



# Análisis de la riqueza léxica en tesis de grado de Ciencias Sociales y Humanidades (2007-2014)<sup>1</sup>

Marco Antonio Pérez Durán 

Universidad Autónoma de San Luis Potosí  
San Luis Potosí, México

## Resumen

Este trabajo es un estudio sobre la riqueza léxica de las tesis de grado de tres licenciaturas de Ciencias Sociales y Humanidades de la Universidad Autónoma de San Luis Potosí, México, del 2007 al 2014. La metodología aplicada es cuantitativa descriptiva no experimental. La muestra está conformada por la introducción y la conclusión de 125 tesis, las cuales se estudian a través de tres variables: género, licenciatura y año en relación con tres dimensiones de estudios de la riqueza léxica: densidad, diversidad y frecuencia. Los resultados muestran que las mujeres son quienes generaron más tesis en este periodo, que la licenciatura en Geografía es la que más tesis produjo y que 2010 y 2011 destacaron como los años con mayor generación de léxico de la tesis de grado.

**Palabras clave:** educación; frecuencia léxica; tesis de grado; *types*; *tokens*.

## Abstract

### Analysis of lexical richness in undergraduate theses in Social Sciences and Humanities (2007-2014)

This study examines the lexical richness of undergraduate theses from three-degree programs in Social Sciences and Humanities at the Universidad Autónoma de San Luis Potosí, Mexico, from 2007 to 2014. The applied methodology is quantitative, descriptive, and non-experimental. The sample consists of the introduction and conclusion sections of 125 theses which are studied through three variables: gender, degree and year in relation to three dimensions of lexical richness: density, diversity,

---

<sup>1</sup> Artículo de investigación  
*Lenguaje* 53(1), 20514606

<https://doi.org/10.25100/lenguaje.v53i1.14606>

and frequency. The results indicate that women produced the highest number of theses during this period, that the bachelor's degree in Geography produced the highest number of works and that 2010 and 2011 stood out as the years with the highest lexical generation in undergraduate theses.

**Keywords:** education; lexical frequency; undergraduate theses; *types*; *tokens*.

### **Résumé**

#### **Analyse de la richesse lexicale dans des mémoires de licence en Sciences sociales et humaines (2007-2014)**

Cette étude examine la richesse lexicale des mémoires de licence de trois programmes en Sciences sociales et humaines à l'Université Autonome de San Luis Potosí, Mexique, de 2007 à 2014. La méthodologie appliquée est quantitative, descriptive et non expérimentale. L'échantillon est constitué par l'introduction et la conclusion de 125 mémoires, qui sont étudiées à travers trois variables : genre, programme de licence et année en relation avec trois dimensions de la richesse lexicale : densité, diversité et fréquence. Les résultats montrent que les femmes ont produit le plus grand nombre de travaux au cours de cette période, que la licence en Géographie a été la plus prolifique en nombre de mémoires et que les années 2010 et 2011 se sont distinguées comme les années avec la génération lexicale la plus élevée dans les thèses de licence.

**Mots-clés :** éducation ; fréquence lexicale ; mémoires de licence ; *types* ; *tokens*.

## **SOBRE EL AUTOR**

### **Marco Antonio Pérez Durán**

Doctor en Lingüística por la Universidad Nacional Autónoma de México (UNAM). Es Profesor-Investigador de la Licenciatura en Lengua y Literatura Hispanoamericanas y profesor invitado en la Facultad de Psicología de la Universidad Autónoma de San Luis Potosí, México. Es miembro del Sistema Nacional de Investigadores, nivel I, y pertenece al Programa de Fortalecimiento del Profesorado. Su línea de investigación está centrada en la lexicología aplicada en la generación de corpus y en el análisis de discursos académicos.

Correo: marco.duran@uaslp.mx, C.P. 78394

## **CÓMO CITAR ESTE ARTÍCULO**

Pérez Durán, M. A. (2025). Análisis de la riqueza léxica en tesis de grado de Ciencias Sociales y Humanidades (2007-2014). *Lenguaje*, 53(1), e20514606. <https://doi.org/10.25100/lenguaje.v53i1.14606>

## INTRODUCCIÓN

En el ámbito universitario mexicano, la tesis de grado es un trabajo de investigación utilizado para obtener un grado académico al finalizar los estudios superiores y de grado. La tesis tiene como finalidad desarrollar y renovar el conocimiento de un tema desde una perspectiva innovadora y actualizada del fenómeno lingüístico a través de la generación de hipótesis que validen el resultado de la investigación.

Este trabajo se ha realizado en el marco del proyecto de investigación *Vocabulario académico universitario de la Universidad Autónoma de San Luis Potosí, México* por el Programa Estímulos al Desempeño en 2023, que, actualmente, se encuentra vigente y en el que se estudia la riqueza léxica de las tesis de distintas áreas disciplinares. El propósito de este proyecto general es contar con una base de datos del tipo lexical que eficiente el uso adecuado del léxico utilizado no solo para planes y programas de estudio, sino también para todo aquello formal-administrativo-académico que requiera emplear un léxico acorde con cada área de estudio.

Específicamente, este artículo estudia, desde el enfoque de la riqueza léxica, el léxico que aparece en dos secciones de la tesis de grado: introducción y conclusión. Se consideraron estas dos secciones debido a que 1) hay mayor libertad para utilizar un léxico académico y disciplinar propio del área de estudio por parte de quien escribe este texto; 2) en estas dos secciones se ve reflejada la voz de quien ha escrito el proyecto (Ibarra-López, 2017) y su punto de vista y 3) las unidades de estudio que conforman un corpus para su estudio son *types* y *tokens*. El objetivo de este artículo es hacer un estudio de tipo preliminar sobre la riqueza léxica (densidad, diversidad y frecuencia léxica) de la introducción y conclusión de 125 tesis de grado de tres licenciaturas pertenecientes a la Facultad de Ciencias Sociales y Humanidades de la Universidad Autónoma de San Luis Potosí, México, durante el periodo 2007 al 2014. Las licenciaturas de esta facultad pertenecen a las áreas de estudio que ocupan el tercer lugar de elección entre los estudiantes de México, solo detrás de las áreas de Salud y Económicas-Administrativas (INEGI, 2024)<sup>2</sup>.

La hipótesis de la que parte esta investigación es que, independientemente del número de tesis analizadas y conformadas como corpus de estudio, el número de *types-tokens* es parecido en las licenciaturas porque todas ellas comparten una base disciplinar en Ciencias Sociales y Humanidades y su variación léxica puede no ser tan diferente. Las preguntas que se tratarán de responder en este trabajo son ¿Cuántas palabras nocionales tiene una tesis de grado de estas disciplinas en estas dos secciones?, ¿Qué licenciatura presenta mayor riqueza léxica? y ¿Qué licenciatura presenta mayor variación de léxico por cada una de las variables?

Este trabajo presenta las siguientes implicaciones: 1) el análisis de las dos secciones de las tesis como corpus reflejará una primera aproximación real sobre el número de *types-tokens* utilizados por los universitarios en sus trabajos de grado, 2) el

---

<sup>2</sup> Instituto Nacional de Estadística y Geografía.

análisis de las dos secciones sienta un precedente real sobre la constitución de corpus léxicos disciplinares que ayudarían a la programación temática dentro del aula de clase y 3) hace más evidente la relación entre la lectura disciplinar con el léxico que aparece en cada sección, lo que generará un modo de construir un panorama del universo lingüístico y discursivo a través del cual los sujetos en formación acceden al conocimiento disciplinar (Parodi, 2007).

