léxica se han dado en la enseñanza de segundas lenguas –L2– (JIMÉNEZ, 2002; ŠIŠKOVÁ, 2012). Por otra parte, en lengua materna los estudios han estado enfocados en las etapas de formación primaria y secundaria o de enseñanza media (REYES, 2010), y un número reducido ha indagado en el campo de la riqueza léxica considerando al lenguaje periodístico como un tipo de discurso especializado.

El discurso especializado, abordado desde el ámbito léxico-estadístico, ha sido poco estudiado teniendo como foco de estudio el discurso

periodístico. Quizás uno de los estudios más emblemáticos en habla hispana fue el coordinado por Raúl Ávila, del Centro de Estudios Lingüísticos y Literarios de El Colegio de México, que junto a un grupo de lingüistas de distintos países se plantearon estudiar el español utilizado por los medios de comunicación de Hispanoamérica y España.

En Chile, Echeverría (1997) estudió la riqueza léxica en las noticias emitidas por el canal TVN y por el del deporte en vivo transmitido por dos radioemisoras del país. En dicho trabajo se analizaron 48 unidades textuales, de 1.200 palabras cada una, con un total aproximado de 72.000 palabras gráficas (también llamadas *running words*). Los resultados de esta muestra señalan que en los noticieros aparecen 4.259 tipos (formas distintas) y 1.965 voces distintas en deportes (ECHEVERRÍA, 1997).

En 2016 la Fundación del Español Urgente,[3] Fundéu BBVA, publicó los resultados del Proyecto Aracné. Este estudio tuvo por objetivo realizar una medición de la variación de la riqueza lingüística en la prensa española escrita desde 1914 hasta 2014. A grandes rasgos, los resultados mostraron que los valores de la riqueza lingüística de los medios de comunicación españoles se mantienen en general muy estables, con diferencias mínimas entre épocas. Los valores además están muy concentrados, las diferencias respecto a la media nunca superan el 10%, y se mantienen sorprendentemente uniformes en todas las épocas y para todas las longitudes.[4]

2 Antecedentes teóricos

2.1 Discurso especializado

La principal función de las noticias es satisfacer la necesidad de información de la sociedad mediante un lenguaje funcional en un texto que, por una parte, debe ser de fácil comprensión para el lector y, por otra, debe cumplir con una serie de requisitos y cualidades expuestos en diversos manuales de redacción periodística. Entre ellos se destacan aspectos como

[3] Fundéu BBVA es una institución sin fines de lucro que tiene como principal objetivo impulsar el buen uso del español en los medios de comunicación. Nacida en el año 2005 fruto de un acuerdo entre la Agencia Efe y el banco BBVA, trabaja asesorada por la Real Academia Española de la Lengua. Enlace de acceso: http://www.fundeu.es

[4] Ver informe completo en el sitio web http://www.fundeu.es/aracne/index.html, revisado nuevamente el 4 de enero de 2018.

claridad, concisión, densidad, exactitud, precisión, sencillez, originalidad, brevedad, variedad, corrección y propiedad (MARTÍN VIVALDI, 1993, 2004; MARTÍNEZ ALBERTOS, 1989).

Uno de los aspectos novedosos de este trabajo consiste en circunscribir la riqueza léxica al marco del discurso especializado. Si bien los estudios tradicionales han analizado la riqueza léxica relacionada a la adquisición de la L1 o al aprendizaje de una L2, un nuevo ámbito de la adquisición del léxico se abre con el surgimiento del llamado discurso especializado, que se entiende como un conjunto variado de tipos de textos, pero con rasgos prototípicos.

Parodi (2006) plantea que el discurso especializado es un "continuum", en el que se organizan los textos a lo largo de una gradiente diversificada que va desde un alto hasta un bajo grado de especialización. En este contexto, Gómez (2002) sostiene que el léxico específico designa las voces utilizadas en situaciones de comunicación que implican la transmisión relevante de un campo de conocimiento y de experiencia particular. Este vocabulario va siempre acompañado de una exigencia de exactitud y precisión, característico de los campos científicos, técnicos y profesionales a los que el léxico específico hace referencia y de los que se nutre.

Gómez propone que la competencia léxica, enmarcada en un discurso especializado, tiene características propias y relevantes, entre las cuales se puede ejemplificar el tipo de léxico específico. Este es uno de los fundamentos que investigadores sobre discurso especializado (CABRÉ, 1993; PARODI, 2005a, 2005b, 2006) utilizan para definirlo como un objeto de estudio científico.

Parodi (2006) define al discurso especializado como una hiperonimia de los llamados discurso académico y discurso profesional. Cabré (1993) explica algunos fundamentos de los desacuerdos, que surgieron de la consideración de los aspectos comunes entre la lengua general y este lenguaje especializado.

No obstante, resaltando los puntos coincidentes, se puede llegar a la siguiente definición de lenguaje especializado:

> Se trata de conjuntos "especializados", ya sea por la temática, la experiencia, el ámbito de uso o los usuarios; se presentan como un conjunto con características interrelacionadas, no como fenómenos aislados; mantienen la función comunicativa como predominante, por encima de otras funciones complementarias (CABRÉ, 1993, p. 19).

El lenguaje especializado está caracterizado, pragmáticamente, por las variables temáticas, usuario y situación de comunicación, que implican a su vez unas peculiaridades lingüísticas y textuales. Es decir, que frente a la lengua general los lenguajes especializados se desarrollan en función de una temática determinada, siendo especiales en cuanto al contenido de su discurso ya que transmiten un conocimiento específico (PARODI, 2006).

2.1.1 La noticia como discurso especializado

La noticia es la materia prima del trabajo del periodista (MARTÍNEZ ALBERTOS, 1989), la cual se concreta en dos productos contrastados: el relato y el comentario. Mientras que en el primero el fin es transmitir hechos que se consideran de interés para el público, en el segundo el objetivo es la expresión de ideas, juicios o pensamientos. Ambos ofrecen una gama de "submodelos" particulares: la noticia, el reportaje, la crónica y el artículo; géneros concretos y fundamentales del periodismo (MARTÍNEZ ALBERTOS, 1989, 1991).

Este trabajo utiliza la noticia escrita, un tipo de texto perteneciente al género llamado Periodismo Informativo. Una las principales características estructurales de este tipo de texto es la organización de la información, que se le ha asemejado a una pirámide invertida. Esta pirámide es una estructura en la cual, desde el primer momento, el lector encuentra toda la información relevante del acontecimiento, mientras que en los párrafos posteriores se ven anexados datos que complementan dicha noticia.

La estructura consiste en que entre el titular y el *lead* del texto (sus dos primeras partes) debe entregarse la información más relevante de la noticia al lector. Esta pirámide invertida se contrapone a la estructura de pirámide tradicional, utilizada por la ficción y el drama, en que la tensión del relato se dispone de manera creciente a medida que éste avanza.

La redacción de noticias bajo la concepción de pirámide invertida ha comprobado su eficacia y utilidad para la transmisión de datos y tiene, a la vez, ventajas para conseguir ahorro de espacio en las páginas impresas de los periódicos. Por otro lado, se debe tener en cuenta que este estilo expositivo capta de inmediato el interés del lector al entregar, desde el titular de la noticia, lo más relevante y atrayente de un hecho noticioso (MARTÍNEZ ALBERTOS, 1991).

