I+D Revista de Investigaciones
ISSN 2256-1676 / ISSN en línea 2539-519X
Volumen 21 Número 2 Julio-Diciembre de 2026 pp. 83-94

# Identificación de temáticas discursivas basada en métricas de similitud textual[1]

# Identification of discursive topics based on textual similarity

Antonio Reyes Pérez[2]

Artículo recibido el 28 de octubre de 2025; artículo aceptado el 14 de agosto de 2026



**Resumen**

Una característica del análisis lingüístico es la evidencia de rasgos prominentes de una comunidad de hablantes a partir de las observaciones que, apoyadas en la lingüística de campo, en la sociolingüística, en la etnolingüística o en la lingüística de corpus, resumen los elementos más representativos de tales comunidades. Estos rasgos, en su mayoría, están fundamentados en una representatividad de corte estadístico, la cual se sustenta en datos frecuenciales. Es decir, a mayor frecuencia de un rasgo, más representativo será, no solo del conjunto, sino de la generalidad. Es innegable que este tipo de representación ofrece un panorama global acerca de las generalidades lingüísticas de una comunidad; sin embargo, la frecuencia no es un absoluto en cuanto a sus particularidades lingüísticas; hay palabras cuya frecuencia puede resultar baja, pero eso no significa que sean menos representativas de cierto fenómeno. Con base en esta premisa, en este estudio se describe una metodología para la identificación de temáticas a partir del reconocimiento de patrones semántico-discursivos que no se sustentan, necesariamente, en la mayor frecuencia que las palabras tienen en un corpus. Para ello, se aplica un conjunto de métricas de similitud textual a los datos del Corpus del Habla de Baja California; la hipótesis es que este tipo de métricas permite identificar relaciones lingüísticas que no están dadas por la mera frecuencia, sino que son reflejo de comportamientos subyacentes que caracterizan los intereses discursivos de los hablantes. Como hallazgo principal, se destaca el reconocimiento de las temáticas discursivas más representativas de este conjunto de datos, las cuales responden a tópicos claramente diferenciables. Ello permite concluir que la similitud textual es eficiente para identificar temáticas explícitas, basadas en la mayor frecuencia de aparición, así como implícitas, sustentadas en relaciones semántico-discursivas más profundas.

**Palabras clave:** lenguaje hablado, lingüística informática, análisis lingüístico, discurso.

**Abstract**

A major outcome of linguistic research is the characterization of a community based on observations that, supported by different specialized areas, such as field linguistics, sociolinguistics, ethnolinguistics, and even corpus linguistics, summarize the most representative features of such communities. Most of these features are based on descriptive statistics, in which frequency plays a relevant role: The higher the frequency of a feature, the more representative it will be. In this regard, it is undeniable that any statistical description offers a global overview of the lexical generalities of a community; however, frequency is not an absolute in terms of its linguistic particularities. Words with low frequency may still be highly representative of a specific phenomenon. In this respect, this study addresses the identification of discursive topics from the recognition of semantic and discursive patterns, which not necessarily rely on high word frequency. To this end, a vectorial representation process is described. This is based on the application of textual similarity metrics to automatically recognize the underlying topics regarding the interviews of the Corpus del Habla de Baja California. Our hypothesis is that such metrics are suitable to identify linguistic relationships that

[1] Artículo de investigación con enfoque mixto que presenta resultados del proyecto de investigación "FONFIVE-UAQ. *Procesar y comprender información: generación de conocimiento para las Humanidades desde las Humanidades Digitales*, FLL-2024-104". Proyecto financiado por la Universidad Autónoma de Querétaro. Fecha de inicio: agosto 2024. Fecha de conclusión: enero 2026.

[2] Doctor en Informática de la Universidad Politécnica de Valencia. Integrante del Laboratorio de Estudios del Lenguaje de la Facultad de Lenguas y Letras de la Universidad Autónoma de Querétaro (Querétaro, México). Dirección: Cerro de las Campanas s/n, Col. Las Campanas, C.P. 76010, Santiago de Querétaro. PBX: +52 (442) 192-1200. ORCID ID: https://orcid.org/0000-0001-9560-1674. Correo electrónico institucional: antonio.reyesp@uaq.edu.mx. Rol de autoría según CRediT: Conceptualización, análisis formal, fondos, investigación, metodología, responsable técnico de proyecto, escritura, revisión y edición.

stem not merely from frequency, but from underlying behaviors reflecting the speakers' discursive interests. The study's main finding is the automatic identification of the most representative discursive topics, which are clearly discriminating each other. This leads to the conclusion that textual similarity is an effective tool for identifying both explicit topics, based on high frequency, as well as implicit ones, grounded in deeper semantic-discursive relationships.
**Keywords:** spoken language, computational linguistics, linguistic research, speech.

## Introducción

Una característica del análisis lingüístico es que debe sustentarse en evidencia de rasgos prominentes, basados en observaciones empíricas, sea de una comunidad de hablantes o sea de los fenómenos que los caracterizan, con el fin de describir sus lenguajes y sus formas de comunicación. Ejemplo de ello son los estudios sobre léxico, los cuales se han caracterizado por ofrecer un amplio panorama respecto de los rasgos que, a diferente nivel, constituyen a las comunidades lingüísticas. Por ejemplo, Duffé Montalván (2016) presenta una perspectiva general sobre diversas investigaciones que tienen como punto focal el léxico. Los trabajos ahí presentados proporcionan una visión actual, suscrita a la lengua española, en los que se abordan problemáticas relacionadas con disponibilidad, préstamos, colocaciones sintácticas, traducción o enfoques didácticos. Común a todos estos trabajos y, en sentido amplio, a toda investigación que se base o trate sobre el lenguaje humano, es el uso de corpus. Con base en los ejes metodológicos de la lingüística de campo, la sociolingüística, la etnolingüística y, por supuesto, la lingüística de corpus, se han compilado varios corpus, cuyo fin es presentar evidencia sólida que sirva de base para el desarrollo de investigaciones que abordan, entre diferentes aspectos, el léxico (considérense, por citar solo algunos proyectos recientes, las aplicaciones, los repositorios de datos y los corpus generados en proyectos tales como Antenas neológicas descrito por Adelstein y Boschiroli (2021), Proyecto panhispánico de disponibilidad léxica expuesto por López (2015) o el proyecto comandado por Moreno Fernández (2021) y que reúne a más de una veintena de equipos internacionales para crear un corpus para el estudio sociolingüístico del español de España y de América (PRESEEA).