Este trabajo se enmarca en las líneas de investigación de Cuba Vega (2007) y López Morales (2011), quienes abordan el concepto de riqueza léxica a través del grado de madurez y dominio léxico general y nocional de textos orales o escritos. Estos textos se actualizan según el tema del discurso y las circunstancias comunicativas y su análisis se realiza mediante índices estadísticos aplicados al análisis de corpus textuales (Read, 2010). En este sentido, esta investigación robustece tres líneas de estudio: 1) la línea de análisis sobre los estudios empíricos de riqueza léxica en corpus (Ávila, 1986; Ávila Martín, 2017; Guzzi y Alonso Ramos, 2023; Hernández-Pérez *et al.*, 2023; Leal *et al.*, 2005; Leiva Pizarro, 2019; Madrigal y Vargas, 2016; Medina Urrea, 2014; Ávila Muñoz, 2014; Pérez Durán y Rojas Corona, 2025; Silva-Pereyra *et al.*, 2014; Torres González, 2003; Torruella y Capsada, 2013; Valencia *et al.*, 1992; Venegas *et al.*, 2013); 2) la línea de análisis sobre los estudios del léxico desde una perspectiva académica-disciplinar de corte cuantitativo-descriptivo (Bazerman, 2012; Cadario, 2001; Cisneros Estupiñán y Muñoz Dagua, 2014; Cisneros *et al.*, 2014; Cisneros Estupiñán *et al.*, 2020; Hyland, 2009; Parodi, 2008; Romero, 2024) y 3) la línea de investigación sobre análisis de corpus en trabajos académicos de contextos universitarios (Rojas, 2024; Velasco Luzuriaga y Álvarez Álvarez, 2019; Venegas, 2007).

## **MARCO CONCEPTUAL**

La tesis de grado es un trabajo de investigación que, en el ámbito universitario, posee un reconocimiento académico cuya relevancia se ve reflejada en lo social y en lo cultural. La intención de elaborar una tesis es la actualización del contenido de un tema disciplinar a través de la reflexión y desarrollo del documento académico. La diferencia entre la tesis y cualquier otro trabajo escolar universitario (tareas, ensayos, elaboración de mapas, etc.) es que la tesis de grado 1) presenta varias secciones: la introducción, donde se presenta y sostiene la justificación; planteamiento del problema; objetivos y preguntas de investigación; los antecedentes, el marco teórico; la metodología; el análisis y la discusión de los de los datos; la conclusión y la bibliografía relacionada con el contenido de la misma (Blasco Gil *et al.*, 2024). Además, en la tesis, también, se actualiza la metodología de estudio. En México, este documento se solicita como un requisito indispensable para que el estudiante pueda obtener su grado de estudio y pueda ser considerado como profesionalista en el ámbito laboral.

La elaboración de la tesis presenta un beneficio de naturaleza epistémica, ya que promueve el pensamiento crítico, genera una reflexión sobre el objeto de estudio y propicia la generación de nuevas teorías y métodos de investigación. Además,

puede considerarse como cualquier otro trabajo de investigación del cual se obtendrá información valiosa sobre el dominio que poseen estos futuros profesionistas sobre el lenguaje académico y disciplinar (Ibarra-López, 2017). En este sentido, la tesis de grado puede conformarse como corpus real donde hay posibilidad de realizar estudios de riqueza léxica, maduración sintáctica, maduración cognitiva, desarrollo y habilidades de escritura, entre otros aspectos.

La escritura que permea el trabajo de grado es de tipo académico y disciplinar. La escritura académica hace referencia al uso de unidades léxicas que se aplican en la generación de un trabajo académico. Su característica esencial es que este léxico fortalece a través de su uso las convenciones gramaticales, ortográficas y discursivas que forman parte global del texto, mientras que el léxico disciplinar está relacionado al fortalecimiento temático o disciplinar del área de estudio. Tanto el léxico académico como disciplinar son constituyentes esenciales de la escritura que está presente en cualquier trabajo de investigación (Pérez y Rojas, 2023; Torres Perdigón, 2018).

Sobre el léxico que conforma el trabajo de tesis de grado, este se puede estudiar como un conjunto de entradas léxicas que sirven de base para conocer el comportamiento léxico de los textos orales o escritos que el estudiantado ha aprendido en el ámbito educativo. Uno de tantos análisis que se puede llevar a cabo es el trabajo sobre riqueza léxica. La riqueza léxica es una destreza compleja que involucra todos aquellos aspectos matemáticos que se pueden aplicar a los corpus lingüísticos, razón por la cual se emplea la lexicoestadística (Cuba Vega, 2007; Hoover, 2003; Laufer y Nation, 1995; López Morales, 2011; Vine Jara *et al.*, 2021).

Esta riqueza léxica se puede encontrar en los textos escritos y en la producción oral, lo que permite comprender mejor la competencia léxica del sujeto. En términos específicos, la riqueza léxica tiene que ver con el número de palabras diferentes que se usan en el texto, ya sean palabras simples o compuestas. De la identificación de estos términos se obtienen datos reales sobre los textos. La riqueza léxica incluye unidades menores como la densidad, diversidad y frecuencia léxica. De acuerdo con el Centro Virtual Cervantes, Diccionario de Términos clave de ELE (2024), la diversidad léxica es el número de palabras diferentes que posee un texto en lugar de las palabras repetitivas (Read, 2010), mientras que la densidad léxica que aparece en un texto se obtiene mediante la relación existente entre su extensión (número de palabras totales) y el número de palabras diferentes que contiene (Daller *et al.*, 2003; Johansson, 2008) y la frecuencia léxica resulta del contraste entre el número de palabras utilizadas en el texto frente a otros corpus (Casado-Mancebo, 2023), es decir, son aspectos de la complejidad léxica en una muestra.

Para el análisis de la información sobre riqueza léxica, se utilizan la identificación de los *tokens* y los *types* como medidas de estudio. Su contraparte será el vocablo y la palabra en el ámbito hispánico. El vocablo será considerado *type* porque opera sobre la unidad mínima de léxico y son denominados unidades de escritura, y *token* será considerado palabra porque es una unidad textual y son las unidades mínimas elementales del texto que se distinguen por la tipografía y la escritura

(Carcedo, 1999; Müller, 1973; Torruella y Capsada, 2013). Además, "The term type denotes the different types of words in the text, while token refers to the total quantity of words used" (Djiwandono 2016, p. 209).

La organización del corpus para su estudio está vinculada a criterios lexicoestadísticos que se deben aplicar. Así, por ejemplo, en el caso de los verbos, se usa el lema en infinitivo o, en el caso de los nominales, se utiliza el masculino neutro y singular, etc. Este tipo de decisiones está relacionado con la estructura metodológica para el análisis del corpus, tal como mencionó Nation (2014) quien, además, enfatiza que las medidas cuantitativas son un referente fundamental en la comprensión global del texto de cualquier tipo. Para ello, hay que tener presente 1) el recuento de las palabras que componen el texto que está representado por N y es el encargado de generar una medida de la extensión del texto y 2) el recuento de los vocablos empleados en el texto, representado por V que, al igual que el anterior recuento, mide la extensión del vocabulario.

En el caso de la tesis de grado, un criterio fundamental que se conoce a través de los estudios de riqueza léxica es saber qué ítem léxico aparecerá como parte de su estancia en la universidad. A través de los resultados que se han obtenido de la densidad, la diversidad y la frecuencia léxica, autores como Laufer y Nation (2001) insisten en la importancia de visibilizar los ítems léxicos frecuentes antes de tratar de incorporar nuevas palabras dentro de la escritura y, en la tesis, como es un producto académico, se percibe el amplio dominio lexical que pueden tener los estudiantes cuando han terminado de escribir la tesis.

7/25

## **METODOLOGÍA**

La metodología de este trabajo es de corte cuantitativo descriptivo no experimental aplicado al análisis de la riqueza léxica. La muestra se integró con la sección de la introducción y la conclusión del producto tesis de grado de tres licenciaturas de la Facultad de Ciencias Sociales y Humanidades de la Universidad Autónoma de San Luis Potosí, México. En este trabajo, se obtiene el estudio por conteo de los *types* y *tokens* por cada licenciatura y por cada variable establecida y se lleva a cabo el análisis de la densidad, diversidad y frecuencia léxica de los textos.