2.2 Riqueza léxica

Müller en 1973 propuso la idea de que la estructura de un vocabulario comprende elementos cuantitativos simples: el número de vocablos del texto y la frecuencia de cada uno de ellos. Junto a estos elementos agrega aspectos cualitativos, tales como la naturaleza gramatical de los vocablos y las relaciones de asociación (gramaticales o semánticas, paradigmáticas y sintagmáticas), que existen entre vocablos. Para dicho autor, cuantificar vocabulario de un texto es proceder a dos operaciones distintas que pueden ser sucesivas o simultáneas: a) El recuento de las palabras que componen el texto y cuyo número, representado por N, dará una medida de la extensión del texto; y b) El recuento de los vocablos empleados en el texto y cuyo número, representado por V, mide la extensión del vocabulario.

Por su parte, Read (2010) señala que la riqueza léxica es una medida estadística y supone que la buena escritura tiene características léxicas como la variedad de palabras diferentes, en lugar de un número limitado de palabras utilizadas repetitivamente. La **diversidad léxica** se refiere al número de palabras diferentes utilizadas en un texto, un rango mayor indica una diversidad mayor (JOHANSSON, 2008; LÓPEZ MORALES, 2002; MCCARTHY; JARVIS, 2010), la medida que se aplica en este caso es el *type-token ratio* o TTR, donde *type* corresponde a las palabras distintas del texto, y *token*, al número total de palabras. Existe, además, otra clasificación que para efectos de este trabajo no es significativo, se trata del *hápax legomena*, que corresponde a palabras cuya aparición es única dentro del texto.

Una segunda característica tiene que ver con el porcentaje de palabras léxicas –o de contenido– (verbos, nombres, adjetivos y algunos adverbios) comparadas con las llamadas palabras gramaticales –o funcionales– (artículos, preposiciones, conjunciones, entre otros). Esta medida se reconoce como **densidad léxica** (READ, 2010; JOHANSSON, 2008). La densidad léxica es, además, un indicador de la cualidad informativa del texto: un alto índice de densidad léxica significa que tiene a su haber más palabras de contenido, por lo tanto, entrega mayor información.

La estadística léxica, o léxico-estadística, ha sido el primer peldaño para dar sustento a estudios de este tipo, abarcando el conjunto de operaciones, a veces sumamente complejas, que toman como unidades de trabajo las palabras y los vocablos; la palabra, unidad del texto, y el vocablo, unidad del léxico (LÓPEZ MORALES, 2002). Tanto la diversidad como la densidad léxica son medidas cuya ventaja radica en su sencilla operacionalización y

medición gracias a los avances en técnicas computacionales de análisis, junto a los estudios de corpus (JOHANSSON, 2008).

2.2.1 Índices de riqueza léxica

Para abordar el análisis de riqueza léxica, se siguió la línea tradicional de la léxico-estadística, que en primer lugar busca obtener el porcentaje de vocablos; es decir, la proporción de palabras diferentes del total de palabras de una composición, lo que otorga una visión de la diversidad léxica de un texto y entrega, según López Morales (2002), un indicador grueso del texto. Para calcular este índice se contabiliza el número total de palabras del discurso y el número de vocablos (entendidos como palabras diferentes), con el fin de reconocer la proporción entre ambas cifras en cada uno de los textos.

Este índice, también conocido como TTR –o Type-Token Ratio–, se representa con la fórmula:

$$\mathbf{LV = V/N}$$

(LV) = Diversidad léxica.
(V) = Total de vocablos.
(N) = Extensión total del texto.

Un segundo cálculo, propuesto por las investigaciones de López Morales, para complementar el concepto de riqueza léxica, está dado por el Intervalo de aparición palabras de contenido nocional (IAT), que considera la cantidad total de palabras del texto y la divide por el conjunto formado por nombres, adjetivos, verbos y adverbios, o también conocidas como palabras nocionales (PN) o palabras con contenido semántico:

$$\mathbf{IAT = N/PN}$$

(IAT) = Intervalo de aparición de palabra.
(N) = Extensión del texto.
(PN) = Palabras nocionales.

El resultado de esta operación matemática refleja cifras relacionadas con la proporción de palabras nocionales –o de contenido– en el texto. En otras palabras, mayor número de palabras nocionales, menor es el intervalo; lo que se interpreta como un mejor índice de riqueza léxica. Se trata de una medida que resulta efectiva especialmente cuando se trabaja con textos de un grado académico particular y se pretende identificar la relación de un sujeto

con el resto del grupo. En este caso la relación que se explora es la diferencia entre los dos grupos de estudiantes y la riqueza léxica de los profesionales.

El índice de densidad léxica fue propuesto por Ure en 1971. En sus hallazgos señaló que usualmente más del 40% de un texto corresponde a palabras léxicas, mientras que en el discurso oral este porcentaje no supera dicha cifra. Lo anterior sería reflejo de que la información y las ideas se presentan con mayor concentración en el lenguaje escrito que en el oral (READ, 2010). Para obtener este índice se aplicará al *lead* de las noticias la siguiente fórmula matemática:

$$\mathbf{DL = PN/N}$$

(DL) = Densidad léxica.

(PN) = Total de palabras nocionales.

(N) = Extensión total del texto.

3 Metodología

Este trabajo se basó en un diseño cuantitativo, no experimental y transversal. Se efectuaron diversas comparaciones, la primera fue una general entre la riqueza léxica de los textos elaborados por los estudiantes de periodismo y los profesionales en los tres dominios; una segunda comparación entre los estudiantes agrupados por ciclo universitario, finalmente, una tercera comparación entre los índices obtenidos según temática (ciencia y tecnología, política y deporte). El análisis de los datos y la elaboración de los gráficos se efectuaron con el *software* estadístico SPSS 23.

3.1 Objetivo general

Evaluar la riqueza léxica de estudiantes de periodismo en el discurso especializado escrito, por medio de un modelo cuantitativo, y compararla con la riqueza léxica profesional, para establecer el nivel de competencia léxica que se requiere durante el ejercicio de la profesión.

3.1.1 Preguntas de investigación

- En términos de diversidad y densidad léxicas, ¿cuál es la diferencia en cada uno de los centros de interés?
- ¿Cuál es la diferencia de los índices de riqueza presentada por los textos de la muestra de estudiantes en comparación con el corpus profesional?

3.2 Características metodológicas del estudio

3.2.1 Participantes

El estudio está compuesto por tres corpus temáticos de noticias informativas elaboradas por estudiantes de periodismo pertenecientes a dos niveles distintos de formación universitaria y por profesionales, configurándose tres grupos en total: G1, alumnos de primer año; G2, alumnos de cuarto año; y G3, profesionales, conforme se presenta en la Tabla 1.

