10.15446/fyf.v34n1.80581
Artículos

UNA PROPUESTA DE HERRAMIENTAS INFORMÁTICAS PARA EL TRATAMIENTO ESTADÍSTICO DEL ÍNDICE DE DISPONIBILIDAD LÉXICA EN ESTUDIOS CORRELACIONALES DE EDUCACIÓN Y MOVILIDAD SOCIAL[1]

A PROPOSAL OF COMPUTER TOOLS FOR THE STATISTICAL TREATMENT OF THE LEXICAL AVAILABILITY INDEX IN EDUCATION AND SOCIAL MOBILITY CORRELATION STUDIES

Dalia Reyes Valdés[2]
José R. Reyes Valdés[3]
María Eugenia Flores Treviño[4]
Rina B. Ojeda Castañeda[5]

Cómo citar este artículo:
Reyes Valdés, D., Reyes Valdés, J. R., Flores Treviño, M. E., & Ojeda Castañeda, R. B. (2021). Una propuesta de herramientas informáticas para el tratamiento estadístico del índice de disponibilidad léxica en estudios correlacionales de educación y movilidad social. *Forma y Función*, *34*(1). https://doi.org/10.15446/fyf.v34n1.80581

Recibido: 20-06-2019, aceptado: 21-07-2020

1 Avance de investigación de la tesis doctoral *La escuela secundaria como reguladora de los factores discursivos correlativos entre disponibilidad léxica y prospectiva de movilidad social*, realizada por Dalia Reyes Valdés para la Facultad de Filosofía y Letras de la Universidad Autónoma de Nuevo León (UANL). Este avance se elaboró durante estancia de investigación. Profundos agradecimientos al Centro de Investigación en Matemática Aplicada y Centro de Investigaciones Socioeconómicas, de la Universidad Autónoma de Coahuila (UAdeC).

2 ORCID https://orcid.org/0000-0001-7804-7050
Universidad Autónoma de Nuevo León, Nuevo León, México. dalia68reyes21@gmail.com

3 ORCID https://orcid.org/0000-0003-1014-9063
Universidad Autónoma de Coahuila, Saltillo, México. jose.reyes.valdes@uadec.edu.mx

4 ORCID https://orcid.org/0000-0001-7664-6674
Universidad Autónoma de Nuevo León, Nuevo León, México. maria.florestr@uanl.edu.mx

5 ORCID https://orcid.org/0000-0019152-558X
Universidad Autónoma de Coahuila, Saltillo, México. rinaojeda@uadec.edu.mx

Resumen

El trabajo plantea una propuesta para el procesamiento de índices de disponibilidad léxica (ámbito educativo) mediante herramientas estadísticas más eficientes que las usadas en la última década. El planteamiento asociado deriva una ruta de ejercicio interdisciplinar entre lingüistas y estadísticos que capitaliza los corpus lingüísticos, tratándolos como estudios correlacionales dentro del marco de la minería de datos. Se muestran resultados iniciales de la fase cuantitativa del estudio *La escuela secundaria como reguladora de los factores discursivos correlativos entre disponibilidad léxica y movilidad social*, procesados con metodología de lingüística de corpus y *software* estadístico libre, correlacionando, con alta eficiencia, índices de disponibilidad léxica y perspectiva de movilidad social en alumnos de secundaria en Saltillo, Coahuila, México, y como un corpus viable de ser correlacionado con bases de datos parciales o censuales para la toma de decisiones en el aula y la política pública por la posibilidad de correlación entre bases de datos.

Palabras clave: *educación; estudios correlacionales; herramientas tecnológicas; lingüística de corpus; minería de datos.*

Abstract

The work presents a proposal for the processing of lexical availability indexes (educational field) through the employment of more efficient statistical tools than the ones used in the last decade. The associated approach derives from an interdisciplinary exercise route between linguists and statisticians that capitalizes on a linguistic corpus, treating them as correlational studies within the framework of data mining. Initial results of the quantitative phase of the study *The high school as a regulator of the correlative discursive factors between lexical availability and social mobility* are shown, processed with corpus linguistics methodology and free statistical software, correlating, with high efficiency, indices of lexical availability, and perspective of social mobility in high school students in Saltillo, Coahuila, Mexico, and as a viable corpus to be correlated with partial or census databases for decision-making in the classroom and public policy due to the possibility of a correlation between databases.

Keywords: *corpus linguistics; correlational studies; data mining; education; technological tools.*

1. Introducción

El objetivo general de esta propuesta, derivada de la investigación sobre la influencia de la escuela secundaria en el acrecentamiento léxico de sus alumnos[1], es proponer el uso de herramientas estadísticas más eficientes para el procesamiento de índices de disponibilidad léxica en el ámbito educativo como una alternativa ante los procesadores diseñados *ad hoc* para la obtención exclusiva de índices y frecuencias. La primera instancia de la investigación se conformó con la fase de recogida de datos a partir de las encuestas estandarizadas para la obtención de palabras que, tras su proceso, se conformaron como la base de datos de vocablos que mostraron los índices de disponibilidad léxica en la población muestra, en este caso constituida por 116 estudiantes de secundaria.

Las herramientas usuales para procesar índices de disponibilidad léxica son, actualmente, IDL, Lexidisp y Dispolex, al menos entre las desarrolladas en específico para procesar los corpus léxicos, los cuales utilizan la fórmula para calcular índices de disponibilidad léxica desarrollada por López-Strassburguer (de la cual se habla en el desarrollo de este artículo). Al inicio, los datos recogidos en las encuestas para esta investigación se procesaron con IDL pero, posteriormente, se encontró que el *software* libre R ofrecía mayores ventajas como herramienta tecnológica para obtener el cálculo de los índices. Si bien aquí se presentan algunos de los resultados iniciales al respecto, el propósito de este texto es mostrar nuevas opciones de herramientas para realizar dicho proceso.

Durante la primera fase, consistente en establecer los índices de disponibilidad léxica (IDL, en adelante) contrastados entre estudiantes de escuela secundaria pública y privada, se encontró que las herramientas de procesamiento estadístico específicas para los conteos, que ofrecen la posibilidad de identificar índices de vocabulario frecuente y disponible entre los alumnos, resultan limitadas para la toma de decisiones del investigador cuando se trata de correlacionar o derivar acciones funcionales para el trabajo en aula o la propuesta dirigida a la política pública; aunada a lo anterior, está la ralentización del proceso. Ante esto, se determinó iniciar pruebas de procesamiento bajo la perspectiva de minería de datos utilizando el *software* libre R, mediante el trabajo interdisciplinar con el Centro de Investigación en Matemática Aplicada (CIMA), de la Universidad Autónoma de Coahuila, México.

En el primer apartado, se describen los instrumentos de toma de datos y la herramienta IDL utilizada hasta ahora para procesarlos, esbozando la justificación para el traslado a otras herramientas estadísticas, prioritariamente de ambiente libre. Se profundiza en los apartados posteriores sobre sus ventajas en tiempo y versatilidad.

Los estudios de IDL, hasta ahora, se han limitado a mostrar índices y frecuencias para ser contrastados en planos diatópicos, diacrónicos o diastráticos, sin retomar los

resultados como un corpus correlacionado con otros rubros de estudios sociales, en gran parte porque el procesador no ofrece muchas opciones automatizadas. En este artículo se plantea la propuesta de introducción de esta herramienta de alto desempeño y amplia disponibilidad en una aplicación guiada por la minería de datos.

2. Fundamentos

Los estudios sobre índices de disponibilidad léxica surgen en 1950, en Francia, con la intención única de establecer el vocabulario frecuente entre los franceses y organizar, en consecuencia, un lexicón básico para ser enseñado a extranjeros que aprenden el idioma (López, 2003).

A finales del siglo pasado, los índices de disponibilidad léxica se convirtieron en una propuesta formal para indagar sobre los estándares léxicos en sectores específicos de la sociedad, siendo el ámbito educativo en donde más se realizan estudios; sin embargo, las herramientas para el proceso de datos arrojados por la encuesta estandarizada de disponibilidad surgieron hasta inicios de este siglo. Consisten en *software* funcionales, pero se limitan a identificar frecuencias e índices de presencia de vocablos.