Ahora bien, hablar de corpus implica, en la mayoría de las ocasiones, caracterizar y describir los fenómenos en términos estadísticos. La noción de frecuencia, por tanto, cobra una gran relevancia para determinar la representatividad y la generalización de las observaciones. Es decir, a mayor frecuencia de un elemento o rasgo observado, más representativo será, no solo del conjunto que especifica al fenómeno, sino del universo que lo constituye. En este sentido, es innegable que la representación de corte frecuencial aporta un panorama global acerca de las generalidades de una comunidad; sin embargo, la frecuencia no es un absoluto en cuanto a sus particularidades lingüísticas. Dado este contexto, en este trabajo se aborda la caracterización lingüística de Baja California, representada por las ciudades de Mexicali, Tijuana y Ensenada, a partir de la identificación de patrones semántico-discursivos, no necesariamente frecuenciales, en el Corpus del Habla de Baja California (CHBC). La pregunta de investigación que guía el trabajo consiste en saber si es posible reconocer información que dé cuenta de las características discursivas de los hablantes considerando representaciones textuales que no estén justificadas sólo en información frecuencial. Con base en esta pregunta, el objetivo que se plantea es aplicar un conjunto de métricas de similitud textual para identificar automáticamente las temáticas que subyacen al discurso de los hablantes que participaron en la creación del CHBC. En este sentido, no se trata solo de dar un recuento de las palabras que con mayor frecuencia aparecen en el corpus, sino de aplicar técnicas de *machine learning* o aprendizaje automático para reconocer patrones escondidos entre las palabras y las oraciones, de forma que, con base en las métricas que más adelante se explicarán, se extraigan relaciones semántico-discursivas intrínsecas que, desde una óptica meramente frecuencial, pueden pasar inadvertidas.

Subyace a este trabajo una aproximación basada en los principios metodológicos y las técnicas del Procesamiento del Lenguaje Natural (PLN), el cual, de acuerdo con Jurafsky y Martin (2007), se define a partir de la interacción entre disciplinas de diferentes campos, por ejemplo, la lingüística, las ciencias de la computación o la psicología, y cuyo fin es reproducir la transmisión de información entre un humano y una computadora de la forma más natural posible. Ello, según lo planteado por Hausser (1998), se logra a partir de modelar la producción de los hablantes y la interpretación de los oyentes, sin que para ello importe si la comunicación se da en el plano de lo escrito o de lo hablado, tal como lo resaltan Bird, Klein y Loper (2009). Por otro lado, Clark Fox y Lappin (2010) apuntan que este objetivo se ha logrado, en parte, debido al desarrollo de algoritmos, técnicas y herramientas que pueden recuperar el significado de una expresión a partir del contenido lingüístico de esta, lo cual, dada nuestra aproximación, es fundamental para reconocer patrones a partir de las relaciones que suponen un significado intrínseco común.

Ahora bien, con este trabajo se aborda una de las grandes limitaciones del trabajo con corpus que tiene que ver con el hecho de que un análisis frecuencial, si bien aporta evidencia cuantitativa respecto a qué palabras son las que más aparecen en un conjunto de datos, reduce el discurso a la mera presentación de datos aislados. Si se considera solo el elemento frecuencial, es común que se pierda información relevante para entender el discurso como un

sistema coherente de significados, puesto que las palabras no son tratadas como constituyentes de una unidad mayor, sino que son aisladas en pos de destacar aquello que numéricamente es más común. Asimismo, la información semántica tiende a perderse puesto que en el conteo de frecuencias difícilmente se considera la polisemia y la homonimia; más aún, el alance de la negación, repercutiendo así en una visión segmentada de lo que se busca comunicar en todo discurso. Por último, no se puede obviar que aquello que no se dice (o se dice poco) a veces es tan importante como aquello que sí. Eso significa que lo primero no se reflejará en un análisis de corte meramente frecuencial, reduciendo así las posibilidades de profundizar en un análisis holístico.

Este tipo de limitantes y algunas otras, tanto de carácter conceptual como metodológico, han sido descritas en distintos trabajos, por ejemplo, Molina y Sierra (2015) discuten algunas limitaciones sobre el uso del CREA y el CORDE y proponen un método de normalización de frecuencias por medio del cálculo de medias móviles con el fin de tener representaciones más realistas de los datos almacenados en estos recursos. Atar y Erdem (2019), por su parte, destacan que una de las principales desventajas de este tipo de aproximaciones es la pérdida del contexto y de la estructura discursiva, puesto que el lenguaje es un fenómeno social y no se puede describir solamente a partir de datos cuantitativos, disociándolos de su componente social que determinan, entre otras cosas, cuándo y dónde utilizar ciertas palabras y expresiones. En este mismo sentido, el trabajo de Bond (2025) destaca que las listas de frecuencias rara vez consideran el contexto, lo cual provoca que haya limitantes en relación con aprehender el significado de las palabras o sus matices. Finalmente, en relación con el análisis del discurso, Thornbury (2012) destaca que esta clase de análisis no necesariamente está en función de las relaciones semánticas que subyacen a una estructura discursiva, por lo cual, tampoco pueden [reconocerlas ni] explicarlas. Mautner (2019) profundiza en esta crítica al señalar que con aproximaciones de este tipo se corre el riesgo de crear generalizaciones muy ambiciosas respecto de los supuestos vínculos causales entre el nivel macro de las estructuras sociales y el nivel micro de las elecciones lingüísticas.

En síntesis, es común en los trabajos que sustentan sus hallazgos en aproximaciones basadas o guiadas por corpus que se reconozca, por un lado, la utilidad de herramientas para describir numéricamente aquello que es más frecuente e, hipotéticamente, más representativo; no obstante, también es habitual, por otro lado, que se reporten limitantes en relación con el alcance descriptivo y explicativo que supone basar por completo la argumentación de cualquier fenómeno lingüístico en comportamientos estrictamente frecuenciales. Debido a lo anterior, se han realizado propuestas para minar información de corpus considerando posturas inter y multidisciplinarias, combinaciones de técnicas o representaciones algorítmicas más complejas. En este sentido, las aproximaciones que retoman los supuestos teóricos y metodológicos del PLN han permitido hacer representaciones más robustas del contenido discursivo de un corpus. El trabajo que aquí se presenta, tal como se mencionó, parte de los avances reportados en investigaciones en el campo del PLN, en específico, de la aplicación de métricas de similitud textual, para contribuir a las investigaciones que abordan las limitaciones antes expuestas.