TABLA 1 – Tabla cruzada según sexo y grupo

		GRUPO			Total
		G1	G2	G3	
SEXO	hombre	25	9	5	39
	mujer	33	11	15	59
Total		58	20	20	98

En resumen, los participantes en el estudio fueron 98 personas: 58 estudiantes de primer año; 20 de cuarto año; y 20 profesionales. La Tabla 1 ordena, además, los grupos según sexo. Por otra parte, se obtuvieron datos sobre la procedencia administrativa del establecimiento educacional en el cual cursaron su enseñanza media. De los participantes, 28 cursaron sus estudios secundarios en un establecimiento cuya dependencia administrativa correspondió a privada o particular; 26 a establecimientos municipalizados; y 44 a establecimientos subvencionados.

3.2.2 Corpus

Esta muestra está compuesta por tres corpus fundamentales producidos por cada grupo de sujetos y divididos, además, en los tres ámbitos temáticos de estudio.

I – Corpus de noticias elaboradas por estudiantes de primer año de periodismo. Incluye tres dominios –política, deporte y ciencia y tecnología– y está compuesto por 121 textos.

II – Corpus de noticias elaboradas por estudiantes de cuarto año de periodismo. Incluye tres dominios –política, deporte y ciencia y tecnología–, compuesto por 44 textos.

III – Corpus de noticias publicadas en medios de comunicación escritas por periodistas profesionales. Incluye tres dominios –política, deporte y ciencia y tecnología–, compuesto por 97 noticias.

Como información complementaria se incluyó el tipo administración del establecimiento educacional del cual egresó cada sujeto.

3.2.3 Variables independientes

La investigación queda conformada por las variables predictoras[5] presentes en la Tabla 2.

TABLA 2 – Variables independientes y sus valores

Variable	Etiqueta	Valores
SEXO	*Género*	0 = Hombre
		1 = Mujer
GRUPO	Grupo al que pertenece la muestra	1 = G1 Primer año
		2 = G2 Cuarto año
		3 = G3 Profesionales
EDUC	Pertenencia administrativa de institución educacional de la cual egresó en enseñanza media	1 = Privado
		2 = Municipal
		3 = Subvencionado
TEMA	Temáticas en las cuales de realizaron las noticias	1 = Ciencia y Tecnología
		2 = Política
		3 = Deporte

3.2.4 Variables dependientes

Las variables dependientes en los estudios de riqueza léxica fueron operacionalizadas de la siguiente manera:

* **Diversidad léxica total:** implica la relación entre la extensión del texto y el total de palabras diferentes de dicho texto. Variable cuantitativa.

[5] Debido a que este estudio es una investigación no experimental, lo más adecuado es utilizar la terminología *variables predictoras*; mientras que las variables independientes utilizan la terminología *variables criterio*, también denominadas respuesta o resultado (HERRERA; MARTÍNEZ; AMENGUAL, 2011).

* **Diversidad léxica aleatoria:** implica la relación entre la extensión del texto y el total de palabras diferentes de dicho texto, esta vez tomando tres muestras aleatorias de cada texto y promediando sus resultados. Variable cuantitativa.
* **Diversidad léxica del *lead*:** implica la relación entre las 250 primeras palabras del texto y las palabras distintas contenidas en esa extensión. En una noticia informativa esta variable se considera de relevancia debido a que, teóricamente, la mayor información de una noticia se encuentra contenida en esa primera parte de la estructura textual.
* **Densidad léxica**: relación entre las palabras de contenido semántico y la extensión del texto. Es una variable cuantitativa.

3.3 Pruebas y mediciones

Para realizar el estudio sobre riqueza léxica se utilizaron dos vías: en primer lugar, evaluar la riqueza léxica de los textos producidos por los estudiantes y los profesionales y, en segundo lugar, identificar si existen diferencias –y de existir cuáles son– entre ambos grupos.

La riqueza léxica de los textos se obtuvo por medio de dos aproximaciones cuantitativas, previamente abordadas en el marco teórico, cuales son la diversidad léxica, es decir, la relación entre la extensión total y los vocablos distintos de un texto; y, por otra parte, la densidad léxica de los mismos, es decir, la relación entre palabras con contenido semántico y, nuevamente, la extensión del texto.

3.3.1 Procedimiento para medir las variables diversidad léxica y densidad léxica

La evaluación de la diversidad léxica siguió la línea tradicional de estudio, es decir, se utilizó la fórmula *type-token ratio*, que en términos sencillos significa obtener la relación entre el total de palabras distintas (*type*) y el total de palabras (*token*), indicado en el apartado "Índices de riqueza léxica", expuestos con anterioridad.

Este procedimiento se efectuó en cada uno de los textos, según la temática del corpus y el grupo muestral de sujetos. Además, se analizaron los textos en cada grupo según la extensión de la noticia, evitando así lo que señala la evidencia de diversas investigaciones, con relación a los índices de relación entre las palabras distintas y el total de palabras que disminuyen a

medida que aumenta la extensión de un texto. Esto se abordó en el apartado teórico y se refiere a la distorsión que puede ocurrir con los datos al comparar textos con distintas extensiones.

4 Resultados

El primer objetivo de este apartado es exponer los resultados obtenidos en un nivel descriptivo de la riqueza léxica de los textos, con la idea de presentar una mirada general del corpus. Se expondrán los datos ordenados según género, grupo y pertenencia al establecimiento educacional en cada una de las temáticas estudiadas. Sin embargo, y para entregar una primera aproximación en términos numéricos globales, es importante indicar que el corpus de los estudiantes de primer y cuarto años está compuesto por 106.286 tokens, 12.698 correspondían a types o palabras únicas, y de ellas, 12.541 a palabras nocionales (PN) o con contenido semántico. En el caso de los profesionales, el corpus está compuesto por 37.989 tokens, de estos 6.751 son palabras únicas y 6.602 corresponden a palabras nocionales.

Debido a que los textos fueron recolectados en el contexto de una asignatura práctica, no todos los estudiantes cumplieron con entregar un texto de cada tema, es por ello que en la Tabla 3 se presentan los detalles del corpus según grupo ordenado por cada temática.

TABLA 3 – Resumen de casos según grupo y corpus

	Grupo	Casos					
		Válidos		Perdidos		Total	
		N	Porcentaje	N	Porcentaje	N	Porcentaje
Corpus Ciencia	G1	47	81%	11	19%	58	100%
	G2	11	55%	9	45%	20	100%
	G3	20	100%	0	0%	20	100%
Corpus Deporte	G1	38	65,5%	20	34,5%	58	100%
	G2	20	100%	0	0%	20	100%
	G3	20	100%	0	0%	20	100%
Corpus Política	G1	55	94,8%	3	5,2%	58	100%
	G2	20	100%	0	0%	20	100%
	G3	20	100%	0	0%	20	100%

Como se indica en la Tabla 3, en el corpus "Ciencia", 47 de los 58 alumnos de primer año que participaron en el estudio entregaron los textos requeridos y, por ende, forman parte del corpus; en el caso de los estudiantes de cuarto año 11 de los 20 alumnos lo cumplieron y, por último, los 20 profesionales entregaron sus noticias sin excepción. El corpus "Deporte" quedó conformado por los textos de 38 estudiantes de primer año, de 20 de cuarto año y de 20 profesionales; éste fue el corpus en el que participó un menor número de estudiantes de primer año. Finalmente, en el corpus "Política" se recibieron las noticias de 55 estudiantes de primer año y las de los 20 participantes de cuarto año, así como de los 20 que pertenecían al grupo de los profesionales.