Una de las herramientas más utilizadas en México para el procesamiento de datos cuantitativos arrojados por la encuesta estandarizada para calcular el índice de disponibilidad léxica es el programa IDL, desarrollado en 2007 por Daniel Acosta Escareño, en la Universidad Autónoma de Zacatecas, México, para optimizar los cálculos estadísticos que ya tomaban auge en el país. Esta herramienta básica ayudó a disminuir la gran inversión de tiempo y esfuerzo al realizar tablas de índices de modo manual; el procesador no está disponible de manera pública, sino que se comparte de manera personal entre los investigadores del área. En la actualidad, las bases de datos cuantitativas disponibles para la investigación mixta y correlacional con el índice de disponibilidad léxica se incrementan y diversifican notablemente, rebasando las posibilidades del programa desarrollado por Acosta, circunstancia que detona el objetivo de este estudio para encontrar nuevas y más eficientes opciones para procesar y correlacionar datos generados por corpus léxicos.

Los conteos de disponibilidad léxica conforman un corpus valioso por las posibilidades de estudios interpretativos y correlacionales que surgen a partir de paradigmas metodológicos modernos, como la minería de datos (Shmueli, 2018), en donde se propone el aprovechamiento de bases numéricas. No se han localizado instrumentos metodológicos para correlacionar índices léxicos e interpretación discursiva del mismo universo, como fue el ejercicio realizado en la investigación que propició este artículo. Por ello, se plantea que esos instrumentos deben diseñarse *ad hoc*, como el caso de la

encuesta desarrollada para este estudio; sin embargo, sí se esboza aquí un ejercicio de redes semánticas utilizando también R Statistics.

El programa IDL se alimenta de datos organizadas en hojas de notas tipo texto sin formato, y el *software* arroja tablas de Excel (Tablas 1 y 2) con los índices de disponibilidad léxica, es decir, organiza los vocablos por jerarquía de aparición según el lugar donde fueron escritos en la encuesta. Esto muestra cuáles son los que están en el léxico frecuente del hablante por estar escritos en la parte superior de las columnas y luego, de forma descendente, los que aparecen hasta el final (responde al sustento psicológico de la encuesta, el asociacionismo). Esto permite distinguir entre léxico frecuente y léxico disponible.

Por último, el programa Vocablos, diseñado por Flavio Darío Mirelez en 2012 (también de la Universidad Autónoma de Zacatecas, y tampoco disponible) se diseñó para empatar el léxico compartido entre dos muestras. Fue un atisbo a los estudios cualitativos correlacionales, porque los resultados que ofrece permiten identificar tipos de vocablos comunes en diferentes colectividades escolares; ello permite saber el nivel de influencia de la escuela, en particular, y del contexto social, en general, para la evolución o involución (Arriaga, 2003) en el léxico frecuente. La correlación, sin embargo, seguía siendo al interior de la muestra.

Un punto axial en el objetivo de este estudio fue la conjunción de un modelo estadístico, particularmente en el sentido de distribución de probabilidad, y un conjunto de datos, que aportan los elementos necesarios para predecir comportamientos (de carácter aleatorio) y realizar contrastes entre grupos o momentos en el tiempo cuando estos son generados. La forma como se organizan conlleva el grado de dificultad para extraer información a partir del léxico, esencialmente porque se presentan de manera estructurada (clasificados y organizados en tablas) o no estructurada (textos en prosa), resultando que los datos estructurados facilitan su tratamiento estadístico, porque los no estructurados requieren de un tratamiento previo, que es fundamental para su procesamiento y posterior análisis.

Este tratamiento no solo se reduce a organizarlos, sino que requiere de un trabajo de limpieza y de homogeneización para llevarlos a una organización estructurada sin perder el significado de su forma original. Janert (2010) menciona que un conjunto amorfo de datos puede ser traducido a ideas concretas asociadas al contexto de aplicación, lo que significa que se puede extraer información útil a partir de datos.

2.1. Minería de datos en corpus lingüísticos

En el contexto de análisis de datos, los modelos estadísticos juegan un rol crucial y son utilizados para representar estructuras de naturaleza aleatoria (Konishi & Kitagawa,

2008). Si bien un modelo es una abstracción (Janert, 2010) del fenómeno que representa, una vez establecido permite, entre otras cosas, extraer información de un conjunto de datos. Aquí la palabra *información* se expresa en dos sentidos: revela aspectos que no están explícitos en los datos crudos y como una medida cuantitativa (Ayres, 1994); más aún, los propios datos sirven para validar la calidad del modelo (Torgo, 2011) o bien para hacer predicciones del sistema o fenómeno de estudio.

La minería de datos (Shmueli, 2018; Torgo, 2011) es un conjunto de técnicas que permite realizar este tipo de tareas. La naturaleza de los datos y la información que se pretende extraer de estos determina la complejidad para su procesamiento, análisis e interpretación en el marco del problema de estudio. Los datos recabados en encuestas con preguntas cerradas facilitan tanto la organización como el procesamiento, ya que los resultados obtenidos son generalmente cuantitativos, como es el caso de los índices de disponibilidad léxica.

Cuando los datos provienen de textos abiertos, su tratamiento se encuadra en lo que se define como procesamiento de lenguaje natural (NLP por sus siglas en inglés) y es parte de la denominada minería de textos (Silge & Robinson, 2017). Lo anterior incrementa la dificultad en la interpretación: si un texto se asocia a una opinión del encuestado, crece la complejidad al estar involucrado el factor de subjetividad. En este caso, la metodología para su tratamiento se denomina minería de opinión y, particularmente, análisis de sentimiento (Liu, 2015).

El desarrollo de las tecnologías de información y comunicación (TIC) derivó también en la capacidad de almacenar grandes cantidades de datos en forma digital (Kelleher & Tierney, 2018). Particularmente, a partir del 2000 se empieza a explotar la información generada mediante el uso de técnicas de minería de datos. El análisis de datos lingüísticos (Baayen, 2008) empieza a tomar relevancia, por lo que diversas herramientas informáticas integran la capacidad para procesar datos de esta naturaleza. Ante el volumen de corpus generados, estas herramientas se vuelven imprescindibles y sin ellas no se tendría la capacidad de manejar tal cantidad de información. Para el tratamiento de datos enfocados a la minería de opinión, se dispone de herramientas comerciales y de alternativas libres de alto desempeño.

3. Metodologías usuales para procesar corpus lingüísticos

En la investigación que generó esta propuesta, relativa al incremento léxico en el nivel educativo de secundaria, se priorizó la enseñanza de la lengua y su dimensión social, en el sentido de aportar elementos constructivos para que los estudiantes de secundaria

tengan elementos prospectivos de movilidad social en lo mediato e inmediato. La metodología correlacional propuesta requirió definiciones categoriales para ampliar el alcance de los estudios de disponibilidad léxica hasta la movilidad del individuo en el espacio social.

Michea (1953) aporta la propuesta conceptual de disponibilidad léxica que se conformó como la base para los estudios posteriores a la década de los años cincuenta, cuando el autor inicia con ellos en Francia (López, 2003). La propuesta es la siguiente:

> Una palabra disponible es una palabra que, sin ser particularmente frecuente, siempre está lista para ser utilizada y se presenta de forma inmediata y natural al espíritu en el momento en que se necesita. Es una palabra que, siendo parte de asociaciones de ideas usuales, existe en el poder en el sujeto que habla, tan pronto como estas asociaciones entran en juego. (Michea, 1953, p. 340)

Un importante impulsor de los estudios de disponibilidad léxica en México es López (2003), quien la entiende como el caudal léxico utilizable en una situación comunicativa dada: «Esta percepción implica vislumbrar que la lengua poseía un amplio conjunto de palabras de contenidos semántico muy concreto que únicamente manejaba si lo permitía el tema del discurso» (p. 1). Con fundamento en ambas propuestas conceptuales, para esta investigación se desarrolló la siguiente definición operativa, con una dimensión de simultaneidad constructiva de los símbolos sociales, es decir, en esta definición el léxico adquirido no tiene una frontera funcional para el presente y el aquí, sino que opera al mismo tiempo en retrospectiva y en la posibilidad.

La definición operativa es la siguiente: disponibilidad léxica es el capital de léxico utilizable en el que cabe, además de un número finito de palabras, la representación individual y colectiva del mundo social al que pertenece una persona cuyo acrecentamiento o limitación determinarán el nivel prospectivo que en el imaginario se edifica para aspirar a otros espacios sociales.