El resto del trabajo está organizado atendiendo a los siguientes puntos. En la sección de Fundamentos se reportan algunos de los trabajos que, desde un enfoque de PLN, abordan las limitantes de las representaciones frecuenciales; asimismo, se explican las métricas de similitud que se aplicaron para la identificación de las temáticas discursivas. En la sección de Metodología se detalla el corpus que se usó en los experimentos que se reportan, así como los procedimientos que se llevaron a cabo para identificar las temáticas discursivas. En Resultados y discusión se presentan los resultados y se discuten sus implicaciones y limitaciones. Finalmente, en la sección de Conclusiones, se aporta una síntesis del trabajo realizado, sus alcances y limitaciones; adicionalmente, se listan las líneas de trabajo futuro.

## Fundamentos

El supuesto pragmático de la similitud textual es el de reconocer si dos palabras, frases, oraciones o textos, tienen un significado equivalente, más allá de las características formales que pudiera haber entre estos. Los primeros trabajos usaban enfoques basados en cadenas, es decir, se buscaba identificar rasgos estructurales en las palabras para determinar qué tan semejantes podrían ser. La aplicación de la distancia de Levenshtein, que se ejemplifica más adelante, es un reflejo de estas primeras aproximaciones. Posteriormente, se implementaron técnicas como *Latent Semantic Analysis*, cuyo propósito es descubrir relaciones entre las palabras por medio de la generación de matrices que intentan preservar la información contextual; asimismo, se volvió una constante el uso de ontologías, tipo WordNet, para estructurar el léxico con base en las relaciones semánticas que las palabras tienen en su uso real. Las aproximaciones más actuales subyacen a las técnicas de *word embeddings* y *transformers*, las cuales hacen representaciones más complejas basadas en las propiedades distribucionales de las palabras. Estas técnicas han resultado ser muy eficientes; sin embargo, requieren una gran cantidad de datos para poder entrenarse.

Si bien en la literatura especializada hay una cantidad importante de trabajos relacionados con este tema, aquí se presentan sólo algunos de los trabajos cuyo enfoque está basado en corpus en español. Al respecto, SemEval

(https://semeval.github.io) ha sido uno de los foros más relevantes para la realización de talleres que impulsan, mediante diversas competencias, el análisis semántico para sistemas de PLN. El trabajo de Agirre, Banea, Cardie, Cer, Diab, Gonzalez-Agirre, Guo, Mihalcea, Rigau y Wiebe (2014) condensa el trabajo desarrollado por varios equipos que participaron en la tarea 10 de SemEval 2014 para determinar la similitud textual en datos en español, particularmente, en artículos de Wikipedia y en noticias extraídas de Google. Entre los hallazgos de la competencia se reporta un porcentaje de eficiencia del 80,76 % para identificar contenidos relacionados. Este porcentaje, señalan los autores, es mayor que el que se había obtenido en tareas semejantes realizadas con datos en inglés. En este mismo foro, el trabajo de Cer, Diab, Agirre, Lopez-Gaspio y Specia (2017) presenta un conjunto de datos para calcular la similitud textual en oraciones. Entre los resultados más destacables se cita la mejora que se logra cuando se aplican técnicas robustas, las cuales están en concordancia con los juicios de anotadores humanos respecto a la semejanza semántica. En un contexto diferente, el trabajo de Venegas (2006) aborda la problemática a través del uso de *Latent Semantic Analysis*. En su artículo, el autor aplica esta técnica a un conjunto de datos extraídos de artículos de investigación en ciencias biológicas y ciencias sociales, concluyendo que la técnica es útil para identificar relaciones de semejanza en los texto, principalmente, en los resúmenes y, en menor nivel, en las palabras clave que se aportan para categorizar cada texto. En una aproximación más reciente y dentro de un contexto cerrado, el trabajo de Campillos-Llanos., Terroba, Zakhir, Valverde-Mateos y Capllonch-Carrión (2022) describe la construcción de un corpus paralelo con textos científicos sobre medicina, por un lado, y sus versiones simplificadas, por otro, las cuales fueron pensadas para los pacientes no especialistas. Este corpus se usó para entrenar sistemas que calculen qué tan similar es el texto simplificado respecto del documento médico original. Para ello, los autores calculan la similitud coseno, obteniendo valores de acuerdo entre anotadores (*inter-annotator agreement*) a nivel oracional bastante aceptables.

A diferencia de los trabajos reportados, en este se utiliza un tipo de dato completamente diferente: entrevistas semiestructuradas obtenidas mediante trabajo de campo. Este dato, además de aportar evidencia respecto a características propias del lenguaje oral, muy distintas a las del lenguaje escrito, evidencia rasgos sociolingüísticos y culturales que son muy relevantes para la aproximación que más adelante se detalla.

### *Similitud textual*

De acuerdo con Manning y Schütze (1999), el uso de métodos estadísticos en el campo del PLN ha sido una constante para modelar los diferentes fenómenos que suceden en el lenguaje natural. Ahora bien, al hablar de estadística se implica en automático la presencia de datos, sean estos cualitativos o cuantitativos, y en grandes volúmenes o escasos (véase el trabajo de Lee (1997) como referencia clave al respecto). El punto, independientemente de lo expuesto por Lee, es la necesidad de contar con elementos que reflejen de forma efectiva la realidad que se pretende modelar, en este caso, el lenguaje humano. Los métodos, que pueden ser varios, solo explicitarán el conjunto de relaciones que subyacen a los datos y, con base en tales relaciones, se podrán construir, además de descripciones puntuales, patrones que expliquen cómo funciona la realidad que dichos datos reflejan.

Para el caso de la similitud textual, existen varios métodos de análisis, los cuales, *grosso modo*, difieren en el hecho de cómo se implementa a nivel algorítmico determinada métrica o métricas para reconocer que dos unidades son semejantes, con base, por ejemplo, en su contexto lingüístico. Para ejemplificar cómo funciona la similitud textual, considérense las palabras A (describir) y B (barrer) en un contexto lingüístico X. Mediante la aplicación de una métrica, como la distancia de Levenshtein, se puede determinar si B puede ser sustituta de A (y viceversa) en X y en X'. Ello se hace estimando en una matriz el costo que implica intercambiar A y B sin que el significado en X se vea afectado. Véanse los ejemplos en (1) y (2) para contextualizar la explicación.

(1) X = El sol describe/barre áreas iguales en el mismo tiempo.