En tanto, el ordenamiento según sexo en cada una de las temáticas quedó establecido según lo indicado en la Tabla 4. En el estudio participaron 39 hombres y 59 mujeres. El desglose en este nivel de la muestra, como se presenta en la Tabla 5, señala que de los casos válidos, es decir, de quienes entregaron sus textos y se incluyeron en el corpus "Ciencia", 31 son hombres (79,5%), y 47, mujeres (79,7%). En "Deporte", el 76,9% son varones (30 participaron) y el 81,4%, damas (48); mientras que en el corpus "Política", los 37 hombres que facilitaron sus noticias corresponden al 94,9% de los casos válidos, y las 58 mujeres que participaron corresponden al 98,3% de la muestra.

TABLA 4 – Resumen de casos según sexo y corpus

	Sexo	Casos					
		Válidos		Perdidos		Total	
		N	Porcentaje	N	Porcentaje	N	
Corpus Ciencia	Hombre	31	79,5%	8	20,5%	39	100%
	Mujer	47	79,7%	12	20,3%	59	100%
Corpus Deporte	Hombre	30	76,9%	9	23,1%	39	100%
	Mujer	48	81,4%	11	18,6%	59	100%
Corpus Política	Hombre	37	94,9%	2	5,1%	39	100%
	Mujer	58	98,3%	1	1,7%	59	100%

Un tercer punto que se incluyó fue el tipo de establecimiento educacional (privado, municipal o subvencionado) del cual egresaron los sujetos que participaron en la muestra.

En "Ciencia" fueron 23 los sujetos de establecimientos privados que contribuyeron con sus textos, 24 sujetos provenientes del sistema municipal y 31 del subvencionado. En "Deporte" del total de sujetos, 28 provenían de instituciones educacionales privadas, 16 de establecimientos con dependencia administrativa municipal y 34 de subvencionados. En "Política", 28 participantes eran de colegios privados, 26 de establecimientos municipalizados y 41 sujetos de instituciones subvencionadas.

Finalmente, la Tabla 5 presenta el recuento del corpus según el *cluster* de extensión al cual pertenece y muestra que más de la mitad de los textos (55,6%) se encuentran en X1, es decir, tienen una extensión que va de 250 a 500 palabras. Esta exposición de datos cobra relevancia para comprender el real alcance de las comparaciones debido a que se quiere evitar distorsiones en las mediciones es que se debe conocer a qué *cluster* de extensión corresponde cada noticia.

TABLA 5 – Recuento según extensión de las noticias

		Frecuencia	Porcentaje
Válido	X1 = 250-500	145	55,6%
	X2 = 501-750	63	24,1%
	X3 = 751-1000	34	13%
	X4 = 1001+	19	7,3%
	Total	261	100%
Total		261	100

4.1 Estadísticos descriptivos y pruebas estadísticas

4.1.1 Variable Sexo

Los datos descriptivos de la variable independiente "Sexo" fueron obtenidos del análisis global de los textos. Es decir, se realizó un análisis del total de los textos que componen el corpus de noticias (261 textos), en sus tres ámbitos temáticos. La relación de Sexo con la variable dependiente **Densidad léxica**, indicador que mide la relación de palabras con contenido semántico dentro del texto, muestra que la media obtenida por los hombres (0,42) no difiere en gran medida del obtenido por las mujeres (0,43); aunque la varianza es mayor en el grupo de mujeres (0,008) que en el de los hombres

(0,005). Sin embargo, la desviación estándar de ambos tiene solo una diferencia de 0,01 (Tabla 6).

TABLA 6 – Estadígrafos descriptivos densidad léxica

Sexo		Estadístico	SEM[6]
Hombre	Media	0,4221	0,00719
	Varianza	0,005	
	Desviación estándar	0,07264	
Mujer	Media	0,4361	0,00696
	Varianza	0,008	
	Desviación estándar	0,08802	

Los resultados descriptivos de la variable dependiente diversidad léxica, que mide la relación entre el número de palabras diferentes y el total de los textos, se encuentran descritos en la Tabla 7. Éstos indican que la media fue la misma tanto en hombres como mujeres (0,49), lo que implica que los promedios de riqueza léxica en el indicador diversidad léxica total de las noticias –es decir, sin diferenciar entre la extensión de los textos ni la temática–, tanto hombres como mujeres, no presentan diferencias. Lo mismo ocurre en la varianza, donde ambas varían en 0,004; y en su desviación estándar, que corresponde a 0,06.

Al analizar los estadísticos desde la perspectiva del impacto de la variable independiente Sexo en la variable dependiente Diversidad léxica aleatoria[7] (Tabla 7), las medias obtenidas tanto en hombres como mujeres vuelven a presentar valores similares: 0,60 en hombres y 0,59 mujeres; así como sus varianzas y la desviación estándar. Si bien ambas mediciones de diversidad presentan medias con resultados disímiles, la tendencia entre ambas es la misma: las diferencias no son superlativas para ningún sexo.

[6] SEM = Standard Error of the Mean (Error estándar de la media).

[7] Es decir, la diversidad léxica medida al azar con tres muestras de 200 palabras en cada texto.

TABLA 7 – Sexo y Diversidad léxica

Sexo		Diversidad léxica total		Diversidad léxica aleatoria	
		Estadístico	SEM	Estadístico	SEM
Hombre	Media	0,4988	0,00612	0,6063	0,0038
	Varianza	0,004		0,001	
	Desviación estándar	0,06178		0,03842	
Mujer	Media	0,491	0,00497	0,5963	0,00302
	Varianza	0,004		0,001	
	Desviación estándar	0,06292		0,0382	

4.1.1.1 Prueba no paramétrica para la variable Sexo

Una vez presentados los resultados descriptivos para esta variable, se propuso la indagación de la hipótesis:

> *H1: El Sexo es una variable que incide en la riqueza léxica que presentan los sujetos en sus textos periodísticos.*

Para comprobar si existe un efecto de la variable Sexo, se efectuó la prueba estadística U Mann-Whitney, apropiada para el estudio de muestras independientes cuyos resultados no tienen una distribución normal –como ocurre en este caso–, y que calcula las diferencias estadísticas entre las medianas de los resultados, para así explorar la existencia de diferencias estadísticas entre hombres y mujeres.

La prueba considera que, al no existir diferencias significativas en los resultados de sus medianas para ninguna de las variables dependientes: Densidad léxica, Diversidad léxica y Diversidad léxica aleatoria; se debe aceptar la hipótesis nula:

> *H0: El Sexo no es una variable que incide en la riqueza léxica que presentan los sujetos en sus textos periodísticos.*

4.1.2 Variable Grupo

A continuación, se exponen los estadígrafos descriptivos analizados según la variable "Grupo". Al igual que lo anterior, esta variable fue analizada con la totalidad del corpus, sólo organizado mediante los tres grupos que participaron con sus noticias, sin distinción de la temática de éstas.