### **Contextualización del corpus**

Para el ámbito universitario en México, las áreas de Ciencias Sociales y Humanidades, se ubican en tercer lugar a nivel nacional como opción académica de estudio, solo detrás de las áreas de Salud y Económicas-Administrativas (INEGI, 2024) por el

número de estudiantes<sup>3</sup>. De acuerdo con datos de Secretaría Escolar<sup>4</sup> de la Facultad de Ciencias Sociales y Humanidades de la Universidad Autónoma de San Luis Potosí, los alumnos para titularse tienen distintas opciones como: Alto rendimiento (Ar), Artículo (A.), Catálogo (C.), Curso (Cu.), Estudios de postgrado (Ep.), Informe de actividades académicas (IdAA.), Informe técnico (IT.), Monografía (M.), Otro producto de investigación (O.), Producto audiovisual (Pa.), Reporte de proyecto profesional (Rp), Tesis (T.) y Tesis Etnográfica (Te.). Se presentan en la siguiente tabla los resultados totales de alumnos titulados en 2024. En la Tabla 1 se observa el total de cada opción a titularse por licenciatura.

**Tabla 1.** Total de alumnos titulados por las distintas modalidades en 2024

Lic.	Ar.	A.	C.	Cu.	Ep.	IdAA	IT.	M.	O.	Pa.	Rp.	T.	Te.
An.	21	1	0	1	0	0	1	3	0	1	12	157	1
H.	11	1	0	0	0	1	0	0	27	0	14	149	0
G.	13	0	10	0	1	0	0	0	0	0	18	164	0
T.	45	2	10	1	1	1	1	3	27	1	44	470	1

Sobre los resultados generales, el número total de titulados asciende a 607 estudiantes distribuidos de la siguiente manera: 198 pertenecen a estudiantes de Antropología, 203 a Geografía y 206 a Historia. Sobre las opciones de titulación, la tesis de grado sigue siendo la opción número uno para los estudiantes de esta área con 470. Si se observa, para este trabajo, se han tomado 125 tesis que representa el 26.6 % de la totalidad. Sobre el periodo que se ha elegido para el análisis (2007-2014), solo había cuatro opciones para titularse: alto desempeño académico, monografía, reporte profesional y tesis de grado. En la Tabla 2 se muestran los totales que en ese periodo se encontraban vigentes. Para su análisis se ha utilizado: An (Antropología), H (Historia), G (Geografía), Ar (Alto rendimiento), M (Monografía), Rp (Reporte profesional) y T (tesis).

8/25

<sup>3</sup> Sobre la población de estudiantes universitarios en México, se tiene que esta asciende a 4,051,691 (1,903,882 hombres y 2,147,809 mujeres) de 33,062,447 estudiantes totales que asisten a la escuela en México (16,471,800 son hombres y 16,590,647 son mujeres). En lo que respecta al número de estudiantes universitarios en San Luis Potosí, el número es de 39,464 (19,099 hombres y 20,239 mujeres) de 451,028 (226,789 hombres y 224,239 mujeres) que representa la totalidad en este estado. De los 39,464 estudiantes, 32,594 estudian a la Universidad Autónoma de San Luis Potosí. Esto representa el 80% de la población universitaria del estado. El número de estudiantes que forman parte de estas áreas de estudio representa el 0.15%. Todos estos datos han sido recuperados de la página de INEGI (2004) [https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Educacion\\_Educacion\\_06\\_b6725858-a0f4-47fc-97e7-12018dd1b8dd](https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Educacion_Educacion_06_b6725858-a0f4-47fc-97e7-12018dd1b8dd) (Consultas: 2024).

<sup>4</sup> Agradezco la información proporcionada sobre los datos de titulación de la FCSyH a la responsable Ingeniera Judith Elisa Corpus Saldaña.

**Tabla 2.** Total de alumnos titulados clasificados por licenciatura y por modalidades de titulación durante el periodo de 2007-2014

Lic.	Ar.	M	Rp.	T
An.	6	1	7	60
H.	2	0	7	75
G.	0		1	71
Total	8	1	15	206

En el periodo que se analiza, el número total de titulados fue de 230 distribuidos en las cuatro modalidades que se observaron de la tabla anterior. De estas modalidades, como se aprecia, número de estudiantes que escogieron la opción de la tesis de grado fue de 206 y, para el análisis que se hace, se han utilizado 125 tesis (60 % de la muestra). Esto refleja que los datos que se han obtenido pueden ser la base para otros trabajos similares. En la Tabla 3 aparece la distribución de la información:

**Tabla 3.** Tesis recolectadas en la Facultad de Ciencias Sociales y Humanidades distribuidas por licenciaturas

LICENCIATURA	NO. DE TESIS
Antropología	37
Geografía	45
Historia	43
Total	125

9/25

Para el análisis de las tesis se ha dividido la muestra en dos cohortes: 2007-2014 y 2015-2022. Por cuestiones ajenas al proyecto<sup>5</sup>, no se ha podido trabajar con la totalidad del corpus por lo que se dividió la temporalidad de la población en cada 7 años. Se tomaron como base de la cohorte los primeros trabajos de tesis culminados y aprobados. La Facultad de Ciencias Sociales y Humanidades es una facultad joven dentro de la universidad, tiene 20 años de su fundación y alberga seis licenciaturas que fueron apareciendo poco a poco.<sup>6</sup> Las tesis se encuentran en el repositorio del Centro de Información en Humanidades, Bibliotecología y Psicología (CIHBYP) de la universidad.

### **Anonimización de la información**

Para utilizar la información que sirve de base para este artículo, se solicitó por escrito el permiso a la biblioteca del Campus Oriente de la Universidad Autónoma de San Luis Potosí. En este documento, se indicó el objetivo de la investigación y se evidenció

---

<sup>5</sup> Los factores ajenos fueron: falta de personal, falta de computadoras para digitalizar las secciones que servirían de base para el análisis, falta de tiempo para procesar y analizar los corpus, etc.

<sup>6</sup> Para más información, presenta el enlace de la facultad: <https://sociales.uaslp.mx/#gsc.tab=0>

que los datos personales que rigen los proyectos de tesis se anonimizaron y quedaron resguardados de acuerdo con el reglamento sobre el uso de datos.<sup>7</sup>

### *Variables empleadas*

Las variables que se utilizan para el estudio son: género, licenciatura y año de producción de las tesis. La variable lingüística, como se le denomina, está relacionada con el uso de la lengua condicionada por factores sociales, culturales, geográficos, etc. Esta variación forma parte de la sociolingüística cuantitativa (Labov, 1966). En la variable género, se encuentran las diferencias entre hombres y mujeres; en la variable licenciatura, se ubican las diferencias léxicas de cada área de estudio, y en la variable año de producción de la tesis, se hace referencia a la cantidad de textos que aparecieron en un determinado tiempo y que forman parte de la muestra.

### **Criterios para el análisis cuantitativo**

En esta parte, se describen las fórmulas que se han utilizado para el análisis de la muestra.

#### *Densidad léxica (DL)*

10/25 La densidad léxica (DL) calcula la relación existente entre el número de palabras distintas y el número de palabras totales. El Centro Virtual Cervantes en su Diccionario de términos clave de ELE (revisado el 25 de abril de 2024) define la densidad léxica de un texto como la relación que hay entre su extensión (número de palabras total) y el número de palabras distintas que contiene.