(2) X' = Juan describió/barrió la historia que le pasó en la mañana.

En el caso de (1), tanto A como B pueden ser equivalentes dado que su similitud textual tenderá a ser alta. Ello se deberá a que en su matriz de semejanza el costo de intercambiar un ítem por otro es mínimo y, lo más importante, el que aparezca uno u otro no alterará el significado de X; sin embargo, en (2) sucede lo opuesto, la similitud disminuirá porque, a pesar de que el costo por intercambiar los mismos ítems en la matriz sea igualmente bajo, el significado de X' ahora sí se verá claramente alterado.

Este tipo de métricas se han aplicado en varias tareas relacionadas con el PLN y estudios afines. Por ejemplo, el trabajo de Salcedo, Zambrano y Rojas (2017) detalla cómo es que los autores aplicaron la misma métrica, pero en el contexto de estudios de léxico con el propósito de evaluar la disponibilidad léxica a partir de la comparación de lexicones.

### *Métodos de agrupamiento*

En este trabajo nos interesa aplicar métricas de similitud textual orientadas hacia el agrupamiento automático o *clustering*. En este sentido, los procesos de agrupamiento han sido ampliamente documentados en el PLN (Lee, 1997; Manning & Schütze, 1999) y se han creado varias

herramientas que implementan algoritmos de *clustering.* Al respecto, véase el trabajo de Kulkarni y Pedersen (2005) y el de Pedersen, Patwardhan y Michelizzi (2004). Los resultados de esta clase de algoritmos se han empleado en tareas tales como análisis de sentimientos, desambiguación del sentido de las palabras, detección automática de plagio, clasificación de textos o reconocimiento de autor, por citar algunos.

En este tipo de herramientas, es común que la semántica sea un componente recurrente, ya sea desde una perspectiva que intenta reconocer rasgos intrínsecos asociados con el significado (como en el caso de describir/barrer), o bien, que los algoritmos se apoyen de herramientas que integran información semántica para realizar el cómputo de la similitud considerando, por ejemplo, relaciones del tipo hiperónimo-hipónimo, tal como se estructura en la ontología *WordNet,* descrita por Miller (1995).

### *Word embeddings*

En años recientes, los métodos de agrupamiento han incorporado técnicas de aprendizaje más eficientes con el fin de reconocer rasgos intrínsecos en los datos y generar mejores modelos de representación. Una de estas técnicas es la de *word embeddings* o incrustación de palabras. De acuerdo con lo expuesto en los trabajos de Mikolov, Chen, Corrado y Dean (2013) y Mikolov, Sutskever, Chen, Corrado y Dean (2013), estos modelos de representación lingüística operan a partir del agrupamiento de palabras que, dentro de un espacio vectorial, presentan comportamientos similares basados en relaciones de co-ocurrencia. Tales relaciones, según Bakarov (2018), permiten predecir de forma eficiente propiedades sintácticas, semánticas y discursivas entre las palabras de un corpus. Así, por ejemplo, si se considera que tanto describir como barrer son intercambiables en un contexto que implica movimiento, tal como se aprecia en el ejemplo (1), entonces los *embeddings* podrían evidenciar que además de las palabras sol, área o tiempo, hay otras palabras, tales como planeta, órbita, punto o Tierra que, dada su representación en el espacio vectorial, pueden tener similitud con las primeras.

Ahora, considerando datos del corpus que trabajamos y que será descrito en la siguiente sección, se puede observar en la Figura 1 que las palabras que están más cercanas a Mexicali en el espacio vectorial son formas relacionadas con locativos, tales como Tijuana, Ensenada, Sonora, ahí, etc. Es decir, mediante los *embeddings* se está capturando información que semánticamente manifiesta un vínculo discursivo común: cuando los hablantes se expresan sobre ciudades, según los datos del corpus, es común que lo hagan sobre alguno de estos locativos e, incluso, haciendo uso de deícticos (ahí) o entidades nombradas (UVM, sigla de Universidad del Valle de México). Esto significa, dentro del contexto de este trabajo, que se pueden identificar categorías discursivas a partir de ubicar las palabras cuyos puntos convergen en el mismo espacio vectorial, sin que en ello afecte el hecho de tener una frecuencia de aparición baja (UVM, por ejemplo, solo tuvo dos apariciones en el corpus).

A partir de lo anterior, se puede reconocer una naturaleza semántica subyacente a los *embeddings*, la cual identifica propiedades distribucionales o de co-ocurrencia como bases para determinar la semejanza de significado entre las palabras. Esto ha hecho que este tipo de técnicas sean aplicadas en la generación de modelos fonéticos, sintácticos y semánticos con buenos resultados. Asimismo, se han desarrollado varias implementaciones de *word embeddings* que se encuentran disponibles para entrenar modelos en diferentes lenguas, por ejemplo, *Word2Vec*, *FastText*, o *GloVe*. En este trabajo, se adopta la implementación de *Word2Vec*, tal como se describe en el trabajo de Mikolov *et al*. (2013).

```
In [16]: 1 print(model1.wv.most_similar (positive=['mexicali']))

[('ensenada', 0.8882536292076111), ('tijuana', 0.8627590537071228), ('sonora', 0.7877705097198486), ('tecate', 0.75
15378594398499), ('guadalajara', 0.7309742569923401), ('aquí', 0.7271942496299744), ('michoacán', 0.713373661041259
8), ('contabilidad', 0.7126027941703796), ('irapuato', 0.706836998462677), ('uvm', 0.7031515836715698)]
```

**Figura 1**

*Ejemplo de los embeddings identificados para la palabra Mexicali*

*Fuente*: autor.

## Metodología

A continuación, se describe el corpus que se utilizó en los experimentos que aquí se reportan. Asimismo, se detalla el procesamiento de los datos del corpus y se explica el método de representación vectorial que se llevó a cabo para identificar las temáticas discursivas a partir de la aplicación de la técnica de *word embeddings*.

### *El Corpus del Habla de Baja California*

El corpus que a continuación se describe recoge muestras de una de las regiones más jóvenes, en términos de estudios lingüísticos, de México: la región noroeste. Esta región ha sido descrita desde un punto de vista muy general, por ejemplo, en el trabajo de Henríquez Ureña

(1977) se habla de una sola variante que concentraba toda el habla de la zona norte del país, la cual, posteriormente, se subdividió en variantes y subvariantes específicas. Al respecto, con el Atlas Lingüístico de México de Lope Blanch (1970) se restringió la región a los estados de Sinaloa, Sonora, Baja California Sur y Baja California, generando con ello, trabajos más focalizados como los de Moreno de Alba (1994) o Mendoza (2004, 2006), en los que se puntualizaron características lingüísticas particulares de esta zona, así como líneas de trabajo que, en la actualidad, se orientan hacia la creación de recursos específicos para el estudio de rasgos más particulares y desde diferentes perspectivas. Uno de estos recursos es el CHBC, el cual, de acuerdo con Saldívar, Rábago y Lozano (2017), se compiló con la intención de ser representativo de uno de los estados que integran la región noroeste de México: Baja California.