TABLA 8 – Estadígrafos descriptivos densidad léxica según Grupo

Grupo		Estadístico	Error estándar
G1	Media	0,3697	0,00363
	Varianza	0,002	
	Desviación estándar	0,03989	
G2	Media	0,4355	0,00975
	Varianza	0,004	
	Desviación estándar	0,06468	
G3	Media	0,5045	0,00676
	Varianza	0,004	
	Desviación estándar	0,06662	

Al describir la variable Densidad léxica según el Grupo que produjo el corpus (Tabla 8), se obtiene que el G1 (estudiantes de primer año) presenta una media de 0,36 –la más baja entre las tres–, seguido por G2 (estudiantes de cuarto año), con 0,43 como media de densidad léxica y, finalmente, el grupo conformado por los profesionales, G3, tiene una media de 0,5 en su densidad léxica.

En cuanto a la diversidad léxica total, que mide la relación de la diversidad con la extensión total de las noticias, los estadígrafos descriptivos señalan que el G1 tiene un promedio de diversidad total del texto de 0,47; en cambio el G2 es el grupo que tiene mejor puntaje de diversidad con un promedio grupal de 0,53; luego en G3, los profesionales muestran una media en diversidad léxica de 0,5.

Finalmente, en la variable Densidad léxica aleatoria (Tabla 9), similar a la anterior pero utilizando el procedimiento explicado en el apartado "Pruebas y mediciones", los resultados si bien se modifican en términos numéricos de medias, es posible que lo ocurra debido a que en este análisis no se dividen los textos según su extensión, aspecto que fue abordado en el marco teórico y que es un factor en la distorsión de los datos al analizar cuantitativamente la riqueza léxica. Las medias muestran que en G1 la diversidad léxica es de 0,59; en G2, 0,63 y en G3, 0,59.

TABLA 9 – Estadígrafos descriptivos diversidad léxica aleatoria según Grupo

Grupo		Estadístico	Error estándar
G1	Media	0,5921	0,00298
	Varianza	0,001	
	Desviación estándar	0,03277	
G2	Media	0,6320	0,00508
	Varianza	0,001	
	Desviación estándar	0,03373	
G3	Media	0,5959	0,00408
	Varianza	0,002	
	Desviación estándar	0,04022	

Para abordar la posible distorsión de los datos obtenidos con anterioridad, se presenta en la Tabla 10 un análisis del *lead*[8] o entradilla de las noticias que componen cada corpus. Los resultados indican que los promedios aumentan y superan en los tres grupos la media de 0,65. El G1 presenta en este indicador el promedio más alto obtenido de todas las variables de diversidad, alcanzando una media de 0,67. El G2, por su parte, también tiene el promedio más alto, con una media de 0,74; mientras que el G3 presenta una media de 0,66. Lo interesante de este estadígrafo descriptivo es que si bien el G3 tiene los promedios más altos en las mediciones anteriores, en este indicador queda rezagado al relacionarlo con los otros grupos. Aun así, se debe señalar que este grupo contiene promedios similares en todos los indicadores de diversidad, situación que no ocurre con los otros dos grupos que varían considerablemente en unos indicadores de otros.

Esta situación podría ser resultado de la práctica y experiencia en la redacción de noticias escritas. El grupo compuesto por estudiantes de primer año de la carrera de periodismo está por primera vez elaborando textos de hechos noticiosos bajo una estructura determinada. Son capaces de elaborar un *lead* con una diversidad rica, así como textos que, en general,

[8] Según la estructura canónica de pirámide invertida de una noticia, se debe entregar la información más relevante del hecho noticioso en la parte que comprende el titular de la noticia y el primer párrafo. Esto quiere decir que en el inicio de un texto la mayor diversidad de palabras se encuentra dada en los primeros párrafos, decreciendo a medida que aumenta la extensión del texto.

tienen una variedad léxica en la que 6 de cada 10 palabras son diferentes, demostrando que son capaces de generar una alta diversidad en sus textos. Sin embargo, en el indicador de densidad léxica, es decir, cuando la variedad se mide por palabras distintas, pero con contenido semántico y su relación con la extensión total del texto, sus promedios caen sustancialmente y no superan la diversidad de 4 palabras diferentes por cada 10 presentes en la noticia. Esta probable explicación se puede contrastar con lo que ocurre en el G2, estudiantes de cuarto año, quienes presentan mejores índices que sus pares de primer año en todos los indicadores.

Por otro lado, hay que tener a consideración algunos aspectos de formalidad en la elaboración de las noticias redactadas por ambos grupos. Mientras que al G1 se le solicitó el texto como parte de una tarea de un curso de producción de textos; el G2 debió elaborar sus noticias para una asignatura cuyo objetivo es publicar, semanalmente, una revista de noticias miscelánea, en un ambiente similar al que se vivencia en un medio de comunicación.

TABLA 10 – Estadígrafos descriptivos diversidad léxica del *lead* según Grupo

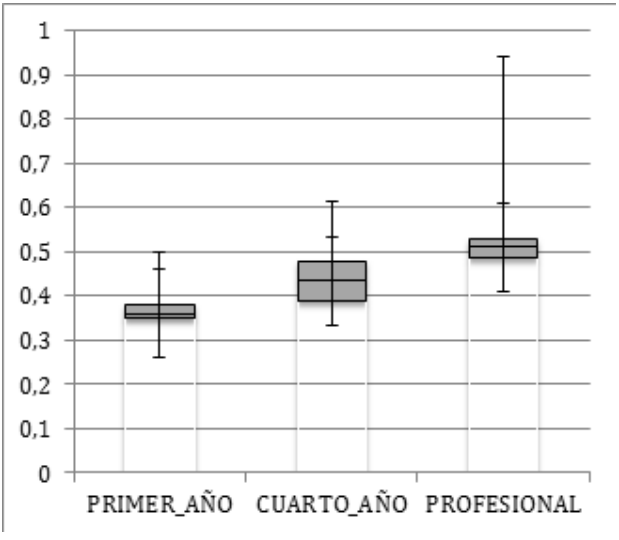
Grupo		Estadístico	Error estándar
G1	Media	0,6773	0,00545
	Varianza	0,004	
	Desviación estándar	0,05999	
G2	Media	0,7436	0,00591
	Varianza	0,002	
	Desviación estándar	0,03918	
G3	Media	0,6688	0,00593
	Varianza	0,003	
	Desviación estándar	0,05842	

4.1.2.1 Pruebas no paramétricas según Grupo

Mediante la prueba estadística no paramétrica de Kruskal-Wallis se puede comprobar que los grupos que se están estudiando son distintos significativamente. Constituye una alternativa no paramétrica al uso del análisis de varianza de un factor (Anova1) y, así como en ésta, las muestras pueden ser de tamaños diferentes.