Como el capital léxico no es inmutable, y en este estudio se busca su dimensión prospectiva, en lo que respecta a la concepción de espacio social, también se propuso una definición operativa. Se parte de las propuestas de Bourdieu (2011), quien entiende el espacio social como un campo virtual de clases en estado potencial, y de Bauman (2005), quien lo asume como un espacio físico asimilado fenomenológicamente para marcar las cercanías o las distancias sociales. La definición operativa para espacio social que surgió en el trayecto de esta investigación se aludió ya en Reyes y Flores (2018), aunque la propuesta final para espacio social se determinó de la siguiente manera: el *topos* en

donde coexisten los productos sociales que conforman cada *habitus* particular y donde actúan una serie de variables correlacionadas a partir de las formaciones imaginarias desarrolladas por una comunidad[2].

La relación entre las propuestas conceptuales anteriores y la metodología de correlación realizada en esta propuesta radica en que los vocablos se entendieron como partículas detonantes de discursos completos. Es decir, cada palabra escrita por el informante se consideró como indicador de una ideología personal respecto de sus posibilidades de movilidad social en el futuro. De esta manera, la relación teórica con la propuesta metodológica radicó en la búsqueda de herramientas que permitieran correlacionar aspectos emanados del léxico con elementos contextuales del espacio social en donde se ubica el universo de estudio.

La aplicación de instrumentos se hizo de manera longitudinal: cuando los alumnos cursaban primer grado de secundaria y, posteriormente, en tercer grado, para conocer la modificación operada en su disponibilidad léxica en el trayecto del nivel escolar. Para la recolección de datos, se diseñaron encuestas con siete centros de interés[3], cuatro de ellos hechos *ad hoc* como motivadores de palabras relativas a movilidad social prospectiva, que fueron «Empleos que ocuparé en el futuro», «Lugar en donde viviré», «Actividades que haré en el futuro» y «Así me veo en comparación con los demás». El último centro de interés se incluyó porque la fundamentación social y pedagógica del estudio completo asume que la construcción de la autoimagen (Gofman, 1997) es determinante en la expresión léxica al reconocer su estatus social un individuo. Los centros de interés «La ciudad», «Profesiones» y «La casa el interior y sus partes» fueron tomados de la encuesta estandarizada propuesta por Michea (1953), quien propuso los 16 centros de interés[4] que se convirtieron en la base instrumental para la recolección de datos en los estudios posteriores.

En su narración, López (2003) reseña el recorrido histórico de los estudios de disponibilidad léxica a partir de la década de los cincuenta y cómo se obtuvieron resultados léxico-estadísticos, con el fin conformar un banco de vocablos que estableciera un estándar para la enseñanza del francés a extranjeros, sin tener, propiamente dicha, una discusión previa para establecer índices. La sistematización de los datos obtenidos con el conservador propósito mencionado se convirtió luego en una propuesta metodológica de enseñanza del francés. Los primeros estudios con la intención de buscar índices se realizaron en Francia y en Canadá entre 1956 y 1971, y eran cómputos de frecuencia; posteriormente, continuaron Inglaterra, Puerto Rico y España, entre 1969 y 1983, periodo cuando se diseñaron las fórmulas para ponderar la frecuencia de las palabras. La primera fórmula fue de Lotán-López Morales; después, apareció la aportación mexi-

cana de López Chávez-Strassburger, en 1987 (López, 2003), de la cual se reconoce lo siguiente: «El manejo de la fórmula Strassburger-López Chávez ha demostrado hasta ahora su superioridad para la lingüística, pues logra producir una adecuación descriptiva sumamente plausible» (López, 1996, pp. 2-3).

Las propuestas de centros de interés adaptados a un objetivo particular de investigación menudean actualmente, sobre todo en el ámbito educativo, en donde los estudios de disponibilidad léxica arrojan un conocimiento previo de ciertos temas sobre los que se debe de incidir en clase. Algunos ejemplos son los realizados por Cortez (2016), con centros de interés para indagar conocimientos previos de literatura en nivel universitario, y Pérez (2020), quien estudia «Insulto» como detonante y, además, propone como instrumento metodológico el análisis de discurso oral, en lugar de la encuesta escrita.

Las definiciones aportadas en este estudio (disponibilidad léxica y espacio social) sustentaron la construcción de los centros de interés y la finalidad correlativa para la que se buscaron las herramientas tecnológicas. Esos conceptos buscaron indagar rasgos discursivos prospectivos de movilidad social en los índices de disponibilidad léxica en estudiantes de secundaria; de ahí que, en la fase de estadística correlacional, se localizaran bases de datos sociodemográficos para contrastarlos con la percepción expuesta, mediante los vocablos, del contexto en el que se desenvuelven. Los vocablos con mayor índice se consideraron como señales emergentes que describen su entorno geográfico y familiar tanto presente como prospectivo. Como se verá, de las bases de datos generadas por el Instituto Nacional de Estadística Geografía e Informática (Inegi) en México, se toman factores sobre tipos de familias en donde está inserta la muestra, así como nivel de estudios de los padres.

En la fase de pruebas del estudio se procesaron, inicialmente en el programa IDL, los resultados de 116 encuestas aplicadas en las escuelas pública y privada que constituyen la muestra, utilizando la encuesta estandarizada, la hoja de notas para la organización de datos y el procesador de Acosta para calcular los índices.

La encuesta estandarizada recogió el vocabulario frecuente de los estudiantes de primero de secundaria bajo los centros de interés descritos para contrastar el avance que, tentativamente, tendrían los alumnos en el plano numérico entre el inicio y el término de la escuela secundaria, ya que la aplicación es longitudinal.

La metodología consiste en recabar las palabras escritas por los estudiantes, organizarlas de forma horizontal en hojas de notas en un formato básico de palabras separadas por comas (Figura 1) y, posteriormente, vaciarlas al programa IDL que procesó, en aproximadamente 25 minutos, los resultados de las 116 encuestas, arrojando los índices como aparecen en la Figura 2.

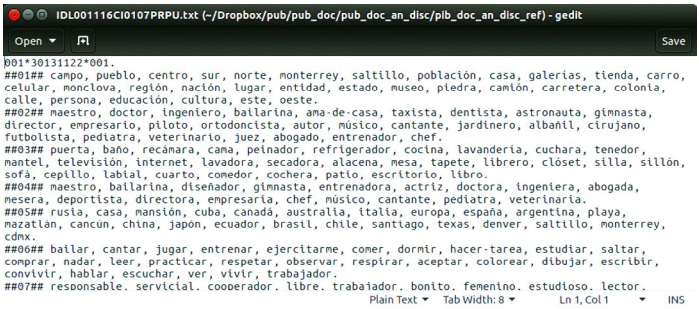
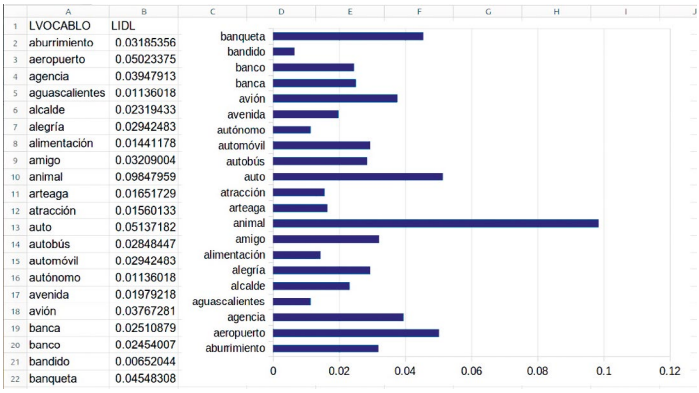


Figura 1. Organización en hoja de notas de palabras recabadas en encuesta estandarizada para ser procesadas en IDL



	LVOCABLO	LIDL
2	aburrimiento	0.03185356
3	aeropuerto	0.05023375
4	agencia	0.03947913
5	aguascalientes	0.01136018
6	alcalde	0.02319433
7	alegría	0.02942483
8	alimentación	0.01441178
9	amigo	0.03209004
10	animal	0.09847959
11	arteaga	0.01651729
12	atracción	0.01560133
13	auto	0.05137182
14	autobús	0.02848447
15	automóvil	0.02942483
16	autónomo	0.01136018
17	avenida	0.01979218
18	avión	0.03767281
19	banca	0.02510879
20	banco	0.02454007
21	bandido	0.00652044
22	banqueta	0.04548308

Figura 2. Índice de disponibilidad léxica, con gráfica, arrojado por IDL

La propuesta de investigación implica correlacionar los vocablos con mayor frecuencia arrojados en la fase de conteo de índices con una segunda fase cualitativa para conocer proyectivas de movilidad social. Por ello, fue necesario utilizar un programa con la capacidad de organizar los datos con mayor flexibilidad para ser relacionados con otras bases. Para tal efecto, se propuso la utilización de la herramienta (lenguaje) R.