***Recopilación de datos***

El CHBC está constituido por datos recabados mediante entrevistas aplicadas a hablantes de las ciudades de Mexicali, Tijuana, Tecate, Rosarito y Ensenada. Para este trabajo, solo se seleccionaron los datos de las tres ciudades más pobladas de la región y que aportan mayor número de entrevistas al corpus; así, quedaron fuera del estudio las entrevistas de las ciudades de Tecate y Rosarito.

Todas las entrevistas que integran el corpus fueron grabadas con base en la metodología del proyecto PRESEEA (Moreno, 2021) durante los años 2017 a 2019. Los datos, por lo tanto, están clasificados a partir de las variables sociales género, edad y nivel de estudios (Saldívar *et al*., 2017). Cabe destacar que, dada la metodología citada, las entrevistas procuran reflejar la forma más natural de expresión de los hablantes.

En cuanto al procesamiento de las entrevistas, estas fueron editadas con el programa *Audacity* para eliminar el ruido y transformar los audios a formato monoaural con el fin de simplificar, en lo posible, el proceso de transcripción. Este último proceso se realizó con el software *Praat*, descrito en Boersma (2014), el cual permite incorporar una interfaz que combina el procesador de textos con el audio a través de espectrogramas. Una vez transcritos, los datos del corpus fueron etiquetados a nivel categoría gramatical con el etiquetador de *Stanford* detallado por Toutanova, Klein, Manning y Singer (2003), el cual determina las categorías gramaticales por medio del cálculo de entropía máxima, alcanzando, según el reporte de los autores, una precisión superior al 95 %. El corpus, en formato textual y etiquetado, se encuentra disponible en el sitio: http://www.corpus.unam.mx/geco/portal/render/chbc. Adicionalmente, las transcripciones y los audios de las ciudades de Mexicali y Tijuana están disponibles en la plataforma del proyecto PRESEEA: https://preseea.uah.es/index.php/corpus-preseea. Por último, para contextualizar el alcance de los resultados, se reporta el tamaño del corpus: 8641 palabras tipo (*types*), 227,593 palabras con repeticiones (*tokens*) y 4157 hápax.

***Procesamiento de datos***

Tal como se mencionó en los párrafos previos, para este trabajo solo se utilizaron los datos de las ciudades de Mexicali, Tijuana y Ensenada. Cada uno de los textos de tales ciudades se preprocesó de forma automática con un script en *AWK* (Robbins, 2001) para convertir los textos a minúsculas y, sobre todo, para eliminar las etiquetas de origen, tales como <risas>, <tiempo>, <respiración> y demás que, para los fines de este trabajo, no aportan contenido que permita identificar la similitud semántica, antes bien, puede generar ruido debido a que el algoritmo consideraría estas etiquetas como parte del contenido léxico expresado por los participantes. Asimismo, se eliminó la información relativa a los turnos de habla del entrevistador, identificados con la etiqueta <E>, con el fin de centrar el análisis solo en el contenido discursivo de los informantes, identificados con la etiqueta <I>. En la Figura 2 se puede contrastar un archivo original (izquierda) y el resultado del preproceso (derecha).

Como fase final, todos los textos se guardaron especificando el nombre de la ciudad como parte del nombre del archivo.

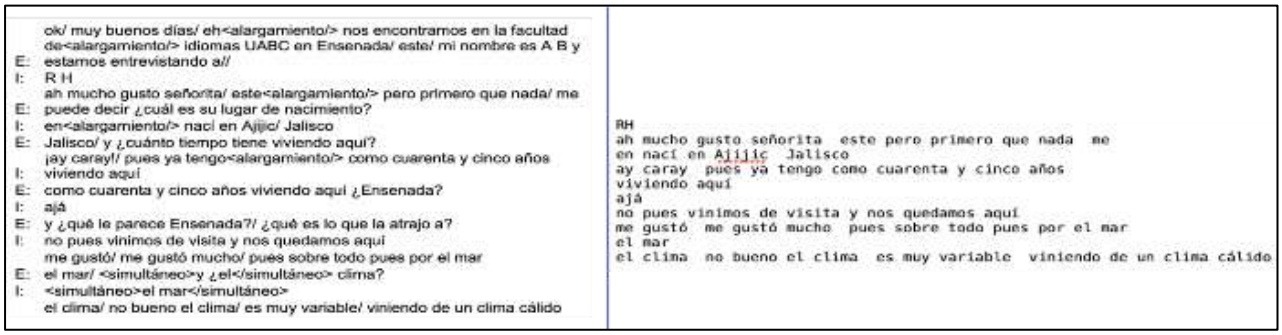


**Figura 2**

*Ejemplo del resultado del preprocesamiento de los textos*

*Fuente*: autor.

***Lematización***

El siguiente proceso que se realizó fue el de lematizar todos los textos con el fin de condensar la representación vectorial en núcleos semánticos y favorecer una distribución más homogénea de los datos, sin importar si alguna palabra tuviera baja frecuencia o fuera un *hápax*. Así, por ejemplo, hubo palabras como hijo, hija, hijitos, hijitas, hijastro que se agruparon en una forma base (*hijo*), la cual apunta a un contenido semántico común (familia) y que, al momento de hacer su representación vectorial, el hecho de que alguna de ellas, por ejemplo, hijastro, tuviera una frecuencia mínima en el corpus, no impactara de manera negativa puesto que con este proceso se mantenía un vínculo con el resto de palabras lematizadas. En la Figura 3 se aporta un ejemplo para mostrar el proceso de lematización.

| Forma | Frec. | Tipos |
|---|---|---|
| vivir | 372 | ver |
| llegar | 362 | ver |
| día | 357 | nom |
| poco | 335 | adj |
| hijo | 333 | nom |
| familia | 328 | nom |
| saber | 323 | ver |
| llamar | 317 | ver |
| mmm | 316 | nr |
| hermano | 312 | nom |
| poner | 311 | ver |
| siempre | 311 | adv |
| salir | 310 | ver |
| estudiar | 299 | ver |
| gente | 299 | nom |
| llevar | 295 | ver |
| verdad | 295 | nom |
| uno | 294 | nom |

```
llegó : 63
llega : 39
llegar : 33
llego : 33
llegué : 31
llegaba : 28
llegamos : 27
llegan : 17
llegas : 15
llegado : 15
llegaron : 13
llegando : 11
llegaban : 11
llegue : 8
llegaste : 8
llegara : 4
llegábamos : 3
llegaran : 2
llegaría : 1
```

**Figura 3**

*Ejemplo del proceso de lematización del verbo llegar*

*Fuente*: autor.