Se trata de una medida estadística que da cuenta de la riqueza léxica de un texto. En otros términos, un texto con una alta densidad léxica posee muchas palabras distintas que se repiten pocas veces, mientras que un texto con una densidad léxica baja contiene pocos términos que se repiten muchas veces. Entonces, a través de la densidad, se conoce la riqueza léxica del texto. Tradicionalmente, a mayor densidad, mayor número de palabras distintas. La fórmula que se utiliza es  $DL = PN/N$ , donde (DL)=Densidad léxica, (PN)=Total de palabras nocionales (palabras que presentan un significado), (N)=Extensión total del texto. Además del análisis de la densidad, se aplica el Type-Token Ratio (TTR) ajustado y denso.

El TTR ajustado calcula la relación entre las palabras únicas de contenido semántico y la extensión total del texto. Ofrece una mirada que relaciona la densidad léxica con la diversidad de cada texto (Vine Jara *et al.*, 2021). Se representa así:  $TTR \text{ ajustado} = \text{type semántico}/\text{token}$ . El TTR denso calcula la relación entre las palabras únicas de contenido semántico y el total de palabras de contenido semántico de cada texto. Con este se observa la densidad léxica del texto al relacionar solo el contenido

---

<sup>7</sup> Este se puede consultar en <https://www.psicologia.uaslp.mx/Paginas/Servicios/3737#gsc.tab=0>.

semántico (Vine Jara *et al.*, 2021). Se representa TTR denso=  $\text{type semántico}/\text{token semántico}$ .

### Diversidad léxica

La diversidad léxica hace referencia al número de palabras diferentes utilizadas en un texto, un mayor rango indica una diversidad mayor (Jarvis, 2013; Johansson, 2008; López Morales, 2011). La medida que se utiliza es *Type Token Ratio* (TTR). El *type* corresponde a las palabras distintas y el *token* al número total de palabras. Para calcular este índice, se utiliza la división del número de *types/token* (total de palabras de un texto). Cabe mencionar que no se considera para el análisis de este trabajo el *hápax legómenon* (palabras únicas dentro del texto). Durán *et al.* (2004) menciona que la dificultad de esta fórmula TTR es la extensión del texto. Mientras más extenso sea el texto, la probabilidad de que aparezcan nuevas o diferentes palabras puede ser más escasa. Esto tiene que ver con el uso de palabras gramaticales y la temática del escrito (Vine Jara *et al.*, 2021).

### Frecuencia léxica

La frecuencia léxica, también llamada frecuencia de uso de palabras, es un indicador de corte estadístico que permite observar el comportamiento del léxico en el corpus. Para el análisis, se consideran dos medidas: media y mediana que pertenecen a la estadística descriptiva (Rodríguez *et al.*, 2020). En el apartado siguiente se describe cada una y se presentan sus fórmulas.

11/25

<p>Media: calcula el valor medio que se encuentra al sumar todos los totales en el conjunto de datos y se divide entre el número de valores del corpus. La fórmula se presenta a continuación:</p> $\text{Media } (X) = \bar{x} = \frac{\sum X_i N_i}{N}$ <p>Siendo <math>(X_1, X_2, \dots, X_N)</math> el conjunto de observaciones.</p>	<p>Mediana: esta medida representa el valor de la variable cuando se ordena de menor a mayor.</p> <p>Se representa así:</p> $Me = Li-1 + \frac{N/2 - Fi-1}{fi} \cdot a$ <p><math>Li-1</math> = Límite inferior del intervalo de la mediana.</p> <p><math>a</math> = Amplitud del intervalo mediana</p> <p><math>Fi-1</math> = Frecuencia acumulada anterior al intervalo mediana</p> <p><math>N</math> = Total de datos.</p>
---	--

Además, se utilizan las medidas de dispersión que indican lo disperso o lo agrupado de un corpus a partir de su medida aritmética (Rodríguez *et al.*, 2020). Esta está integrada por rango, varianza y desviación estándar. Se exponen a continuación los siguientes conceptos:

<p>Rango: se utiliza para obtener la dispersión entre el mínimo y el máximo del corpus. Se presenta a continuación la fórmula:</p> $R = \text{Max } x_i - \text{Min } x_i$	<p>Varianza: es la medida de variabilidad que da cuenta del grado de homogeneidad de un corpus. Su fórmula es:</p> $\text{varianza} = \frac{\sum (x_i - \bar{x})^2}{n - 1}$
<p>Desviación estándar: mide la dispersión de una distribución de datos. Para obtener el resultado, entre más dispersa tenga la distribución de datos, más grande será su desviación estándar.</p> $\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$ <p>Donde:</p> <p><math>\sigma</math> = desviación típica</p> <p><math>n</math> = número de datos u observaciones</p> <p><math>x_i</math> = <math>i</math>ésimo dato</p> <p><math>\bar{x}</math> = media o promedio aritmética</p>	

12/25

### Software utilizado

El análisis se realizó mediante el programa AntConc. Este software es de acceso libre y se encuentra en la siguiente liga: <https://www.laurenceanthony.net/software/antconc/>. Con este programa, se extrajeron los *tokens* y *types*. El resultado de la información se ubica en tablas comparativas para evidenciar el comportamiento léxico.

### Procedimiento

Para obtener el corpus, en primer lugar, se solicitó permiso a la biblioteca de la zona oriente de la Universidad Autónoma de San Luis Potosí a principios de mayo de 2023. Una vez autorizado el uso de la información, se empezó a trabajar y se encontró con que muchas tesis no estaban en formato PDF, por lo que estos escritos tuvieron que

convertirse en ese formato. Después de esto, se generó una lista de las tesis para verificar su disponibilidad, su año y su licenciatura.

En segundo lugar, se elaboró el primer conteo de las tesis que estaban a disposición y se ubicaron las dos secciones que sirven de base para el estudio. Inmediatamente después, se llevó a cabo el proceso editorial de las dos secciones: introducción y conclusión. De las dos secciones, se eliminaron citas, nombres y fechas de los contenidos que pudieran dar pista de los autores utilizados en el texto, de las imágenes y de los parafraseos que aparecían entre comillas simples y dobles, se seccionaron los párrafos en palabras y se eliminaron las palabras funcionales. Todo este procedimiento se hizo de forma manual porque no existe programa que realice de forma detallada este tipo de trabajo. En ese sentido, el corpus que se estudia está conformado exclusivamente por palabras de contenido semántico o nocional (sustantivos, adjetivos y verbos). El proceso de depuración, corrección, agrupación y análisis fue de un año.

Una vez terminado lo anterior, se construyó la base de datos para su análisis y descripción. La información se presenta en tablas.

## RESULTADOS

Sobre el análisis general, se presenta la información recabada en la Tabla 4. Se utiliza L para referirse a Licenciatura, T para Tesis, An. para Antropología, G para Geografía y H para Historia.

13/25

**Tabla 4.** Datos obtenidos en el análisis general sobre los *tokens* y *types*

L	T	Datos generales		Introducción	Conclusión
		Palabras totales ( <i>tokens</i> )	Palabras únicas ( <i>types</i> )	Palabras totales	Palabras totales
An.	37	165,715	17,098	100,165	65,550
G	45	88,275	19,899	48,381	39,894
H	43	143,150	23,457	91,932	51,218
Totales	125	397,140	60,454	240,478	156,662

De la información de *tokens* y *types*, el número total de *tokens* es de 397,140 y el de *types* es de 60,454 (17,1 %). La diferencia asciende a 336,686. Si se comparan los *types* (60,454) con los resultados por licenciatura, Historia es la que más *types* generó con 23,457, después Geografía (19,889) y, por último, Antropología (17,098). Sobre los *tokens*, Antropología acumuló 165,715 que representa 37,5 % del total del corpus y Geografía obtuvo 88,275. Una diferencia de 4,7 % sobre el análisis de las palabras totales (*tokens*). En lo que respecta a las palabras únicas (*types*), Historia ocupó el primer lugar con 23,457 (26,4%), le sigue Geografía con 19,899 y, por último,

Antropología con 17,089. Si se comparan los datos, la diferencia entre Historia y Geografía es de 3,558 y con Antropología es de 6,368.