El lenguaje R es un ambiente libre desarrollado por *R Foundation for Statistical Computing* (R Core Team, 2018) utilizado para cómputo científico, predominantemente para el procesamiento de resultados del ámbito de la estadística, donde se circunscribe la minería de datos. Si bien es un sistema que debe ser manipulado por un experto en estadística, los crecientes proyectos de investigación sobre léxico en redes sociales han construido el foro para el trabajo colaborativo entre numerosas disciplinas: lingüistas, estadísticos, psicólogos, pedagogos y mercadólogos.

R es un lenguaje gratuito, robusto, interpretado por el sistema y dispone de miles de librerías para procesamiento de datos de diversa naturaleza, tanto en formato como en modelos estadísticos. El procesamiento de lenguaje natural no es la excepción en R, ya que cuenta con librerías especializadas para este fin (R Core Team, 2018). Otro lenguaje de programación general que pudiera utilizarse como herramienta de apoyo para trabajar con corpus es Python (Perkins, 2014); sin embargo, requiere de una formación más especializada para su dominio.

Otras herramientas complementarias para la organización, almacenamiento, procesamiento y presentación de resultados son MySQL (Du Bois, 2009), un servidor de datos y QGIS (Graser, 2013) para la representación geoespacial. MySQL concentra datos estructurados y normalizados cuya función es proveer de estos a otros programas para su procesamiento, mientras que QGIS se utiliza para la representación de indicadores en un contexto geoespacial. En este trabajo, por su robustez y flexibilidad, el lenguaje R se utilizó para el núcleo de procesamiento de datos, MySQL para almacenar y proveer datos organizados, y QGIS para la representación de mapas asociados a variables de contexto socioeconómico.

3.1. Las posibilidades de R para los estudios correlacionales de léxico compartido

A diferencia de IDL, R se alimenta no solo de datos con una organización básica en hojas de cálculo, como Excel, por ejemplo, sino de diversos formatos de datos estructurados. En la Figura 3 se observa cómo se ordenaron datos cuantitativos y cualitativos para estudiarse de forma independiente, por variable o confrontados.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CE	CM	CL	NI	CI	NA2	GE	EDAD	SEXO	TEXTO			
2	5	30	1	1	1	3	1	12	2	campo, pueblo, centro, sur, norte, monterrey, saltillo,			
3	5	30	1	1	2	3	1	12	2	maestro, doctor, ingeniero, bailarina, ama-de-casa, ta			
4	5	30	1	1	3	3	1	12	2	puerta, baño, recámara, cama, peinador, refrigerador,			
5	5	30	1	1	4	3	1	12	2	maestro, bailarina, diseñador, gimnasta, entrenadora,			
6	5	30	1	1	5	3	1	12	2	rusia, casa, mansión, cuba, canadá, australia, italia, e			
7	5	30	1	1	6	3	1	12	2	bailar, cantar, jugar, entrenar, ejercitarme, comer, do			
8	5	30	1	1	7	3	1	12	2	responsable, servicial, cooperador, libre, trabajador, b			
9	5	30	1	2	1	3	1	12	2	farola, carro, pavimento, calle, edificio, ventana, puer			
10	5	30	1	2	2	3	1	12	2	bombero, policía, astronauta, pedagogo, médico, mae			
11	5	30	1	2	3	3	1	12	2	lámpara, baño, excusado, lavamanos, cama, buró, esp			
12	5	30	1	2	4	3	1	12	2	astronauta, veterinaria, médica, neuróloga, cantante,			
13	5	30	1	2	5	3	1	12	2	mansión, londres, piscina, colombia, luna, marte, júpi			

Figura 3. Hoja de cálculo (Excel) con datos estructurados obtenidos en la encuesta de disponibilidad léxica, para procesarse en R

En la Figura 4 se aprecia el diagrama de componentes de R para procesar los datos a partir de las librerías desarrolladas e integradas al programa, utilizadas por el especialista en el área de estadística. Las librerías *qdap*, *tm* y *wordcloud* están disponibles en R, y

el resto fueron desarrolladas por el especialista. Mientras IDL calcula exclusivamente el índice de disponibilidad léxica, este programa hará procesos diversos mediante el uso de algoritmos adecuados, de acuerdo con las necesidades del investigador.

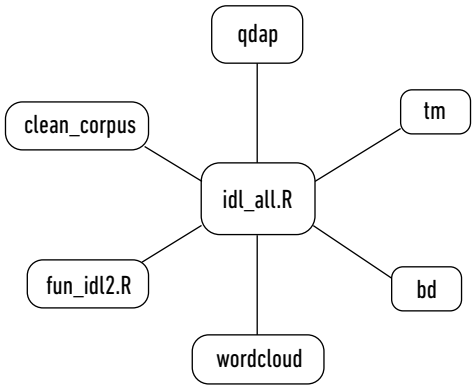


Figura 4. Diagrama de componentes en R para procesar datos de encuesta organizados en hoja de cálculo a partir del uso de librerías específicas

En las Figuras 5 (asociación con un vocablo clave), 6 (frecuencia de vocablos) y 7 (nube de palabras) se observa cómo el resultado indicial se logra igual que en IDL, pero con ventajas de tiempo y la posibilidad de representaciones visuales en las distintas etapas del procesamiento. Si IDL procesa una centena de datos en 25 minutos, R lo realiza en aproximadamente 6.5 segundos por cada centro de interés (un minuto total) y lo muestra como se aprecia en las imágenes, en gráficas de diseño variado, dispersión, datos visuales o frecuencias.

```
$casa
      escuela         tienda       edificio          árbol
         0.59           0.51           0.49           0.47
       carro         parque        persona         calle
         0.47           0.44           0.44           0.43
      puente       hospital    restaurante centrocomercial
         0.41           0.37           0.37           0.35
```

Figura 5. Grado de asociación de vocablos con el vocablo clave (casa) mediante R

En la Figura 6 se organizaron automáticamente en R los datos en barras horizontales para conocer las frecuencias. En IDL, obtener la frecuencia requiere de un segundo cálculo con valores tomados de los automatizados de Excel, cuya fórmula en este caso es más compleja.

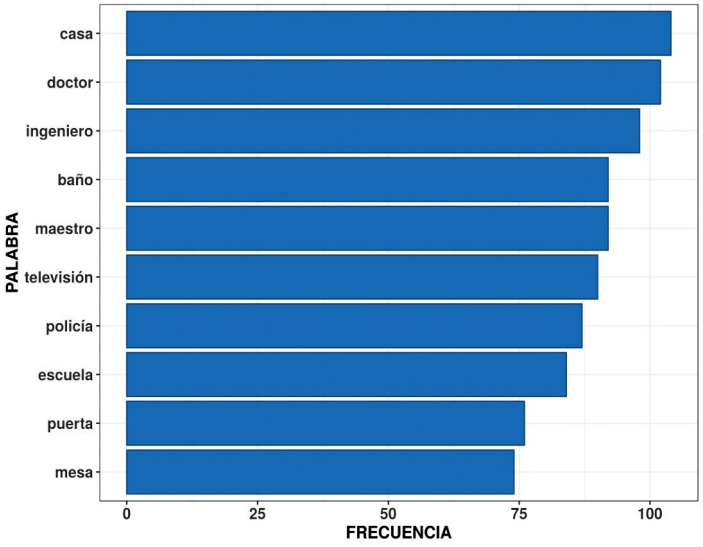


Figura 6. Diagrama de barras de frecuencia de palabras obtenidas en R

Por último, un ejemplo de las representaciones alternativas visuales que arroja R a partir de las tablas de frecuencias se muestra en la Figura 7; este modelo de cuantificar frecuencias se denomina *wordcloud* o *nube de palabras*. Para los fines de este estudio, el recurso gráfico es de gran importancia, considerando que se pretende llevar al docente a que identifique, de forma rápida y sencilla, el vocabulario más usado, el de menor índice y el particular. El programa «traduce» la frecuencia numérica en tamaño de la fuente e intensidad del color.