Finalmente, para evitar que durante la lematización se unieran palabras que, aunque semánticamente tuvieran relación (viaje – viajado), pero pertenecieran a categorías gramaticales diferentes, se realizó un proceso de anotación *POS* o de categoría gramatical.

***Agrupamientos***

Una vez que se procesaron los textos del corpus conforme a lo descrito anteriormente, se aplicó el método de Reniert (1986) para agrupar aquellos textos cuya estructura distribucional, basada en la representación vectorial de las palabras, fuera lo más semejante posible. Para ello, primero se extrajeron las palabras cuya categoría gramatical fuera verbo, sustantivo, adjetivo, adverbio, pronombre e interjección, y que tuvieran una distribución similar. Este proceso permitía tener una idea general de aquellas formas que semánticamente eran más relevantes dada su estructura distribucional, la cual, según nuestra aproximación, empezaba a denotar las temáticas que subyacen a los textos del corpus.

En la Figura 4 se muestra una nube de palabras que representa a estas formas.


**Figura 4**

*Palabras cuya distribución es vectorialmente más similar. A mayor tamaño, mayor frecuencia de aparición en el corpus.*

*Fuente*: autor.

Posteriormente, para entender de manera más clara cuáles eran las temáticas más relevantes en lo particular y no solo en lo general, se hizo una clasificación automática de los textos con el fin de reconocer dichas temáticas con base en las particularidades discursivas de cada ciudad. Para este proceso, cobró importancia el hecho de haber guardado cada texto conservando el nombre de la ciudad. En la Figura 5 se presentan algunas de las palabras más representativas para cada ciudad, en tanto que en la Figura 6 se muestra cómo el conjunto de palabras más similares se distribuye, traslapa o particulariza entre las tres ciudades.

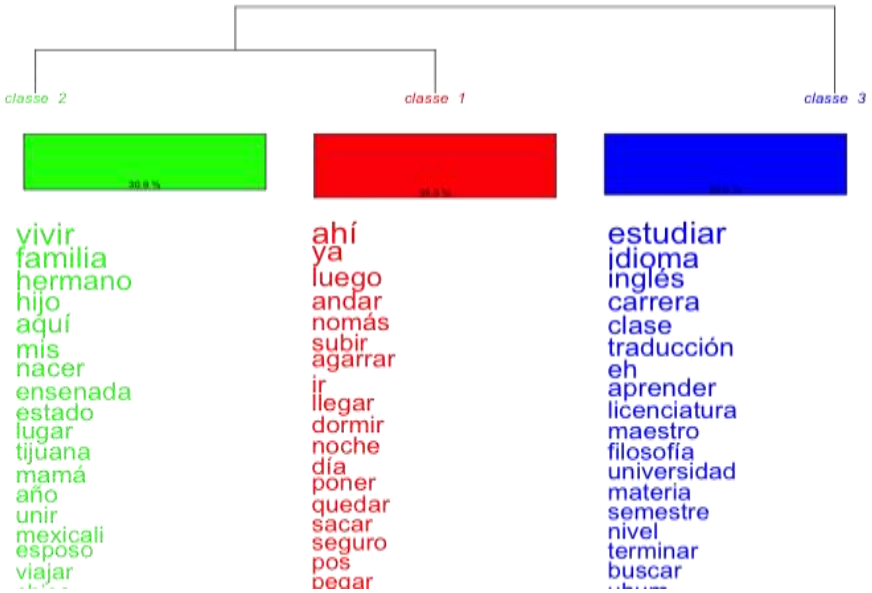


**Figura 5**

*Ejemplo de las palabras más representativas por ciudad: verde-Ensenada, rojo-Mexicali, azul-Tijuana*

Fuente: autor.

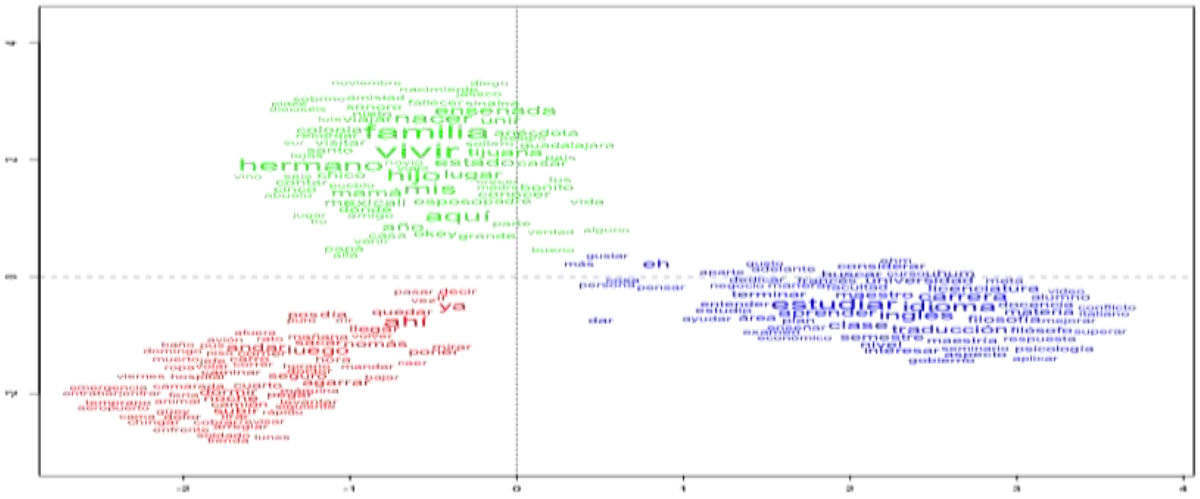


**Figura 6**

*Distribución del conjunto de palabras más similares por ciudad: verde-Ensenada, rojo-Mexicali, azul-Tijuana*

*Fuente*: autor.