Sobre el análisis de la introducción, Antropología es la licenciatura que más palabras totales tiene con 100,165 (39 %), mientras que Geografía ocupó la última posición con 48,381 palabras. Si se comparan ambas cifras, la diferencia es del 51 % más para Antropología y con un menor número de tesis que las otras dos licenciaturas. Si se comparan los corpus de Antropología e Historia, la diferencia es 8,233, lo que representa el 9 % de la totalidad de diferencia. En lo que respecta a la conclusión, los datos son similares a la introducción: Antropología en primer lugar y Geografía en último lugar. La diferencia es de 25,656 entre una y otra. Esta diferencia es de 50 %. Si se comparan Antropología e Historia, la diferencia entre ellas es de 14,332.

En el análisis estadístico de las 125 tesis, la *media* se obtuvo con 28 tesis. La media se aplicó porque los datos que se presentan forman parte de una distribución normal con una cantidad baja de valores atípicos, esto se comprueba al dividir el total general entre el total de tesis (397,140 palabras/125 tesis). El resultado general es de 3,177.12, mientras que, en la introducción, de 1,923.82 y en la conclusión, de 1,253.29. Al comparar los resultados, en la introducción se aprecia una mayor cantidad de palabras que en la conclusión en promedio. En el análisis de la mediana, se menciona que esta es de 3,583 en general, en la introducción es de 2,100 y en la conclusión es de 1,504, por lo tanto, su varianza, como medida de dispersión, se aleja de los valores individuales que aparecen alrededor de la media.

14/25

Además del análisis de la media y mediana, se aplicó el análisis sobre la desviación estándar. De manera concreta, este dato se utiliza para saber cuánto se alejan los datos individuales con respecto a la media. Para ello, si la desviación es baja, los resultados estarán más cercanos a la media y viceversa. En este caso, los resultados generales se ubican más cercanos a la media (2792.047985) que se ve reflejada en la introducción y conclusión. Esto indica que con 28.04 (68,040 palabras) se representa el promedio de las 125 tesis.

Sobre el análisis del rango, el rango máximo en la introducción es de 11,274 y el rango mínimo es de 1,162 palabras. En la introducción, el rango máximo es de 6,940 y de la conclusión es de 5,854 y, en relación con el rango mínimo en la introducción, este es de 563 y en la conclusión de 436.

Se presenta el análisis de las variables sexo y año de cada licenciatura en la Tabla 5. La nomenclatura de la tabla es la siguiente: L (licenciaturas), T (tesis), TGm (*Tokens* generales de la muestra), Tg (*Tokens* por género), I (introducción), Ig (Introducción por género), C (conclusión), Cg (conclusión por género), M (Mujer) y H (Hombre).

**Tabla 5.** Resultados obtenidos de la distribución de *tokens* considerando las variables de sexo y año

N	T	TGm	Tg		I	Ig		C	Cg	
			M	H		M	H		M	H
A	37	165,715	12,923	7,756	100,165	11,455	7,377	65,550	10,826	6,848
G	45	88,275	7,841	7,011	48,381	8,091	6,528	39,894	7,834	6,353
H	43	143,150	9,785	11,481	91,932	9,699	10,367	51,218	9,227	9,670
T	125	397,140	30,548	26,248	240,478	29,245	24,272	156,662	27,887	22,871

En relación con el análisis de los *tokens* por género, las mujeres produjeron tanto en la muestra general como en cada una de las secciones el mayor número de *tokens*. En el análisis general, la cifra fue de 4,300, en la introducción fue de 4,973 y en la conclusión fue de 5,016 *tokens*. En esta variable, el número de participantes “mujer” refleja el resultado en las secciones que se estudian. Por lo que respecta a los resultados de los *types*, se presenta la Tabla 6 de distribución:

**Tabla 6.** Resultados obtenidos de la distribución de *types* considerando las variables de sexo y año

N	T	Palabras totales ( <i>tokens</i> )	<i>Types</i> <i>generales</i>	Introducción		Conclusión	
				M	H	M	H
An	37	165,715	17,098	2,649	1,911	2,551	1,833
G	45	88,275	19,899	2,044	1,757	1,995	1,741
H	43	143,150	23,457	2,368	2,492	2,323	2,408
Totales	125	397,140	60,454	7,061	6,160	6,869	5,982

El número total de *types* es de 60,454. La suma de la introducción y conclusión de Antropología asciende a 8,944, de Geografía a 7,531 y de Historia a 9,591. Esta última licenciatura es la que obtuvo más *types* de la muestra. Sobre los totales, las mujeres son las que más producción presentan en ambas secciones, esto mismo se ha visto en el corpus de los *tokens*. De igual forma, la totalidad de las participantes influyó en el resultado de los *types*.

En relación con el análisis por género, se presenta la Tabla 7 de distribución de datos: T (tesis), M (mujeres), H (hombres) y Mix (mixto).

**Tabla 7.** Resultados obtenidos de la distribución de tesis por la variable de sexo

	T	M	H	Mix
Antropología	37	26	11	
Geografía	45	26	18	1
Historia	43	21	22	
	125	73	51	1

Como se observa, el número de tesis es de 125. La licenciatura en Geografía presenta el mayor número de tesis con 45 y, en último lugar, Antropología con 37. En relación con género, las mujeres son las que presentan más tesis concluidas con 73. Antropología y Geografía son las licenciaturas que acumularon más escritos con 26. Sobre las tesis escritas por hombres, Historia es la que mayor número de tesis presenta con 22 y Antropología se ubicó en la parte final con 11.

Sobre el análisis general de las tesis por año, el mayor número de tesis se ubicó en 2012 con 22, 2011 con 21 y 2010 con 20. Estos tres años representan el 50.4 % de la totalidad de las tesis de 2007 al 2014. En relación con el género, las mujeres fueron las que más tesis produjeron. Los hombres en 2007, 2010 y 2014 fueron los años más productivos en relación con la tesis de grado. En la Tabla 8, se especifica mejor la distribución.

**Tabla 8.** Resultados obtenidos de la distribución de tesis por la variable de año

	M	H	Mix	T
2007	1	7		8
2008	12	5		17
2009	8	4		12
2010	8	12		20
2011	17	4		21
2012	12	9	1	22
2013	12	6		18
2014	1	2		7
	73	51	1	125

Sobre la distribución por licenciatura, las mujeres de la licenciatura en Antropología en 2010 y 2011 generaron el mayor número de tesis: 6 y 5 respectivamente, en Geografía, también las mujeres en 2011 produjeron 11 tesis y, en Historia, en 2008, alcanzaron 5 tesis. Por lo que respecta a los hombres, en la licenciatura en Historia, alcanzaron la mayor cantidad de producción en 2010 y en 2011 con 6 tesis por año. Para visualizar mejor la distribución se utiliza M (mujer), H (hombre), An. (Antropología), G (Geografía) y H (Historia). Obsérvese la Tabla 9:

**Tabla 9.** Resultados obtenidos de la distribución de tesis por la variable de licenciatura

An	M	H	G	M	H	Mix	H	M	H
2007	0	2		1	2			0	3
2008	4	2		3	2			5	1
2009	3	0		3	2			2	2
2010	5	3		0	3			3	6
2011	6	1		8	2			3	1
2012	3	2		4	1	1		5	6
2013	4	1		7	4			1	1
2014	1	0		0	2			2	2
TOTAL	26	11		26	18	1		21	22

17/25

Sobre la distribución de las palabras únicas de contenido semántico, se presenta la Tabla 10.