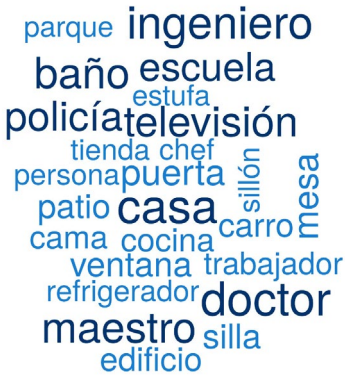


Figura 7. Nube de palabras (*wordcloud*) de R con índices representados por tamaños y colores

Existen algunos ejercicios para procesar corpus léxicos con LexiDisp (Moreno, 1995) y DispoLex (Echeverría & Parada, 1990), herramientas generadas *ad hoc* para resultados en índices de disponibilidad léxica; SPSS tiene herramientas para minería de datos enfocada a aplicaciones de corte empresarial, teniendo como desventaja que es un programa comercial con costos privativos, además de otras limitaciones como el manejo eficiente de grandes bases de datos.

En contraste, el lenguaje R, aparte de ser gratuito, cuenta con miles de librerías dentro de las cuales existe una gran cantidad enfocada a minería de datos (Baayen, 2008; Torgo, 2011) y tiene la capacidad de manejar grandes volúmenes de datos, millones de registros y cientos de variables; además, es un lenguaje factible para combinar librerías existentes y generar nuevas librerías *ad hoc* para procesos definidos por el usuario, como lo sería el IDL, en particular. Más aún, R actúa como una plataforma de desarrollo, ya que interactúa con otras herramientas libres y formatos de programas comerciales.

Algunos de los estudios realizados con los *software* arriba mencionados son *Incidencia de la variable* «sexo» *en la disponibilidad léxica de estudiantes preuniversitarios en Pinar del Río, Cuba* (Pacheco, 2016), *Metodología de análisis de disponibilidad léxica en alumnos de Pedagogía a través de la comparación jerárquica de léxicos* (Rojas, 2017) o *Estudio de disponibilidad léxica en 43 estudiantes de ELE* (López, 2008); todos ellos utilizaron paquetes estadísticos diferentes a IDL, pero diseñados específicamente para el cálculo de índices de disponibilidad léxica. Como se puede inferir por los títulos, los propósitos de investigación estuvieron orientados hacia estudios intrínsecos sobre el mismo corpus.

3.2. Obtención del IDL mediante lenguaje R

En la sección anterior, básicamente se contabilizaron los vocablos por su frecuencia en todos los centros de interés; sin embargo, cada vocablo tiene relevancia basado en su correspondiente índice de disponibilidad. Mediante el lenguaje R, se desarrolló el código necesario para calcularlo; el tiempo de procesamiento para cada uno de los centros de interés fue de 6.5 segundos aproximadamente. La fórmula utilizada como referencia para el cálculo del IDL es la propuesta por López-Strassburguer (López, 2003):

$$D(P_i)=\sum_{j=1}^{n_i} e^{(-2,3)\left(\frac{j-1}{n_i-1}\right)}\left(\frac{f_{ij}}{I}\right)$$

Como muestra de los resultados obtenidos, se construyeron las Tablas 1 y 2, que contienen los vocablos e índice de disponibilidad, respectivamente, con los primeros 10 valores de mayor disponibilidad dados en forma decreciente.

Tabla 1. Diez vocablos con mayor IDL en cada uno de los centros de interés

ID	C1	C2	C3	C4	C5	C6	C7
1	edificio	doctor	baño	doctora	parís	viajar	feliz
2	casa	maestro	puerta	maestra	mansión	trabajador	alto
3	carro	ingeniero	televisión	chef	canadá	estudiar	enojón
4	escuela	policía	mesa	veterinaria	españa	dormir	alegre
5	persona	abogado	cocina	ingeniero	japón	comer	inteligente
6	calle	arquitecto	patio	doctor	playa	jugar	amigable
7	árbol	bombero	silla	policía	argentina	bailar	divertido
8	parque	licenciado	sillón	licenciada	cdmx	futbol	amable
9	tienda	veterinario	ventana	arquitecto	italia	correr	bueno
10	contaminación	médico	cama	maestro	francia	cantar	risueño

Tabla 2. Diez valores más altos de IDL en cada uno de los centros de interés

ID	C1	C2	C3	C4	C5	C6	C7
1	0.4232	0.5013	0.5094	0.1774	0.3088	0.3025	0.1972
2	0.4030	0.4587	0.3877	0.1739	0.1613	0.2654	0.1880
3	0.3379	0.3654	0.3869	0.1621	0.1523	0.2344	0.1752
4	0.2867	0.2619	0.3704	0.1475	0.1465	0.1692	0.1610
5	0.2593	0.2506	0.3657	0.1397	0.1419	0.1619	0.1568
6	0.2313	0.2131	0.3493	0.1295	0.1346	0.1400	0.1288
7	0.2206	0.2005	0.3470	0.1117	0.1325	0.1035	0.1166
8	0.2203	0.1974	0.3468	0.0931	0.1236	0.0985	0.0977
9	0.2201	0.1503	0.3369	0.0881	0.1233	0.0928	0.0853
10	0.1795	0.1330	0.3152	0.0873	0.1209	0.0854	0.0839

La relevancia del C7 («Así soy en comparación con los demás») radica en localizar la individualidad del informante, datos que, para la segunda fase de la investigación, permitieron la selección de la muestra para el proceso interpretativo a través del análisis del discurso. Es importante mencionar que los índices de disponibilidad léxica se han utilizado, prioritariamente, para la generalización, pero en esta propuesta la individualidad es relevante para la toma de decisiones en el aula.

En la Figura 8 se representan los valores de índices mayores a cero en forma decreciente para el primer centro de interés. Una tendencia similar se observa en los restantes centros de interés.

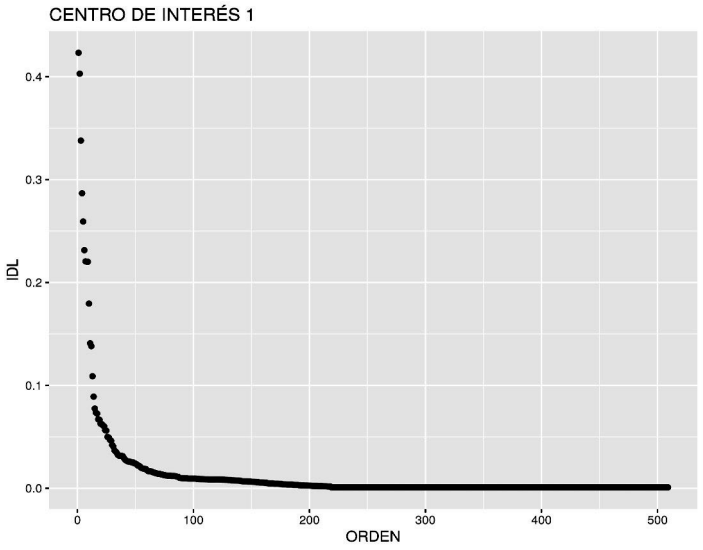


Figura 8. ***Índice de disponibilidad léxica en forma decreciente para el primer centro de interés***

El comportamiento del índice es similar al observado directamente en las frecuencias de vocablos, aunque no necesariamente en el mismo orden. La ley de Zipf (Urbizagástegui & Restrepo, 2011) se aplica a la relación de la posición con el valor calculado; dado este comportamiento, es factible hacer comparaciones entre centros de interés y vocablos, tanto con centros de interés como con segmentos de población diferenciados por alguna característica.

4. Análisis lingüístico y de sentimientos con R

En el análisis de textos, sean estos estructurados o discontinuos, existe una variedad de resultados estadísticos que se pueden derivar. Los más generales son de tipo descriptivo, tanto numérico como gráfico; entre estos, la frecuencia con que aparece cada palabra, o secuencia de palabras o n-gramas (Silge & Robinson, 2017), aporta información descriptiva preliminar para guiar, en primera instancia, cálculos más elaborados. Esta primera aproximación basada en frecuencias de palabras se le denomina *wordcloud* (se mantiene el término en inglés, ya que se asocia al nombre de la técnica más que a una traducción literal de la expresión).

La variedad de cálculos estadísticos que se pueden realizar en la minería de datos depende del programa en que se realicen; los programas dedicados a una tarea específica suelen ser más acotados que los de propósito general. En el contexto de minería de datos, el lenguaje R presenta algunas ventajas sobre los de propósito general y los especializados: es libre; se enfoca al procesamiento estadístico en general; existen miles de librerías (Adler, 2009) aplicadas a distintas disciplinas; genera gráficas complejas y de alta calidad; interactúa con otros programas (incluidos los comerciales); facilita realizar los procesos bajo un esquema de investigación reproducible; es factible crear librerías propias, y la curva de aprendizaje para usuarios finales es de corto plazo.