**Resultados y discusión**

El interés de este trabajo consiste en saber de qué habla la gente, es decir, sus temáticas discursivas, por medio de la convergencia distribucional que las palabras tienen un mismo espacio vectorial. Esta convergencia está dada por la similitud semántico-discursiva que se asume tienen las palabras cuando las usamos para referirnos a diferentes tópicos. En este sentido, tal como se estipuló, la aproximación no basa la identificación de las temáticas en información meramente frecuencial, sino en identificar rasgos intrínsecos que, a nivel semántico, denoten cierta similitud discursiva a partir de la aplicación del método de Reniert.

La información que se presenta en las Figuras 4 a 6 busca reflejar cómo es que se construye el proceso de identificación de temáticas. Por un lado, la Figura 4 aporta una representación de las palabras que vectorialmente son más similares dentro del corpus, pero lo más importante es notar que las frecuencias no son el rasgo por el cual se determina la similitud. Al contrario, se les asume como anclas que permiten reconocer plantillas temáticas, las cuales se van a formalizar con diferente léxico. Así, una palabra con frecuencia alta (en fuente más grande en la figura) determina la plantilla temática, la cual va a ser llenada con diferentes palabras (porque, a pesar de tener intereses semejantes, todos los expresamos de forma diferente) que tienden hacia un contenido discursivo subyacente común, no importa, si como en el ejemplo de la Figura 1, alguna de esas palabras tiene una frecuencia muy baja.

Por otro lado, la Figura 5 evidencia cuáles son las palabras más representativas en función de cada ciudad. Se puede observar que las palabras que caracterizan las temáticas de estas tres ciudades tienden hacia tópicos específicos. Por ejemplo, para el caso de Ensenada, el conjunto de palabras apunta hacia una temática muy clara: familia (nacer, vivir, hermano, hijo, esposo, mamá). Para el caso de Tijuana, las palabras denotan una temática diferente: estudio (idioma, carrera, aprender, semestre, universidad). En el caso de Mexicali, no es evidente que haya una temática particular; sin embargo, si se observa que tanto esta ciudad como Ensenada comparten un mismo nodo (parte superior de la figura), se podría asumir cierta correlación discursiva, la cual debe probarse con trabajos futuros, aplicando métricas de similitud más discriminantes.

Ahora bien, es importante mencionar que estas palabras, aunque no son privativas de cada ciudad, sí son las que sobresalen en cuanto al contenido discursivo que los informantes del CHBC perfilaron a través de sus entrevistas. Este contenido se logra capturar de forma eficiente mediante la representación distribucional de los *embeddings*. En este sentido, en la Figura 6 se puede apreciar cómo las palabras tienden a ser más generales (distribución cercana a cero) o más particulares (distribución más lejana de cero) en el discurso de cada ciudad. Asimismo, en esta figura queda de manifiesto que la identificación de temáticas no se guía a partir de la frecuencia, sino de los rasgos discursivos intrínsecos que, tanto a nivel explícito (palabras con mayor frecuencia y fuente más grande en la figura) como implícito (palabras con menor frecuencia y fuente más pequeña en la figura), caracterizan los temas que les interesan a los bajacalifornianos.

Finalmente, es necesario señalar que mediante los procesos aquí descritos se ha mostrado que es posible

identificar elementos léxicos que, dadas las métricas y técnicas utilizadas, evidencian patrones discursivos que son intrínsecos a la forma en que los datos se distribuyen. Si bien cada una de las palabras identificadas puede ser un tema en sí misma, lo más importante es que en su conjunto, son representativas de aquello de lo que comúnmente le interesa y, por lo tanto, verbaliza, la gente de Mexicali, Tijuana y Ensenada. En este sentido, se ha aportado evidencia, basada en los comportamientos antes descritos, para alcanzar el objetivo planteado al inicio del trabajo: aplicar un conjunto de métricas de similitud textual para identificar automáticamente las temáticas discursivas en los datos del CHBC. Al respecto, si bien la aproximación descrita presenta limitaciones, la principal, tener un corpus relativamente pequeño en comparación con el tamaño de los corpus con que los que se entrenan de forma más eficiente los modelos basados en *word embeddigns* (Mikolov, Chen, Corrado y Dean (2013); Mikolov, Sutskever, Chen, Corrado y Dean (2013); Bakarov (2018)), los cuales están constituidos por millones de palabras, se ha demostrado que la aplicación de métricas más complejas, tales como las expuestas en torno a la similitud textual, son eficaces para hacer representaciones que complementan el alcance de la información frecuencial, mejorando la capacidad descriptiva y explicativa que se puede hacer sobre cualquier fenómeno de carácter lingüístico en un corpus.

## Conclusiones

En este trabajo se presentó una aproximación para reconocer las temáticas que subyacen a las entrevistas que conforman el CHBC. En particular, se aplicaron métricas de similitud textual y técnicas de aprendizaje automático para identificar el léxico representativo, tanto del corpus en lo general, como de las ciudades de Mexicali, Tijuana y Ensenada, en lo específico. Como hallazgo principal, se destaca el reconocimiento de las temáticas discursivas más representativas de este conjunto de datos, las cuales responden a tópicos claramente diferenciables. Estos tópicos evidenciaron las temáticas que son intrínsecas a los intereses de los hablantes de cada ciudad, las cuales manifiestan a través de sus discursos. Asimismo, se mostró que la información de corte frecuencial no fue determinante para la identificación de estos tópicos y, en consecuencia, de las temáticas discursivas. Ello permite concluir que la similitud textual es eficiente para identificar temáticas explícitas, basadas en la mayor frecuencia de aparición, así como implícitas, sustentadas en relaciones semántico-discursivas más profundas. Por último, se subraya que esta aproximación puede replicarse en otros corpus puesto que no es dependiente de la lengua o de un dominio específico.

Dados los alcances y limitaciones del trabajo, se plantea, como trabajo futuro, la aplicación de métricas más discriminantes con el fin de reconocer con mayor detalle la distribución del léxico y la forma en la que este se particulariza en cada ciudad. Asimismo, se busca recuperar las variables sociales que contiene el CHBC para expandir la identificación de temáticas y especificarlas en función del género, la edad y el nivel de estudios, lo cual, como se reportó al inicio del documento, es una limitante de los trabajos de corte cuantitativo. Finalmente, se busca extender el alcance de esta aproximación, haciendo una comparación de temáticas en las diferentes particiones que integran el corpus del proyecto PRESEEA con el fin de aumentar el tamaño de corpus y, sobre todo, tener un conjunto de datos más amplio sobre el que se puedan entrenar los *word embeddings* con mejores y más profundas estimaciones.