**Tabla 10.** Resultados obtenidos sobre la distribución de las palabras únicas de contenido semántico

Datos generales				
N	T	Palabras totales (tokens)	Palabras únicas (types)	Palabras únicas de contenido semántico (Types contenido semántico, PuCS)
An	37	165,715	17,098	6.49
G	45	88,275	19,899	4.43
H	43	143,150	23,457	6.10
Totales	125	397,140	60,454	6.56

Como se aprecia, la licenciatura que presentó el número mayor de palabras de contenido semántico es Antropología con 6.49 y le sigue Historia con 6.1 *types*. Los *types* de contenido semántico son fundamentales para obtener el TTR ajustado y el TTR denso. Para calcular el TTR ajustado se utiliza la siguiente fórmula: palabras únicas de contenido semántico / palabras totales y, para el TTR denso, se divide palabras semánticas / total de las palabras semánticas. Obsérvese los resultados en la Tabla 11:

**Tabla 11.** Resultados obtenidos de TTR ajustado y TTR denso

		Datos generales				Densidad léxica		
N	T	Palabras totales ( <i>tokens</i> )	Palabras distintas ( <i>tokens</i> )	Palabras únicas( <i>types</i> )	PuCS ( <i>types</i> )	TTR ajustado	TTR denso.	DL (0-1)
An	37	165,715	148,617	17,098	8,042	0.018	0.47	0.89
G	45	88,275	68,376	19,899	7,191	0.016	0.36	0.77
H	43	143,150	119,693	23,457	7,992	0.018	0.34	0.83

18/25

Para el análisis de densidad léxica, hay que considerar que, a mayor densidad léxica, mayor número de palabras distintas de un texto. Si el resultado está más cercano a 1, esto indicará mayor densidad léxica y viceversa. Para obtener el cálculo, se divide el número de palabras distintas entre el número de palabras totales. Todas las licenciaturas se ubican en el rango de 0,7 y 0,9 de densidad léxica. La diferencia es que Antropología e Historia están más cercanas a 1 que Geografía.

Sobre el análisis de los resultados de las medidas TTR denso y TTR ajustado, se ubican ambas medidas en el rango 0-1. Entre más se aproxime a 1, mayor será el número de palabras de contenido semántico que posea el texto y viceversa. En el análisis del TTR ajustado, Antropología e Historia tuvieron el mismo resultado 0.018, mientras que Geografía apareció en tercer lugar con 0.016. Sobre el TTR denso, Antropología ocupó la primera posición (0.47), mientras que Geografía e Historia presentan un TTR denso parecido con 0,36 y 0.34. Si se comparan las licenciaturas, Antropología estaría más cerca al número de palabras de contenido semántico que las otras dos licenciaturas.

El análisis de la diversidad léxica hace referencia a la diversidad y variedad de palabras de contenido. En palabras de Torruella y Capsada (2013), la TTR no es un valor constante, sino que va disminuyendo a medida que va aumentando la cantidad de *tokens* del texto que cogemos para su medida. Esta disminución de la TTR se explica a partir de la relación existente entre el número de *types* que van apareciendo con respecto al número de *tokens* del texto. En la Tabla 12 se presenta la distribución:

**Tabla 12.** Resultados del análisis de la diversidad léxica en las tesis de licenciatura

N	T	Datos generales		
		Palabras totales ( <i>tokens</i> )	Palabras únicas ( <i>types</i> )	Diversidad léxica <i>Types/tokens</i>
An	37	165,715	17,098	0.10
G	45	88,275	19,899	0.22
H	43	143,150	23,457	0.16

En el análisis anterior, Geografía es la licenciatura que presenta mayor diversidad que las otras dos licenciaturas. Esto tiene que ver con el tamaño de *tokens* de las licenciaturas que condiciona de manera directa el resultado.

## DISCUSIÓN

La hipótesis planteada al inicio de este trabajo indicaba una respuesta distinta a la que se creía en un principio. El análisis de los datos permitió conocer que la licenciatura en Antropología presentó mayor riqueza léxica en los datos relacionados con las variables (género, licenciatura y año de la tesis). En relación con los *types*, la licenciatura en Historia ocupó el primer lugar en uso. Esto demuestra un alto grado de heterogeneidad entre los *types* y *tokens* que aparecen en todas las licenciaturas tal como también apareció en el trabajo de Parodi (2007).

La interpretación del comparativo lleva a establecer que la extensión de la introducción y la conclusión es un factor específico en la generación del proyecto de tesis y que los estudiantes universitarios tienden a escribir mucho para sostener argumentativamente su proyecto de investigación. Sin considerar el número de tesis y su extensión acumulada como corpus y como factor específico de análisis, es un hecho que Antropología es una licenciatura que propicia mayor uso de vocabulario en este periodo que se estudia. De este modo, resulta claro mencionar que las secciones analizadas en este trabajo son el resultado de una amplia cantidad de estrategias educativas que fueron puestas a disposición de los estudiantes y que los resultados son parcialidades que servirán como primer acercamiento sobre *types-tokens* que aparecen en estos discursos académicos.

Sobre la relación tesis y *types-tokens*, los datos que se reflejan en las tablas 4, 5 y 6 de este trabajo revelan que sí hay una tendencia en el incremento y variación léxica de los datos entre licenciaturas y la variable propuesta (hombre-mujer). Esta tendencia, en términos de análisis, establece que las mujeres son las que más tesis generaron con 73 en comparación con los hombres (51). Sobre el periodo de generación de la tesis, hay cuatro años 2008, 2011, 2012 y 2013 en los que se evidencia un incremento en el número de tesis que concluyeron el trabajo. De estos años, el 2011 fue en el que más investigaciones se generaron con 17 tesis escritas por mujeres. Sobre

los resultados de la introducción, como se ha evidenciado, las estudiantes son muy productivas en cuanto al léxico utilizado. Las estudiantes de Antropología en ambas secciones son las que mayor cantidad de *tokens* han obtenido. Esto coloca a esta licenciatura como la más productiva de esta comparativa y de esta área de estudio. De manera específica, Historia obtuvo el mayor número de *types* con 23,457. Al hacer el análisis por introducción y conclusión de *types*, de nueva cuenta, Antropología obtuvo el mayor número: 2,649 en la introducción y 2,551 en la conclusión. Este es otro dato revelador del comportamiento léxico que se ha venido observado a lo largo del estudio y que, también, aparece en los trabajos de Cuba Vega (2007) en el que, según el grado escolar, aumenta la riqueza léxica y alcanza una proporción entre palabras de contenido nocional que permite generar y producir textos académicos. Esto se aprecia cuando se afirma que Antropología ha sido la licenciatura que más palabras nocionales generó con 6.49.

Los datos presentados en el párrafo anterior demuestran además que esta licenciatura ha sido productiva en cuanto a *types-tokens*, también los es en cuanto a la obtención de palabras de contenido semántico. Este número se encuentra en el rango de las respuestas que ha sido de 6.56. La obtención de las palabras de contenido semántico es un claro ejemplo de que los resultados generales son evidentemente similares en las tres licenciaturas. Sobre los resultados de densidad léxica, Antropología sigue demostrando que es la licenciatura que mayor frecuencia y densidad léxica presenta, esto se demuestra al obtener en TTR ajustado y denso los datos más elevados del corpus y por la diversidad léxica.

20/25

## CONCLUSIONES

La relación entre léxico y educación ha sido un fenómeno que está presente en cualquier tipo de estudio. El análisis realizado forma parte de los resultados del trabajo general. Se estudia la riqueza léxica como parte importante de la estructura y competencia léxica que debe mostrar un estudiante que ha dejado de serlo para convertirse en un profesional. Con el análisis léxico hecho, se ha obtenido el primer corpus sobre léxico universitario extraído directamente de un proyecto de tesis, algo que no se había hecho y que es necesario para comprender desde otra vertiente de estudio el desarrollo de la escritura académica en el contexto universitario.