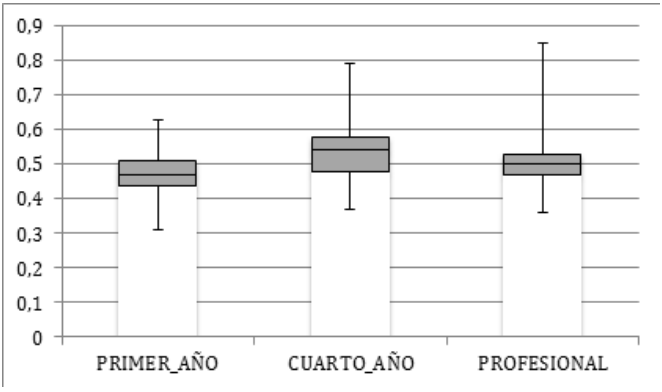
Se pudo comprobar que existen diferencias significativas entre los grupos y la riqueza léxica de sus textos, tanto en la variable diversidad léxica como en la variable densidad léxica, al desechar en ambos casos la hipótesis nula sobre la no existencia de diferencias significativas en la riqueza léxica en los textos producidos por estudiantes de primer año, de cuarto año y por los profesionales, hecho que se comprueba al medir los cuatro indicadores de riqueza léxica. Según lo presentado, existen diferencias estadísticas significativas no atribuibles al azar y, por ende, se rechazan la hipótesis nula indicada en la misma tabla, tras comparar las medianas de cada grupo.

FIGURA 1 – Diferencias de densidad léxica según Grupo



El Figura 1 presenta las diferencias existentes entre los tres grupos que participaron con sus noticias escritas. Los datos indican que, efectivamente, existen diferencias significativas entre las muestras y que la variable predictora Grupo es un indicador de que los estudiantes de primer año cuentan con una riqueza léxica, en el plano de la densidad, menor a la de los estudiantes de cuarto año y de los profesionales. Por otra parte, evidencia una alta dispersión en el resultado de la densidad dentro de sus propios grupos, principalmente entre los estudiantes de primer año y los profesionales. Aun así, los puntajes límites, es decir, los más altos y más bajos tienen una mejor puntuación entre los profesionales.

FIGURA 2 – Diferencias según Grupo en diversidad léxica



Con la Figura 2 se expone, nuevamente, que existe una diferencia estadísticamente significativa entre las medias de cada grupo y que dicha diferencia no corresponde al azar. En esta oportunidad, a diferencia del indicador de densidad léxica, lo que se aprecia es que el grupo conformado por estudiantes de cuarto año presenta un mayor puntaje. Por su parte, los profesionales, si bien no presentan una diversidad mayor que el grupo anterior, sí presentan una menor dispersión en sus puntajes, lo que evidencia a un grupo más cohesionado.

4.1.3 Variable Tipo de establecimiento educacional

En Chile, la dependencia administrativa de las instituciones que componen el sistema educacional se divide en tres grandes grupos. El primero de ellos es el sistema público, en el que las instituciones educacionales (colegios, escuelas, liceos) dependen de fondos del Estado para funcionar. Estos fondos son administrados por los municipios, por medio de las direcciones de educación municipal, que tienen por objetivo procurar las condiciones óptimas para un eficiente y eficaz desarrollo del proceso educativo en estos establecimientos. El otro gran grupo –el que, debido a las modificaciones de la Ley de Educación en Chile, está próximo a desaparecer– es la educación particular subvencionada. Las instituciones que dependen de este tipo de administración son responsabilidad de un ente sostenedor que recibe aportes o subsidios del Estado, los cuales deben ser cofinanciados por la comunidad que asiste a dichos establecimientos. El tercer grupo está compuesto por las instituciones privadas que se sostienen con el pago particular de la matrícula y aranceles.

Debido a las características descriptivas de esta investigación, se propuso abordar el aspecto socioeconómico como una variable que influye en la competencia léxica de los estudiantes. Sin embargo, puesto que el constructo socioeconómico es una variable compleja, que aborda elementos financieros del grupo familiar de cada sujeto, la escolaridad de los padres, junto a otro tipo de datos, se decidió constatar sólo la dependencia administrativa de la institución de egreso de los participantes de la muestra, entendiendo que esto puede significar algún tipo de sesgo en el resultado del efecto de esta variable.

TABLA 11 – Estadísticos descriptivos de densidad léxica

Sistema educacional		Estadístico	Error estándar
Privada	Media	0,46	0,01
	Varianza	0,008	
	Desviación estándar	0,09	
Municipal	Media	0,40	0,009
	Varianza	0,006	
	Desviación estándar	0,078	
Subvencionada	Media	0,42	0,006
	Varianza	0,004	
	Desviación estándar	0,06	

La Tabla 11 expone las medias de la muestra ordenadas según el tipo de establecimiento educacional. Tal como se explicó anteriormente, se encuentra dividida en los tres tipos de dependencia administrativa. La densidad léxica que presenta el grupo de egresados del sistema privado muestra una media de 0,46; la sigue en promedio los egresados del sistema subvencionado con una media de 4,42; y luego los egresados del sistema municipal, con una media de 0,4. En el caso del efecto de la variable dependencia administrativa de los establecimientos educacionales de egreso de enseñanza media, sólo se considerará esta tabla de estadígrafos descriptivos ya que, a continuación, en el punto *Pruebas no paramétricas*, no hay una diferencia significativa en las otras variables dependientes para medir la riqueza léxica.

4.1.3.1 Prueba no paramétrica para el tipo de establecimiento educacional

La prueba no paramétrica de Kruskal-Wallis señala que las diferencias estadísticamente significativas se encuentran en los indicadores diversidad léxica total y densidad léxica. Sin embargo, hay que reconocer que el nivel de significación de la variable diversidad léxica total se encuentra muy cercano al límite de significación de 0,5 utilizado en las investigaciones en ciencias sociales y humanidades, por lo tanto, no es de extrañar que en la medida diversidad léxica aleatoria, la prueba no alcance la significancia necesaria para establecer estadísticamente una diferencia significativa.

4.1.4 Variable Extensión

Para comprender las comparaciones realizadas en esta variable, se debe señalar que los análisis se efectuaron con los corpus *X1* (con menos de 500 palabras) y *X2* (entre 501 y 750 palabras), debido a que en los restantes la cantidad de texto fue insuficiente para obtener datos concluyentes ya que al trabajar con sus promedios podrían alterar la tendencia de los resultados, induciendo al error en la interpretación.

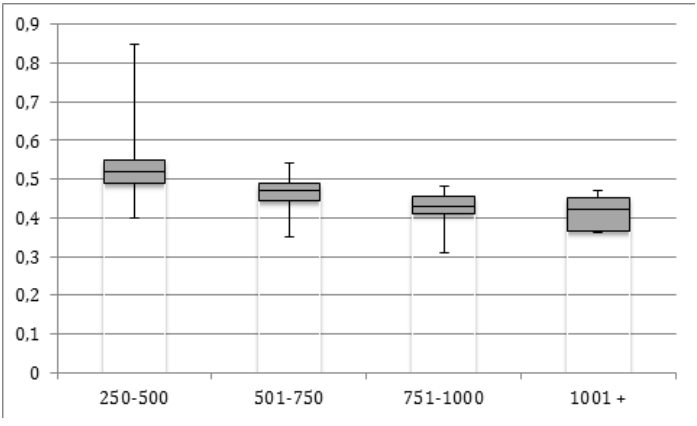
Como se ha visto, la extensión del texto constituye una variable predictora ante la cual las variables dependientes son sensibles. No es de extrañar que en cada uno de los *clusters* propuestos se hayan presentado diferencias significativas en los diversos indicadores de riqueza léxica, tal como se muestra en la Tabla 12, lo que confirma los hallazgos de estudios previos que indican que en textos de menor extensión los índices de diversidad léxica son mayores que en textos más extensos.