Para alcanzar el objetivo de este estudio, consistente en hacer más eficientes los procesos de léxico con herramientas tecnológicas, las características enlistadas fueron determinantes en particular por la flexibilidad que ofrece para generar librerías específicas que permitieron el procesamiento de los corpus, desde el cálculo del índice, pasado por la estadística correlacional con el contexto sociodemográfico, hasta la posibilidad de resaltar análisis léxico de importante relación con el ámbito interpretativo y de análisis del discurso.

Una librería o paquete (Adler, 2009) en R es un conjunto de funciones que realizan una tarea especializada. R viene con cientos de librerías preinstaladas en el programa base y, dependiendo de las necesidades, se pueden instalar miles más. Por ejemplo, la librería *tm* se enfoca en tareas relacionadas con minería de textos; *ggplot2* se utiliza para generar gráficas complejas; y *dplyr*, para realizar filtros y operaciones de datos agrupados a partir de tablas o *data frames*. Las librerías permiten la automatización de ciertos procesos a partir de algoritmos predeterminados en estas, y eso es lo que hace tan eficiente en tiempo y versatilidad a esta herramienta.

El segundo nivel de procesamiento es encontrar asociaciones entre palabras que, en el campo de la estadística, se denominan *correlaciones*; en esta fase, la representación gráfica de resultados tomó relevancia. Por ejemplo, una gráfica de Paretto consiste en ordenar las palabras en sentido decreciente, de acuerdo con su frecuencia, para luego representarlas mediante un gráfico de barras. Este tipo de gráfica da una perspectiva inmediata de las palabras más utilizadas, además de que se pueden construir variantes de esta para ser aún más informativas, inclusive sin observar los valores numéricos correspondientes.

El manejo de datos obtenidos a partir de corpus de disponibilidad léxica se ha limitado al comparativo al interior de la muestra o a la detección de coincidencias y divergencias entre dos muestras similares. La relevancia que implica manejar los datos con otras herramientas de procesamiento estadístico alcanza la correlacionalidad entre

corpus diversos en cuanto a tipo de informante, ubicación, género, posición económica y nivel académico; incluso, se abre una vertiente de análisis discursivo, considerando que ya se trabaja el parámetro «emocionalidad» usando herramientas disponibles en el lenguaje R para realizar lo que se denomina *análisis de opinión* y *análisis de sentimiento* (Liu, 2015), que se enfocan en extraer información a partir de textos no estructurados basados en la percepción de las personas.

Entre las grandes empresas que hacen minería de datos en tiempo real se destacan Amazon, Google, Netflix, Facebook y redes sociales en general (Russell, 2014). El factor común de estas empresas es que utilizan los perfiles personales de los usuarios y sus hábitos de navegación en web, para crear lo que se denomina *ávatar* del usuario, el cual es un vector de características que lo tipifican. En el caso de Amazon y Netflix, hacen sugerencias de consumo a los usuarios, inclusive dando una estimación en la similitud con los productos o servicio en su historial de consumo. Por su parte, Netflix da, de manera porcentual, la similitud de un contenido propuesto con los ya consumidos.

El análisis de sentimiento es un ámbito de relevancia para la psicopedagogía en las actuales consideraciones programáticas de la Secretaría de Educación Pública en México; una evidencia de ello es la inclusión de la educación socioemocional como un rubro explícito en los planes de educación básica, por lo que las estrategias neuroeducativas, estrechamente ligadas al léxico, se vuelven una condición para alcanzar aprendizajes esperados. Los corpus de disponibilidad léxica, fundamentados en el asociacionismo, tienen gran relación con el contexto psicosocial del cual está rodeado el estudiante; por lo tanto, se convierten en signos interpretables de una realidad viable de ser intervenida por la escuela.

En relación con lo anterior, la propuesta en este documento consiste en ubicar el manejo de datos de disponibilidad léxica dentro de la lingüística de corpus, entendido su interés como dentro del análisis lingüístico en los ámbitos léxico, gramatical, semántico y pragmático mediante herramientas diseñadas para este fin (Bolaños, 2015). Se propone su procesamiento como minería de datos desarrollado de modo correlacional contextual al estudiarla simultáneamente con corpus censales en zonas de incidencia socioeconómica a las cuales pertenecen los informantes.

La correlación contextual, en el caso específico de este estudio, atiende a que los corpus censales se analizan como el contexto de origen y perspectiva de movilidad en el cual se desenvuelven los informantes con quienes se construyeron los corpus de índices de disponibilidad léxica, considerando que esta población, integrada por menores de edad, no fungió como informante para el corpus censal, pero sí está inserta en el Área Geográfica Estadística Básica (AGEB) de la cual se toman los rubros para la correlación.

5. Modelo correlacional a partir de corpus construidos con índices de disponibilidad léxica

Las tendencias de investigación hoy en día incurren en estudios correlacionales por las posibilidades que ofrece el acceso a la información. Si la recogida de datos y la construcción de corpus era trabajo de un investigador en particular, hoy el volumen de datos públicos y accesibles que existen en la red y en los portales de información pública abren una vertiente de estudios correlacionales colaborativos con la aportación personal de datos y su enfrentamiento a los publicados en censos diversos.

La propuesta metodológica central en este proyecto apuesta por generar un corpus viable de ser correlacionado con (1) un segundo corpus cualitativo generado por los mismos informantes, (2) bases de datos en coincidencia con franjas sociogeográficas-económicas y (3) algunos rubros de censos nacionales de población.

Es factible establecer, en una primera instancia, una correlación del índice de disponibilidad a nivel de persona, asociándolo al rendimiento escolar. En una segunda instancia, la correlación se obtiene en relación con el desempeño de la escuela en evaluaciones oficiales o las internacionales. A partir de información de datos censales, se correlaciona con unidades geográficas que contienen una o más escuelas; particularmente en zonas urbanas, se obtiene el Grado de Rezago Social (GRS) a nivel AGEB formada a partir de un conjunto de manzanas contiguas.

Conapo (2019) construye el GRS a partir de once variables agrupadas en cuatro ejes: (1) rezago educativo, (2) acceso a los servicios de salud, (3) calidad y espacios de la vivienda, y (4) servicios básicos en la vivienda e ingreso (bienes del hogar). El censo 2010 contiene 51.034 AGEB a nivel nacional y, dado que las condiciones socioeconómicas no cambian a corto plazo, puede ser utilizado durante el periodo previo al siguiente censo en 2020.

5.1. Ámbitos de aplicación

Los ámbitos de aplicación de un procesamiento estadístico correlacional extendido de corpus de disponibilidad léxica se prospectan, prioritariamente, en el de minería de datos cuyos resultados se abocan a beneficiar a la educación, los estudios psicopedagógicos y sociales, y la toma de decisiones en política pública.

En la escuela, la obtención de índices visualmente accesibles, correlacionados entre léxico y resultados de evaluaciones estandarizadas de la institución correspondiente o características socioeconómicas del entorno escolar, permitirá desde la toma de decisiones en el aula y estrategias alternativas en la escuela, hasta iniciativas en la política pública. Los estudios de involución léxica marcaron un punto de partida en este sentido (Arriaga, 2003).

Para obtener los primeros resultados de correlación contextual entre el corpus de este estudio y los datos del Censo Inegi 2010, se ubicaron las AGEB de donde provienen los informantes de la escuela secundaria tomada como muestra para el estudio. Se tomaron como referencia un conjunto de 13 AGEB (Figura 9) en el entorno de la escuela donde se aplicó el instrumento IDL.