A través del conocimiento de la frecuencia, densidad y diversidad léxica, se obtiene información valiosa que evidencia el comportamiento léxico de esta comunidad universitaria (Crossley *et al.*, 2017). En ese sentido, el haber hecho un análisis de riqueza léxica obliga a entender la necesidad y la importancia que presenta este trabajo de investigación por encima de otras opciones de titulación. Aunque hay que reconocer que la tesis de grado representa un reto de tipo académico que muchos estudiantes no logran superar porque 1) es un trabajo organizado conceptualmente, 2) es un trabajo que requiere de mucho tiempo para desarrollar y comprobar lo que se

espera obtener y 3) es un trabajo que requiere del dominio y aprendizaje de teorías y métodos que se aprendieron durante la vida universitaria.

En lo que concierne al objeto de estudio, la introducción y la conclusión son dos secciones particularmente interesantes porque en estas se puede apreciar la voz de quienes escriben, lo que permite conocer el dominio, la percepción y la maduración cognitiva que son fundamentales para afrontar la vida laboral. Los beneficios que esto arroja a mediano y largo plazo son diversos, entre ellos, hace consciente a los directores de tesis y a los programas educativos de la importancia que se tiene en mejorar los procesos de enseñanza y aprendizaje durante la carrera universitaria. Además, a través de este tipo de estudios, se puede observar la relación entre decodificación léxica, habilidad léxica y comprensión léxica para lectura y escritura, procesos educativos y cognitivos que no se han desarrollado a través del uso de la frecuencia léxica y que sería de gran ayuda para la planificación de materiales académicos, la elaboración de programas y guías de estudio, la elaboración de exámenes para ingreso y egreso al ámbito universitario, la elaboración de informes académicos, la generación de bases de datos para inteligencia artificial, etc.

Por último, los resultados que se han analizado en este artículo reflejan una posible realidad sobre la capacidad de escritura que presentan los estudiantes de Ciencias Sociales y Humanidades. Esta capacidad está condicionada por la cantidad de lectura que lleva a cabo el estudiante durante su estancia en la universidad y que se refleja en la tesis de grado. Esto se comprobará cuando se tenga de manera completa la base de datos de todas las áreas de estudio de esta universidad.

21/25

### **Financiación**

Este trabajo forma parte del proyecto titulado Vocabulario académico universitario de la Universidad Autónoma de San Luis Potosí, México y financiado por la misma universidad, con código UASLP-PREDO-2023-874. Este inició el 1 de enero de 2024 y finalizó el 31 de diciembre del mismo año. El responsable principal es el doctor Marco Antonio Pérez Durán, profesor-investigador de tiempo completo de la Facultad de Ciencias Sociales y Humanidades, UASLP.

### **Contribución de los autores**

Marco Antonio Pérez Durán: conceptualización, metodología, software, análisis formal, investigación, recursos, curaduría de datos, escritura – borrador original, escritura, revisión y edición, visualización, supervisión, administración del proyecto, adquisición de fondos.

### **Conflictos de interés**

El autor declara que no tiene ningún conflicto de interés en la escritura y publicación de este artículo.

## Implicaciones éticas

El autor declara no tener ningún tipo de implicación ética que se deba declarar en la escritura y publicación de este artículo.

## REFERENCIAS

- Ávila, R. (1986). Léxico infantil de México: Palabras, tipos, vocablos. En J. G. Moreno de Alba (Ed.), *Actas del Congreso del II Congreso Internacional sobre el español de América* (pp. 510-517). Universidad Nacional Autónoma de México
- Ávila Martín, M. C. (2017). Aportaciones de la lexicografía en línea del español. *Itinerarios. Revista de estudios lingüísticos, literarios, históricos y antropológicos*, (25), 101-113. <https://itinerarios.uw.edu.pl/resources/html/article/details?id=222732>
- Ávila Muñoz, A. (2014). Patrones sociolingüísticos de la riqueza léxica. Estudio basado en una propuesta original para el cálculo del índice de la densidad léxica virtual de los hablantes. *Lingüística Española Actual*, 36, 171-194.
- Bazerman, C. (2012). *Géneros textuales, tipificación y actividad*. Benemérita Universidad Autónoma de Puebla.
- Blasco Gil, Y., Pavón Romero A., Peset, F. y García Peralta T. (2024). La tesis que me parió. Manual para abordar, elaborar y terminar un trabajo académico. *Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales*, 48(187), 451-452.
- Cadario, E. (2001). La Lingüística del corpus como herramienta de investigación en escritura académica en campos disciplinares específicos al área de formación docente. <https://feeye.uncu.edu.ar/web/posjornadasinve/area3/Lengua%20-%20Didactica%20de%20la%20lengua%20-%20TICs/176%20-%20Cadario%20-%20UN%20Rio%20Cuarto.pdf>
- Carcedo, A. (1999). Índices léxicos-estadísticos y graduación del vocabulario en la enseñanza de ELE (aspectos culturales). En M. Franco Figueroa, C. Soler Cantos, F. J. de Cos Ruíz, M. Rivas Zancarrón y F. Ruiz Hernández, (eds.), *Nuevas perspectivas en la enseñanza del español como lengua extranjera* (tomo I, pp. 174-183). Universidad de Cádiz.
- Casado-Mancebo, M. (2023). Una aplicación para explorar la frecuencia léxica a partir de corpus de referencia. *CHIMERA: Revista de Corpus de Lenguas Romanes y Estudios Lingüísticos*, 10, 89-95. <https://revistas.uam.es/chimera/article/view/15525>
- Centro Virtual Cervantes. (s.f.). CVC. *Diccionario de términos clave de ELE*. Recuperado en diciembre de 2024, de [https://cvc.cervantes.es/ensenanza/biblioteca\\_ele/diccio\\_ele/diccionario/densidadlexica.htm](https://cvc.cervantes.es/ensenanza/biblioteca_ele/diccio_ele/diccionario/densidadlexica.htm)
- Cisneros Estupiñán, M. y Muñoz Dagua, C. (2014). Los mecanismos de la valoración en la construcción del discurso en el aula universitaria. *Revista Colombiana de Educación*, (66), 247-263.

- Cisneros, M., Olave, G. y Rojas, I. (2014). *Cómo escribir la investigación académica. Desde el proyecto hasta la defensa* (1ed.). Ediciones de la U.
- Cisneros Estupiñán, M., Olave Arias, G. y Serna Pinto, M. A. (2020). Un estudio lexicométrico a la escritura de los aspirantes a un postgrado en lingüística. *Lingüística y literatura*, (77), 33-53.
- Crossley, S., Liu, R. y McNamara, D. (2017). Predicting math performance using natural language processing tools [Predicción del rendimiento matemático mediante herramientas de procesamiento del lenguaje natural]. In 7th International Learning Analytics and Knowledge Conference (pp. 339-347). <https://doi.org/10.1145/3027385.3027399>
- Cuba Vega, L. (2007). Índices de riqueza léxica en escolares de primaria de ciudad de La Habana. En M. A. Domínguez Hernández (coord.), *La lengua en Cuba* (pp. 238-308). Universidad Santiago de Compostela / Servizo de Publicacións.
- Daller, H., Van Hout, R. y Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, 24(2), 127-222. <https://doi.org/10.1093/applin/24.2.197>
- Djiwandono, P. (2016). Lexical Richness in Academic Papers: a Comparison Between Students' and Lecturers' Essays. *Indonesian Journal of Applied Linguistics*, 5(2), 209-216. <https://ejournal.upi.edu/index.php/IJAL/article/view/1345>
- Durán, P., Malvern, D., Richards B. y Chipere, N. (2004). Developmental Trends in Lexical Diversity. *Applied Linguistics*, 25(2), 220-242. <https://doi.org/10.1093/applin/25.2.220>
- Guzzi, E. y Alonso Ramos, M. (2023). Sofisticación y diversidad como medidas de complejidad léxica para determinar el perfil colocacional de textos académicos en español. *Revista signos*, 56(112), 282-305.
- Hernández-Pérez, Y. A., Alaniz-Álvarez, S. A., Martínez-Troncoso, J. C., Martini, M. y Esparza-Herrera, A. D. (2023). Detección de las frecuencias del léxico de un texto científico con la aplicación web LEX-A para apoyar el aprendizaje de lenguas extranjeras. *Enseñanza y Comunicación de las Geociencias*, 2(1), 28-31.
- Hoover, D. L. (2003). Another perspective on vocabulary richness. *Computers and Humanities*, 37(2), 151-178.
- Hyland, K. (2009). *Academic Discourse. English In A Global Context*. Bloomsbury Publishing.
- Ibarra-López, P. (2017). La titulación por tesis en México: el problema de su conceptualización. *Dilemas contemporáneos: Educación, Política y Valores*, 31(3), 05-22.
- Instituto Nacional de Estadística y Geografía. (2024). *Tabulados Interactivos-Genéricos*. Recuperado en diciembre de 2024, de [https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Educacion\\_Educacion\\_06\\_b6725858-a0f4-47fc-97e7-12018dd1b8dd](https://www.inegi.org.mx/app/tabulados/interactivos/?pxq=Educacion_Educacion_06_b6725858-a0f4-47fc-97e7-12018dd1b8dd).
- Jarvis, S. (2013). Capturing diversity in lexical diversity. *Language Learning*, 63, 87-106.
- Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. *Working Papers*, 53, 61-79.