TABLA 12 – Estadísticos descriptivos de diversidad léxica, según Extensión

Extensión		Estadístico	Error estándar
X1 (250-500)	Media	0,5426	0,00407
	Varianza	0,002	
	Desviación estándar	0,04898	
X2 (501-750)	Media	0,4846	0,00513
	Varianza	0,002	
	Desviación estándar	0,04075	
X3 (751-1000)	Media	0,4582	0,00450
	Varianza	0,001	
	Desviación estándar	0,02622	
X4 (1001+)	Media	0,4153	0,01186
	Varianza	0,003	
	Desviación estándar	0,05168	

Mientras que en *X1* la media supera el 50% de variedad de palabras, en *X4* apenas supera la relación de cuatro palabras diferentes por cada diez presentes en el texto. El descenso de este indicador es consistente en la medida que aumenta la extensión del texto. Este hecho queda mejor explicado en el Figura 3.

FIGURA 3 – Relación entre extensión y diversidad léxica total



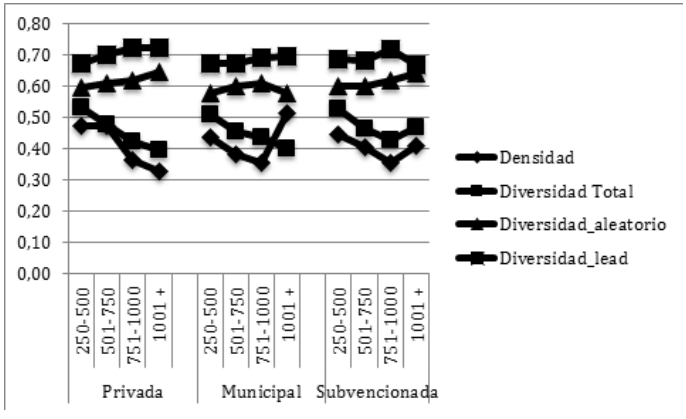
Un comportamiento similar ocurre con el indicador de densidad léxica, que disminuye en textos más extensos, aunque de una manera menos

marcada. El promedio de los textos del *cluster X1* supera el 40% de variedad de palabras, mientras que esta relación disminuye a medida que el texto es más extenso. El error en estos estadígrafos es bajo, lo que se ve reflejado, a la vez, en la varianza de cada *cluster*, lo que implica que se trata de una medida muy estable, cuyos valores se encuentran concentrados en torno a su media.

Al igual que la diversidad, la densidad léxica es también susceptible a la extensión de los textos. Lo anterior se expresa en las medias obtenidas por cada grupo de extensión, cuyos datos estadísticos indican una varianza –es decir, el grado en el que los resultados se alejan de la media del grupo– que no supera el 0,002, lo que indica una baja dispersión de los resultados.

Una mirada más completa se presentará en la Figura 4, donde se muestra la relación entre las medidas de riqueza léxica obtenidas del corpus y el tipo de establecimiento educacional del cual egresó el sujeto que escribe la noticia, esta vez tomando como valor la extensión de los textos. Se puede observar que el promedio en los cuatro indicadores de riqueza léxica –en densidad léxica, en diversidad léxica total, en diversidad léxica aleatoria promedio y en diversidad léxica del *lead*– es superior entre quienes egresaron de un establecimiento educacional particular, decreciendo el promedio en los egresados de educación municipal y volviendo a aumentar entre quienes egresaron de educación particular subvencionada. Se debe indicar que en los datos obtenidos del *cluster* X4, el promedio de cada indicador no refleja la tendencia del resto de los *clusters,* ya que los datos de *X4* corresponden solo al 7,3% del total del corpus.

FIGURA 4 – Indicadores de riqueza léxica relacionados con extensión y tipo de establecimiento educacional



5 Conclusiones

En resumen, se efectuaron mediciones basadas en los indicadores diversidad léxica y densidad léxica con la variable *Sexo*, en que no hubo diferencias significativas, lo que fue comprobado por pruebas estadísticas no paramétricas. La variable que sí incidió estadísticamente en los resultados fue el efecto de *Grupo* como variable predictora. Las diferencias, no atribuibles al azar o a la coincidencia, indican que en términos de densidad léxica los estudiantes de los dos períodos evaluados y los profesionales presentan diferencias significativas; también ocurre al analizar los indicadores de diversidad léxica total, diversidad léxica aleatoria y diversidad léxica del *lead*.

Otro indicador que es sensible en este estudio es la variable *Dependencia administrativa del establecimiento educacional*, especialmente en los indicadores densidad léxica y diversidad léxica total. En esta variable quedó comprobado que, para esta muestra, quienes egresaron de una institución privada cuentan con una diversidad léxica y densidad léxica mayores de la que presentan los egresados de enseñanza municipalizada y subvencionada.

En el plano de la riqueza léxica se puede describir al grupo de primer año, G1, como un grupo capaz de elaborar textos con un alto nivel de diversidad léxica, especialmente cuando la medición se realiza con el sistema de diversidad léxica aleatoria –explicada en el apartado de la Metodología–, índice bajo el cual se acercan, en promedio, a un TTR = 0,6 en los tres ámbitos temáticos abordados: *ciencia y tecnología*, *política* y *deporte*. Por otra parte, este grupo presenta una alta diversidad léxica en el *lead*, que sobrepasa el TTR = 0,65, lo que implicaría una entrada de la noticia con una importante fuente informativa, ya que se podría suponer que, en esas 200 primeras palabras, los estudiantes son capaces de organizar una serie significativa de hechos.

El problema surge cuando el índice de riqueza léxica es abordado también por la densidad léxica de un texto. Se ha definido como densidad léxica la relación que existe entre el total de palabras de contenido semántico y la extensión de un texto. Al remover todas las palabras gramaticales –artículos, conjunciones, y verbos auxiliares, entre otras categorías– en el texto solo quedan vocablos con significado, los que contienen la información relevante de un hecho. Este indicador es en G1 = 0,3; es decir, por cada 10 palabras que hay en un texto solo 3 son palabras de contenido que aportan, en este caso, a la función informativa del texto noticioso.

Bajo estos mismos indicadores, el G2, compuesto por estudiantes de cuarto año, demuestra una riqueza léxica más elevada en sus textos, tanto en la diversidad léxica aleatoria como en la diversidad léxica en el *lead*, de

las noticias que elaboran. Por otra parte, la densidad léxica es 0,53. Si lo visualizamos en términos porcentuales, este grupo tiene aproximadamente un 20% mejor reporte en sus noticias escritas debido a que su utilización de palabras con contenido semántico se ve más equilibrada con relación a las palabras gramaticales.

Dentro de los objetivos específicos de este trabajo, se propuso comparar la riqueza léxica de los estudiantes con la de profesionales. Al respecto, los índices que arroja el G3, compuesto por periodistas, no parecían muy distanciados de los obtenidos por los estudiantes; incluso en la diversidad léxica del *lead* fue menor que lo obtenido por los estudiantes de cuarto año (G2). Los profesionales marcan una diferencia significativa en el índice de densidad léxica, en que G1 = 0,33; G2 = 0,43 y G3 = 0,5.