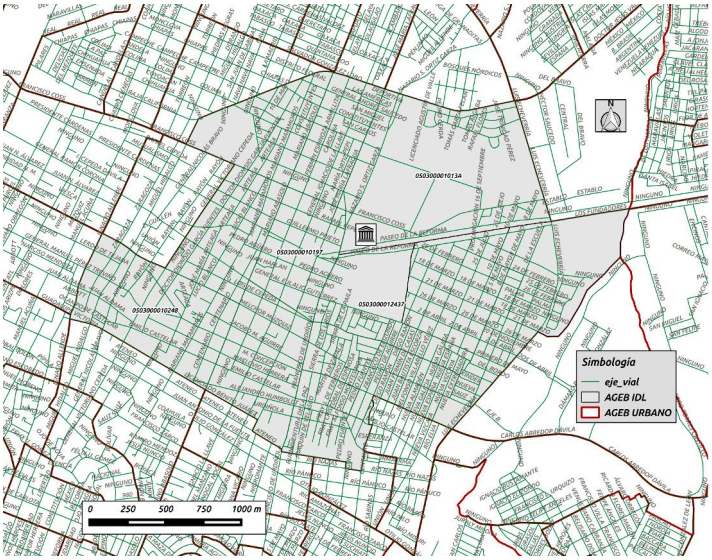


Figura 9. AGEB urbano en el entorno de la escuela piloto

A partir de datos del Censo 2010 de Inegi (2010), se generan las cifras de las Tablas 3 y 4. Este conjunto se conforma de 13 AGEB con una población de 29.446 personas. Se incluyen los siguientes campos: CVEGEO: clave de referencia; POB1: población total; POB7: población de 12 a 14 años; POB20: población de 15 años o más; EDU13: población de 12 a 14 años que asiste a la escuela; EDU15: población de 15 años y más alfabeta; EDU25: población de 15 años y más alfabeta; EDU31: Población masculina de 15 años y más sin escolaridad; EDU34: Población de 15 años y más con educación básica incompleta; EDU46: Población de 25 años y más con al menos un grado aprobado en educación superior; ECO26: población femenina desocupada; SCONY2: población femenina soltera o nunca unida de 12 años y más; SCONY7: población que estuvo casada o unida de 12 años y más, SCONY2_R: porcentaje de población femenina soltera o nunca unida de 12 años y más; SCONY8_R: porcentaje de población femenina que estuvo casada o unida de 12 años y más; HOGAR2: hogares censales con jefatura femenina; HOGAR13: hogares censales con jefa(e) de 30 a 59 años; HOGAR14_R: porcentaje de hogares censales con jefa de 30 a 59 años

Tabla 3. Cifras del Censo 2010 de AGEB en el entorno de escuela piloto (parte uno)

CVEGEO	POB1	POB7	POB20	EDU13	EDU25	EDU31	EDU34	EDU46	ECO26
0503000010013A	902	33	763	32	755	7	140	336	5
0503000010182	2209	109	1752	106	1719	26	396	502	13
0503000010197	1612	60	1303	55	1275	25	286	377	9
0503000010214	3114	152	2428	142	2341	72	685	416	9
0503000010248	2925	125	2349	119	2297	41	442	831	30
0503000010252	3445	168	2739	164	2670	57	666	771	18
0503000012371	1714	78	1384	74	1356	23	351	346	8
0503000012386	2028	94	1627	94	1586	45	500	354	9
0503000012437	2435	108	1926	103	1838	61	536	448	19
0503000012441	2237	128	1695	120	1569	100	578	212	7
0503000012456	632	20	499	20	472	15	173	55	-6
0503000012494	4138	173	3234	163	3158	67	766	985	49
0503000012526	2055	101	1629	100	1605	20	213	661	18

Nota: se filtraron los 13 AGEB aledaños

Tabla 4. Cifras del Censo 2010 de AGEB en el entorno de escuela piloto (parte dos)

CVEGEO	SCONY2	SCONY7	SCONY2_R	SCONY8_R	HOGAR2	HOGAR13	HOGAR14_R
0503000010013A	139	102	33.2	17.7	81	150	42
0503000010182	373	279	37.9	20	214	340	42.5
0503000010197	287	207	37.7	20.6	162	229	42
0503000010214	449	390	33.7	20.3	272	490	48.5
0503000010248	522	428	39.1	21.7	340	502	44.1
0503000010252	577	499	37.2	21.8	327	508	37.3
0503000012371	266	240	35.2	22.1	178	249	37.6
0503000012386	352	276	38.2	20.6	189	289	40.7
0503000012437	379	276	34.8	18.6	191	313	35.1
0503000012441	339	208	35.3	15	156	314	53.8
0503000012456	88	53	31.4	11.8	42	94	54.8
0503000012494	590	484	33.1	19.9	348	548	42.8
0503000012526	335	181	36.9	15.6	164	373	56.1

Nota: se filtraron los 13 AGEB aledaños

Los campos seleccionados están estrechamente relacionados con los centros de interés utilizados en este estudio, temas sobre los cuales los informantes respondieron

a las encuestas. Para el proceso de investigación, fue relevante saber la percepción que tiene el estudiante de la ciudad, las características de la casa familiar, el entorno de profesiones-estudio en el que se mueve y el grado de autoestima-seguridad. La justificación de elección de variables y su relevancia para la toma de decisiones en la escuela se ejemplifica en la Tabla 5.

Tabla 5. Selección de variables y su valor para la toma de decisiones en el aula a partir de los Centros de Interés, los datos censales y la realidad contextual del estudiante

Variables	Pertinencia	Aplicación educación
EDU13: población de 12 a 14 años que asiste a la escuela.	Los informantes están integrados a este grupo.	116 informantes insertos en una población de 1292 que actualmente estudian, lo que permitiría acompañamiento de pares en actividades educativas.
EDU15: Población de 15 años y más alfabeta. EDU31: población de 15 años y más sin escolaridad. EDU34: población de 15 años y más con educación básica incompleta.	Los potenciales padres de familia, abuelos, familiares adultos de los informantes.	Aunque el contexto de alfabetismo de la población que conforma la muestra es del 97.06%, hay 7192 potenciales padres sin escolaridad alguna y 5732 sin educación básica completa. La calidad del acompañamiento en tareas escolares se permea a partir de la preparación parental.
HOGAR14_R: porcentaje de hogares censales con jefa de 30 a 59 años.	El ambiente familiar en donde está inserto el estudiante.	Los AGEB circundantes arrojan datos censales de entre 35 y 56 por ciento de hogares con jefatura femenina que, en el ámbito educativo, tiene implicaciones de manejo de la información, tipo de tareas para hacer en casa, responsabilidades familiares de los adolescentes, emocionalidad del adolescente respecto de su rol como hijo y acompañamiento académico.

De la población total, 1349 tienen de 12 a 14 años, que corresponde al 4.58%. El 2.94% de esa población de 15 años o más es analfabeta, y el 4.33% de la población tiene de 12 a 14 años. Dado que el conjunto de AGEB se inserta en un entorno urbano con perfil socioeconómico de clase media, el porcentaje de analfabetas de adolescentes se puede considerar alto; más aún, algunos de los AGEB corresponden a la zona centro.

La toma de decisiones en el aula (aunque no es motivo de este artículo) estaría estrechamente ligada a la comprensión del contexto generalizado y particular en donde se desenvuelve el estudiante para buscar que cada uno alcance, según sus posibilidades, los aprendizajes esperados y el buen desenvolvimiento en la práctica social. Las estrategias

docentes, opacadas por el aparato público, son en realidad los detonantes para el alcance de aprendizajes esperados y el acrecentamiento en la prospectiva de movilidad social.

6. Conclusiones

Conforme se avanzó en el proceso de los datos, mediado por R, se encontró altamente satisfactorio el alcance del objetivo planteado respecto a la eficiencia en la velocidad de procesar los índices, la dimensión de datos que permitió manejar y la flexibilidad para generalizar el tipo de procesos. Se encontró también que la propuesta se apega al principio de investigación reproducible, porque es viable procesar nuevas bases de datos sin modificar el código. A lo anterior, se agregan las posibilidades de cálculos diversos derivados de los índices (se estudian ya para publicarse en el futuro mediato), la variedad de representaciones gráficas, así como la disponibilidad del *software* libre.

La aportación de este estudio, cuyo avance aquí se muestra, es valiosa si se considera como un antecedente del doble propósito para la transición en el uso de herramientas tecnológicas tradicionales y, en consecuencia, la propuesta de trabajo correlacional o interdisciplinario que permita aprovechar tanto la labor del lingüista y sus corpus, como la del estadístico y sus habilidades en el manejo de herramientas que nos ocupa en este proyecto.

La investigación interdisciplinaria pasó de ser una posibilidad a una exigencia cuando entre sus propósitos se encuentra el impacto social. La especialización académica se atomiza con el paso del tiempo, pero las propuestas que requieren las diversas instituciones deberán ser integrales. En este ámbito se ubica la propuesta de trabajo compartido a partir del manejo de corpus lingüísticos.