- Labov, W. (1966). *The Social Stratification of English in New York City*. Center for Applied Linguistics.
- Laufer, B. y Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, 16(3), 307-322. <https://doi.org/10.1093/applin/16.3.307>
- Laufer, B. y Nation, P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? *EUROSLA yearbook*, 1(1), 7-28.
- Leal, F., Matute, E. y Zarabozo, D. (2005). La transparencia del sistema ortográfico del español de México y su efecto en el aprendizaje de la escritura. *Estudios de lingüística aplicada*, 23(42), 127-145.
- Leiva Pizarro, A. V. (2019). *Riqueza léxica: diversidad y densidad léxica en las narraciones orales de los niños con desarrollo típico y trastorno específico del lenguaje* [Tesis de licenciatura, Pontificia Universidad Católica de Valparaíso]. Repositorio Institucional PUCV. [http://opac.pucv.cl/pucv\\_txt/txt-7500/UCC7956\\_01.pdf](http://opac.pucv.cl/pucv_txt/txt-7500/UCC7956_01.pdf)
- López Morales, H. (2011). Los índices de riqueza léxica y la enseñanza de lenguas. En J. de Santiago Guervós, H. Bongaerts, J. Sánchez y M. Seseña (coords.), *Del texto a la lengua: La aplicación de los textos a la enseñanza-aprendizaje del español L2-LE* (pp.15-28). Asociación para la enseñanza del español como lengua extranjera.
- Madrigal, M. y Vargas, E. (2016). Índice de riqueza léxica en redacciones escritas por estudiantes universitarios. *Káñina*, 40(1), 139-147. <http://dx.doi.org/10.15517/rk.v40i3.29260>
- Medina Urrea, A. (2014). Análisis de consultas al Diccionario del español de México en línea. *Andamios*, 11(26), 119-139.
- Müller, C. (1973). *Estadística Lingüística*. Biblioteca Románica Hispánica. Editorial Gredos.
- Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, 26(2), 1-16.
- Parodi, G. (2007). El discurso especializado escrito en el ámbito universitario y profesional: Constitución de un corpus de estudio. *Revista Signos*, 40(63), 147-178. <https://dx.doi.org/10.4067/S0718-09342007000100008>
- Parodi, G. (2008). *Géneros académicos y géneros profesionales. Accesos discursivos para saber y hacer*. Ediciones Universitarias de Valparaíso.
- Pérez, M. y Rojas, G. (2023). Análisis de las frecuencias léxicas verbales de redacciones universitarias de estudiantes asesorados de la Clínica de Escritura de la Facultad de Psicología de la UASLP, México. *Lengua y Sociedad*, 22(2), 357-375. <https://doi.org/10.15381/lengsoc.v22i2.24993>
- Pérez Durán, M. A. y Rojas Corona, M. G. (2025). Análisis de las frecuencias léxicas nominales, adjetivales y adverbiales presentes en escritos de estudiantes universitarios de la clínica de escritura de la Facultad de Psicología de la UASLP. *RILEX. Revista Sobre Investigaciones léxicas*, 8(1), 69-91. <https://doi.org/10.17561/rilex.8.1.8473>
- Read, J. (2010). *Assessing vocabulary* (9ª ed.). Cambridge University Press.

- Rodríguez, A. R., Pisco, R. J. L., Gómez, P. Á. P., y Quimis, O. (2020). Comprensión y manejo de la media aritmética, mediana y moda con datos agrupados en intervalos. (Original). *Roca: Revista Científico-Educaciones de la provincia de Granma*, 16(1), 1470-1483.
- Rojas, N. B. (2024). Conocimiento terminológico y escritura disciplinar: ¿Cómo se usan los términos para resolver tareas de escritura disciplinar en Psicología? *Debate Terminológico*, 2(1).
- Romero, E. I. (2024). La enseñanza del léxico en español de los negocios. *RILEX. Revista sobre investigaciones léxicas*, e8970-e8970.
- Silva-Pereyra, J., Rodríguez-Camacho, M. Prieto-Corona, B. y Aubert, E. (2014). *LEXMEX: Diccionario de frecuencias del español de México*. UNAM Editorial FES Iztacala.
- Spanish Learner Language Oral Corpora (SPLLOC). (s.f.). <http://www.sploc.soton.ac.uk/index.html>.
- Torres González, A. (2003). Riqueza léxica en textos narrativos escritos por estudiantes de Tenerife. En F. Moreno Fernández, F. Gimeno Menéndez, J. A. Samper, M. L. Gutiérrez, M. Vaquero y C. Hernández Alonso (Eds.), *Lengua, variación y contexto. Estudios dedicados a Humberto López Morales* (pp.435-449). Arco/Libros.
- Torres Perdígón, A. (2018). ¿Escritura disciplinar? Una experiencia a partir de la lectura crítica y la habilidad argumentativa en la universidad. *Revista mexicana de investigación educativa*, 23(76), 95-124.
- Torruella, J. y Capsada, R. (2013). Lexical statistics and typological structures: a measure of lexical richness. *Procedia-Social and Behavioral Sciences*, 95, 447-454.
- Valencia, A., Ávila, E., Muñoz Rigollet, G., Núñez, N., Echeverría M. S. y Veliz, M. (1992). Evaluación de la riqueza léxica en estudiantes de último año de enseñanza media. *Estudios Filológicos*, 27, 59-72.
- Velasco Luzuriaga, E. y Álvarez Álvarez, M. (2019). El resumen de trabajos fin de grado: diferencias discursivas según áreas disciplinares. *Traslaciones. Revista Latinoamericana De Lectura Y Escritura*, 6(11), 160-177. <https://revistas.uncu.edu.ar/ojs3/index.php/traslaciones/article/view/2006>
- Venegas, R. (2007). Clasificación de textos académicos en función de su contenido léxico-semántico. *Revista Signos*, 40(63), 239-271. <http://dx.doi.org/10.4067/S0718-09342007000100012>
- Venegas, R., Meza, P. y Martínez, J. (2013). Procedimientos discursivos en la atribución del conocimiento en tesis de lingüística y filosofía en dos niveles académicos. *RLA, Revista de Lingüística Teórica y Aplicada*, 51(1), 153-179. <https://dx.doi.org/10.4067/S0718-48832013000100008>
- Vine Jara, A., Fuentes Riffo, K., Neira Martínez, A., y Poza Molina, C. (2021). Riqueza léxica en textos escritos por aprendientes de ELE. Una contribución a las nociones específicas del plan curricular del Instituto Cervantes. *Nueva Rev. Pac.*, 7, 201-226. <http://dx.doi.org/10.4067/S0719-51762021000200201>