## Agradecimientos

Este trabajo fue realizado con el financiamiento otorgado por la Universidad Autónoma de Querétaro, en el marco del proyecto FONFIVE-UAQ, FLL-2024-104.

## Conflictos de interés

Se declara que no existe conflicto de interés de ninguna índole.

## Referencias

Adelstein, A. y Boschiroli, V. de los Á. (2021). Semantic Aspects of National Varieties of Spanish in a Dictionary of Neologisms, the Antenario. *International Journal of Lexicography*, *34*(3), 336–357. https://doi.org/10.1093/ijl/ecab010

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Perinan-Pascual, F., Mihalcea, R., y Wiebe, J. (2014). SemEval-2014 Task 10: Multilingual semantic textual similarity. En *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pp. 81–91. Association for Computational Linguistics. https://doi.org/10.3115/v1/S14-2010.

Atar, C. y Erdem, C. (2019). The advantages and disadvantages of corpus linguistics and conversation analysis in second language studies. Proceedings of IX Scientific and Practical Internet Conference of Young Scientists and Students, November. Ukraine.

Bakarov, A. (2018). A Survey of Word Embeddings Evaluation Methods. *CoRR*, *abs/1801.09536*. http://arxiv.org/abs/1801.09536.

Bird, S., Klein, E. y Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly.

Bond C. (2025). Corpus Linguistics as a Research Method in Nursing: A Practical Approach to Analysing Language Data. *Journal of advanced nursing*, 81(10), 6960–6967. https://doi.org/10.1111/jan.16659.

Boersma, P. (2014). The Use of Praat in Corpus Research. En U. G. and G. K. Jacques Durand (Ed.), *The*

*Oxford Handbook of Corpus Phonology*. Oxford University Press. https://doi.org/10.1093/OXFORDHB/9780199571932.013.016.

Campillos-Llanos, L., Terroba, A., Zakhir, S., Valverde-Mateos, A. y Capllonch-Carrión, A. (2022). Building a comparable corpus and a benchmark for Spanish medical text simplification. *Procesamiento del Lenguaje Natural*, (69), 189–196.

Cer, D., Diab, M., Agirre, E., Lopez-Gazpio, I., y Specia, L. (2017). SemEval-2017 Task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. En *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval 2017),* pp. 1–14. Association for Computational Linguistics. https://doi.org/10.18653/v1/S17-2001.

Clark, A., Fox, C. y Lappin, S. (2010). The Handbook of Computational Linguistics and Natural Language Processing. En *Blackwell Handbooks in Linguistics*. John Wiley & Sons.

Duffé Montalván, A. L. (2016). *Estudios sobre el léxico*. Peter Lang Verlag. https://www.peterlang.com/document/1053487.

Hausser, R. (1998). *Foundations of Computational Linguistics*. Springer-Verlag.

Henríquez Ureña, P. (1977). Observaciones sobre el español de América. En J.C. Ghiano (Comp.) *Observaciones sobre el español de América y otros estudios filológicos*. Academia Argentina de Letras.

Jurafsky, D. y Martin, J. (2007). *Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall.

Kulkarni, A. y Pedersen, T. (2005). SenseClusters: Unsupervised Clustering and Labeling of Similar Contexts. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 105–108.

Lee, L. (1997). Similarity-Based Approaches to Natural Language Processing. *arXiv*.

Lope Blanch, J. M. (1970). Las zonas dialectales de México. *Nueva Revista De Filología Hispánica (NRFH)*, 19(1), 1-11. https://doi.org/10.24201/nrfh.v19i1.436.

López, H. (2015). *Proyecto Panhispánico de disponibilidad léxica*. http://www.dispolex.com.

Manning, C., y Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Mautner, G. (2019). A research note on corpora and discourse: Points to ponder in research design. *Journal of Corpora and Discourse Studies*. 2(2). https://doi.org/10.18573/jcads.32

Mendoza, E. (2004). *Notas sobre el español del noroeste*. Culiacán: El Colegio de Sinaloa.

Mendoza, E. (2006). El español del noroeste mexicano. En Cestero, A., Molina, I., Paredes, F. (Eds.) *Estudios sociolingüísticos del español de España y América*. Madrid: Arco Libros.

Mikolov, T., Chen, K., Corrado, G. S. y Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ICLR*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. y Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 3111--3119.

Miller, G. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, *38*(11), 39–41.

Molina, C., y Sierra, G. (2015). Hacia una normalización de la frecuencia de los corpus CREA y CORDE. *Revista Signos. Estudios de Lingüística*, *48*(89), 307-331. http://dx.doi.org/10.4067/S0718-09342015000300002

Moreno Fernández, F. (2021). *Metodología del Proyecto para el estudio sociolingüístico del español de España y de América (PRESEEA)*. Universidad de Alcalá.

Pedersen, T., Patwardhan, S. y Michelizzi, J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. *Proceeding of the 9th National Conference on Artificial Intelligence (AAAI-04)*, 1024–1025.

Reinert, M. (1986). Un logiciel d'analyse lexicale. *Cahiers de l'analyse des données*, Tome 11 no. 4, pp. 471-481.

Robbins, Arnold. (2001). *Effective awk programming* (3°). O'Reilly.

Salcedo P., Zambrano C. y Rojas D. (2017). Metodología de Análisis de Disponibilidad Léxica en Alumnos de Pedagogía a través de la Comparación Jerárquica de Lexicones. *Formación Universitaria*.

Saldívar, R., Rábago, A. y Lozano, E. (2017). Corpus para el análisis de la variación lingüística: el diseño del corpus del habla de Baja California. En *Investigación y praxis contemporáneas en torno a las lenguas modernas*. Perales, M; Hernández, E. (Coord.). México: UQroo. pp. 250-264.

Thornbury, S. (2012). What can a corpus tell us about discourse? En A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* pp. 270–287). Routledge

Toutanova, K., Klein, D., Manning, C. D. y Singer, Y. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational*

*Linguistics on Human Language Technology - Volume 1*, 173–180. https://doi.org/10.3115/1073445.1073478.

Venegas, R. (2006). La similitud léxico-semántica en artículos de investigación científica en español: Una aproximación desde el Análisis Semántico Latente. *Revista Signos*, 39(60), 75–106. https://doi.org/10.4067/S0718-09342006000100004.