Actualmente, no se concibe un estudio de cualquier área del conocimiento, con aspiraciones holísticas, sin incorporar un trabajo colaborativo entre diversos especialistas. Existe un estrecho vínculo entre disciplinas aparentemente desconectadas que, al integrarse en un trabajo conjunto, logran resultados más profundos y significativos.

El acompañamiento de las nuevas herramientas informáticas, emergentes o enlazadas, se vuelve una condición *sine qua non* por la vertiginosidad con la cual las sociedades avanzan en la producción de información y su disponibilidad. Si bien hace algunos años las herramientas informáticas disponibles eran limitadas en su desempeño o costo, el desarrollo en la actualidad ha planteado el reto de seleccionar la herramienta adecuada ante una gran diversidad de programas comerciales y libres; estos últimos han sido un parteaguas en lo que a cómputo científico se refiere.

El tratamiento estadístico que da pie a los estudios cualitativos, sus procesos interpretativos y las posibles decisiones pragmáticas se plantean como una metodología

que puede aportar con amplitud y profundidad en el ámbito educativo, no solo para alcanzar lo que prometen los fundamentos, sino para la toma de decisiones en el aula y la política pública. En el aula, el manejo y control de las emociones son parámetros transversales a la enseñanza de los aprendizajes esperados y ambos solo se pueden percibir desde el lenguaje y los posicionamientos del hablante, de ahí la relevancia y pertinencia de considerar los índices de disponibilidad léxica como datos de diagnóstico sistematizado para el trabajo docente.

Para la política pública, en México, la Secretaría de Educación Pública (2017) determinó la dimensión emocional como uno de los enfoques básicos para la educación obligatoria, de manera que la regulación del lenguaje desde la escuela se ha convertido en un eje determinante no solo para alcanzar el perfil de egreso, sino para la construcción de la ciudadanía, estatus cuyos rasgos implícitos pueden ser detectados desde los corpus lingüísticos.

La minería de datos es una oportunidad de capitalización de los corpus lingüísticos acumulados a lo largo de casi cincuenta años, cuyos diagnósticos han sido valiosos, mas no se ha trascendido a la toma de decisiones de política educativa (como se muestra en este estudio que es viable), a la correlación de datos para hacer un trabajo áulico apegado a la realidad de los estudiantes o al acercamiento para acrecentar la individualidad de los alumnos.

El uso de minería de datos es una alternativa para extraer información más rica que puede estar limitada, o no considerada, al aplicar instrumentos estandarizados. Si bien esta propuesta puede representar un reto estadístico por los procesos de su análisis, sí se muestra como una fuente potencial de información, útil en la toma de decisiones al usar corpus lingüísticos, y una oportunidad en la investigación reproducible y la tendencia emergente denominada ciencia abierta.

7. Referencias

Adler, J. (2009). *R in Nutshell.* O´Reilly.

Arriaga, R. (2003). Memorias del Encuentro sobre problemas para la enseñanza del español. En *Educación e involución de la complejidad lingüística* (pp. 33-46). UAZ.

Ayres, R. (1994). *Information, Entropy, and Progress.* AIP Press American Institute of Physics.

Bauman, Z. (2005). *Tiempos líquidos*. Tusquets.

Baayen, R. (2008). *Analyzing Linguistic Data.* Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Bolaños, S. (2015). La Lingüística de Corpus, perspectivas para la investigación lingüistica contemporánea. *Forma y Función*, *28*(1), 31-54. https://doi.org/10.15446/fyf.v28n1.51970

Bourdieu, P. (2011). *Capital cultural, escuela y espacio social*. Siglo XXI.

Conapo. (2019). *Informe sobre el consumo de drogas en México y su atención integral*. Conadic.

Cortez, G. (2016). *Una aplicación de la disponibilidad léxica*. UAZ.

Du Bois, P. (2009). *MySQL.* Addison-Wesley.

Echeverría, M., & Parada, C. (1990). *DispoLex. Programa de cómputo.* Universidad de la Concepción.

Graser, A. (2013). *Learning QGIS 2.0.* Packt Publishing.

Goffman, E. (1997). *La presentación de la persona en la vida cotidiana*. Amorrotu.

INEGI. (2010). *Censo 2010.* Instituto Nacional de Estadística, Geografía e Informática. http://www.beta.inegi.org.mx/proyectos/enchogares/especiales/intercensal/

Janert, P. (2010). *Data Analysis with Open Source Tools.* O'Reilly.

Kelleher, J., & Tierney, B. (2018). *Data science.* MIT Press. https://doi.org/10.7551/mitpress/11140.001.0001

Konishi, S., & Kitagawa, G. (2008). *Information Criteria and Statitical Modeling.* Springer. https://doi.org/10.1007/978-0-387-71887-3

Liu, B. (2015). *Sentiment Analysis. Mining Opinions, Sentiment, and Emotions.* Cambridge University Press. https://doi.org/10.1017/CBO9781139084789.003

López, H. (1996). Los estudios de disponibilidad léxica, pasado y presente. *Boletín de Filología*, *35*(1), 245-259.

López, J. (2003). *¿Qué te viene a la memoria?* UAZ.

López, J. (2008). *Estudio de disponibilidad léxica en 43 estudiantes de ELE.* Universidad de Nebrija.

Michea, R. (1953). Mots fréquents et mots disponibles. Un aspecto Nouveau de la statistique du language. *Les Langues Modernes, 47*(1), 338-344.

Moreno, F. (1995). Cálculo de disponibilidad léxica. El programa LexiDisp. *Lingüística*, *51*(2), 243-250.

Pacheco, C. (2016). Incidencia de la vairable «sexo» en la dsponibilidad léxica de estudiantes preuniversitarios en Pinar del Río, Cuba. *Íkala*, *22*(2), 237-253. https://doi.org/10.17533/udea.ikala.v22n02a05

Pérez, M. (2020). Análisis del léxico disponible del centro de interés del insulto en estudiantes de secundaria de San Luis Potosí, México. *Revista de Filología y Lingüística de la Universidad de Costa Rica, 46*(1), 261-278. https://doi.org/10.15517/rfl.v46i1.41164

Perkins, J. (2014). *Python 3 Text Processing with NLTK3 Cookbook.* Packt Publishing.

R Core Team. (2018). *Language and Enviroment for Statistical Computing.* R Foundation for Statistical Computing.

Reyes, D., & Flores, M. (2018). La movilidad social en los implícitos discursivos de estudiantes de secundaria en México. De la escuela pública a la privada. *Oxímora, 13*(2), 58-80. https://doi.org/10.1344/oxi.2018.i13.22341

Rojas, D. (2017). Metodología de análisis de disponibilidad léxica en alumnos de Pedagogía a través de la comparación jerárquica de lexicones. *Formación universitaria*, *10*(4), 3-14. https://doi.org/10.4067/S0718-50062017000400002

Russell, M. (2014). *Mining the Social Web.* O'Reilly.

Shmueli, G. (2018). *Data Mining for Bussines Analytics. Concepts, Techniques and Applications in R.* Wiley.

Silge, J., & Robinson, D. (2017). *Text Mining with R.* O'Reilly.

Torgo, L. (2011). *Mining with R.* Chapnam and Hall/CRC. https://doi.org/10.1201/b10328

Urbizagástegui, R., & Restrepo, C. (2011). La ley de Zipf y el punto de transición de Goffman en la indización automática. *Revista de Investigación Bibliotecnológica*, *25*(54), 25-32. https://doi.org/10.22201/iibi.0187358xp.2011.54.27482

NOTAS

1 La escuela secundaria como reguladora de los factores discursivos correlativos entre disponibilidad léxica y prospectiva de movilidad social, de Dalia Reyes Valdés, Universidad Autónoma de Nuevo León, México, Doctorado en Filosofía con acentuación en estudios de la Educación.

2 Propuesta conceptual desarrollada para esta investigación.

3 Centro de interés es una palabra o sintagma detonante de campos semánticos, a partir del cual el informante escribirá todas las palabras que relacione con este.

4 Los 16 centros de interés propuestos por Michea (1953, pp. 338-344) son los siguientes:

- Las partes del cuerpo
- La ropa: vestido y calzado
- La casa: el interior y sus partes
- Muebles y enseres domésticos
- Alimentos: comidas y bebidas
- Objetos colocados sobre la mesa
- La cocina y sus utensilios
- La escuela: muebles y útiles
- Electricidad y aire acondicionado
- La ciudad
- La naturaleza
- Medios de transporte
- Trabajos de campo y jardín, los animales
- Los animales
- Diversiones y deportes
- Profesiones y oficios