Cuando se aborda la competencia de estudiantes de periodismo, en términos de riqueza léxica en textos de un discurso especializado, éstos se pueden caracterizar como estudiantes cuya competencia léxica les permite elaborar noticias informativas con un nivel de diversidad léxica medio-alto, con buenos promedios de diversidad especialmente en la elaboración de *lead* noticioso, situación que se mantiene entre los estudiantes de cuarto año. La diferencia radica en la utilización de palabras con contenido que aporten informativamente al texto que elaboran. Se puede hipotetizar, entonces, que los textos periodísticos tienen una configuración lingüística especial, que debe adecuarse para transmitir la mayor cantidad posible de información en textos de extensión reducida, imposición que los convierte precisamente en prototipos de los textos informativamente densos, a los cuales los estudiantes, principalmente los de primer año, no logran alcanzar el grado de destreza comunicativa necesaria para entregar la información sustancial de un hecho. Los resultados de este trabajo sostienen dicha hipótesis: los textos elaborados por estudiantes, especialmente por los de primer año, no logran cumplir con un requisito comunicativo adecuado. Sin embargo, se puede inferir que, en el transcurso de su formación profesional, el alumno de periodismo mejora esta capacidad hasta llegar a la que un periodista profesional debe ser capaz de demostrar.

Contribución de los autores

Karina Fuentes Riffo: autora principal, realiza la fundamentación teórica, el diseño del *script* que evalúa de manera automática la diversidad y la densidad léxica del corpus. Realiza de manera colaborativa el diseño metodológico de la investigación, así como el análisis de los resultados y la elaboración de las conclusiones. Efectúa la redacción y edición del artículo.

Dr. Sergio Hernández Osuna: colabora con la elaboración del *script* que evalúa de manera automática la riqueza y densidad léxicas. Realiza la toma de muestra de los textos y coopera con la evaluación de estos.
Dr. Pedro Salcedo Lagos: realiza en conjunto con la autora principal el diseño metodológico de la investigación, el análisis estadístico de los resultados y las conclusiones. Trabaja en la redacción y edición del artículo.

Referencias

CABRÉ, M. *La terminología*: teoría, metodología, aplicaciones. Barcelona: Antártida, 1993.

DALLER, H.; VAN HOUT, R.; TREFFERS-DALLER, J. Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, Cambridge, v. 24 n. 2, p. 127-222, 2003. Doi:10.1093/applin/24.2.197

ECHEVERRÍA, M. Noticias y deportes en el español público de Chile. *In*: CONGRESO INTERNACIONAL DE LA LENGUA ESPAÑOLA, 1., 1997, Zacatecas. *Anais* […]. Madrid: Instituto Cervantes, 1997. Disponible en: https://bit.ly/2Db0Rs5. Acceso el: 10 abr. 2019.

GREGORY-SIGNES, C.; CLAVEL-ARROITÍA, V. Analyzing lexical density and lexical diversity in university students' written discourse. *Procedia – Social and Behavioral Science*, [*S.l.*], v. 198, p. 546-556, 2015. Doi:10.1016/j.sbspro.2015.07.477

GÓMEZ, J. La competencia léxica en el currículo de español para fines específicos. *In*: CONGRESO INTERNACIONAL DE ESPAÑOL PARA FINES ESPECÍFICOS, 2., 2002, Amsterdam. *Anais* […]. Amsterdam: Embajada de España, 2002. Disponible en: https://bit.ly/2D8Mkx7. Acceso el: 10 abr. 2019.

HERRERA, H.; MARTÍNEZ, R.; AMENGUAL, M. Estadística aplicada a la investigación lingüística. Madrid: EOS Universitaria, 2011.

JIMÉNEZ, R. El concepto de competencia léxica en los estudios de aprendizaje y enseñanza de segundas lenguas. *Atlantis English Studies*, Madrid, v. 24, n. 1, p. 149-162, 2002. Disponible en: https://bit.ly/2UJWgXX. Acceso el: 10 abr. 2019.

JOHANSSON, V. Lexical diversity and lexical density in speech and writing: a develop-mental perspective. *Working Papers in Linguistics*, Lund, v. 53, p. 61-79, 2008. Disponible en: https://bit.ly/2GfxTcx. Acceso el: 10 abr. 2019.

MARTÍN VIVALDI, G. *Géneros periodísticos.* Madrid: Paraninfo 1993.

MARTÍN VIVALDI, G. *Curso de redacción: teoría y práctica de la composición y del estilo.* Madrid: Paraninfo 2004.

MARTÍNEZ ALBERTOS, J. M. *El lenguaje periodístico*: estudios sobre el mensaje y la producción de textos. Madrid: Paraninfo, 1989.

MARTÍNEZ ALBERTOS, J. M. *Curso general de redacción periodística*. 5. ed. Madrid: Paraninfo, 1991.

MCCARTHY, P.; JARVIS, S. Vocd: a theoretical and empirical evaluation. *Language Testing*, California, v. 24, n. 4, p. 459-488, 2007. Doi:10.1177/0265532207080767

MCCARTHY, P.; JARVIS, S. MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, [*S.l.*], v. 42, n. 2, p. 381-392, 2010. Doi:10.3758/BRM.42.2.381

LÓPEZ MORALES, H. Los índices de 'Riqueza léxica' y la enseñanza de lenguas. In: DEL TEXTO A LA LENGUA: LA APLICACIÓN DE LOS TEXTOS A LA ENSEÑANZA-APRENDIZAJE DEL ESPAÑOL L2-LE. Vol 1, 2002. Disponible en: https://dialnet.unirioja.es/descarga/articulo/5419218.pdf. Acceso el: 26 mayo de 2019.

PARODI, G. *Discurso especializado e instituciones formadoras*. Valparaíso: Ediciones Universitarias de Valparaíso, 2005a.

PARODI, G. La comprensión del discurso especializado escrito en ámbitos técnico-profesionales: ¿aprendiendo a partir del texto? *Signos*, Valparaíso, v. 38, n. 58, p. 221-267, 2005b. Doi: 10.4067/S0718-09342005000200005

PARODI, G. Discurso especializado y lengua escrita: foco y variación. *Estudios Filológicos*, Valdivia, n. 41, p. 165-204, 2006. Doi: 10.4067/S0071-17132006000100012

READ, J. Assessing vocabulary. 9. ed. London: Cambrigde University Press, 2010.

REYES, J. Riqueza léxica de textos redactados por alumnos del bachillerato de las Palmas Gran Canaria. *Anuario de Lingüística Hispánica*, Valladolid, v. XXIII-XXIV, p. 147-163, 2010.

TORRUELLA, J.; CAPSADA, R. Lexical statistics and tipological structures: a measure of lexical richness. *Procedia – Social and Behavioral Sciences*, [*S.l.*], v. 95, p. 447-454, 2013. Doi:10.1016/j.sbspro.2013.10.668

ŠIŠKOVÁ, Z. Lexical richness in EFL students' narratives. *Language Studies Working Papers*, Reading, v. 4, p. 26-36, 2012. Disponible en: http://bit.ly/2IdgbJ4. Acceso el: 10 abr. 2019.

Data de submissão: 24/08/2018. Data de aprovação: 18/03/2019.