NUEVA REVISTA DE FILOLOGÍA HISPÁNICA

Nueva revista de filología hispánica
ISSN: 0185-0121
ISSN: 2448-6558
El Colegio de México A.C., Centro de Estudios Lingüísticos y Literarios

Ávila Muñoz, Antonio Manuel
Avance de una propuesta para el desarrollo de la tradición lexicoestadística hispánica: el índice de centralidad léxica
Nueva revista de filología hispánica, vol. LXXI, núm. 1, 2023, Enero-Junio, pp. 3-30
El Colegio de México A.C., Centro de Estudios Lingüísticos y Literarios

DOI: https://doi.org/10.24201/nrfh.v71i1.3838

Disponible en: https://www.redalyc.org/articulo.oa?id=60274089001

Nueva Revista de Filología Hispánica (*NRFH*), LXXI, 2023, núm. 1, 3-29
ISSN 0185-0121; e-ISSN 2448-6558; DOI: 10.24201/nrfh.v71i1.3838

AVANCE DE UNA PROPUESTA PARA EL DESARROLLO DE LA TRADICIÓN LEXICOESTADÍSTICA HISPÁNICA: EL ÍNDICE DE CENTRALIDAD LÉXICA

INITIAL PROPOSAL FOR THE DEVELOPMENT OF THE HISPANIC LEXICO-STATISTICAL TRADITION: THE LEXICAL CENTRALITY INDEX

Antonio Manuel Ávila Muñoz
Universidad de Málaga
amavila@uma.es
orcid: 0000-0002-5239-2670

Resumen: El propósito de este trabajo es presentar un nuevo índice estadístico para el estudio del léxico. El índice de centralidad léxica (ICL) surgió de una reflexión histórico-crítica constructiva a propósito de la tradición lexicoestadística hispánica, en particular de la naturaleza del índice de disponibilidad léxica que, en su momento, apareció como complemento del índice de frecuencia léxica. Esta revisión nos conduce a los orígenes de proyectos panhispánicos clásicos de estudio léxico cuya relevancia es bien conocida en el ámbito de la lexicoestadística. Las posibilidades de aplicación del ICL continúan esta línea y abren vías de estudio complementarias a las emprendidas en trabajos precedentes.

Palabras clave: índice de centralidad léxica; índice de disponibilidad; modelo teórico; proyectos panhispánicos de estudios léxicos; lexicoestadística.

Abstract: Our aim is to present a new statistical index for lexical studies. The lexical centrality index (LCI) stems from a constructive historical-critical revision of the Hispanic lexico-statistical tradition, centering especially on the nature of the lexical availability index which, when it appeared, served as a complement to the lexical frequency index. The revision goes back to the origins of traditional pan-Hispanic lexical projects, whose relevance is well known in the field of lexico-statistics. The LCI has possible applications which continue in this tradition and which open up ways of studying that complement those used in previous projects.

Keywords: lexical centrality index; availability index; lexical statistics; pan-Hispanic projects for lexical studies; theoretical model.

Recepción: 27 de julio de 2020; aceptación: 11 de mayo de 2021.

Introducción*

Trazar el desarrollo de la investigación lexicoestadística hispánica supone transitar a través de un considerable número de experiencias compartidas marcadas por algunos hitos históricos. Aunque cada una de ellas tiene sus características propias, todas comparten un factor común: la ilusión por conocer la organización léxica de nuestra lengua. El objetivo principal de este trabajo es presentar un novedoso índice estadístico para el análisis del léxico: el índice de centralidad léxica (ICL). Se trata de un factor que surgió de la reflexión en torno al conocido índice de disponibilidad léxica (IDL), el cual, a su vez, se originó hace décadas como complemento del índice de frecuencia léxica (IFL). Pero para entender plenamente la naturaleza del ICL es necesario situarlo en el amplio panorama de investigación léxica panhispánica. Por eso, este artículo se estructura según dos bloques principales: en el primero presentamos algunos proyectos panhispánicos sobre análisis estadísticos del léxico vertebrados a partir de índices matemáticos que facilitaron la coordinación de grupos de investigación por todo el mundo. Las limitaciones de espacio propias de trabajos como el que aquí ofrecemos imponen que no se tengan en cuenta todos los estudios precedentes, aunque consideramos que los que mostraremos líneas abajo son los que han tenido mayor recorrido cronológico y mayor trascendencia investigadora.

Puesto que el ICL surge de una profunda y constructiva reflexión crítica del IDL, en el segundo bloque repasamos los principales índices estadísticos que se han utilizado en la tradición lexicoestadística hispánica: el índice de densidad (riqueza) léxica, los índices de frecuencia, dispersión y uso, y, en especial, el índice de disponibilidad léxica. Trataremos de situar el análisis de la centralidad en la evolución de estos índices, en general, pero particularmente, desde luego, en el de disponibilidad. Asimismo, presentamos los conceptos básicos, las propuestas que los sustentan y las herramientas que ayudan a obtenerlos. Por medio de un ejemplo práctico mostraremos los beneficios que aporta el ICL al ámbito de estudio léxico.

* Este trabajo se ha realizado en el marco del Proyecto de Investigación *Observación del Pulso Social en Andalucía a través del Análisis Léxico (PULSO Andaluz). Primera fase* (UMA20-FEDERJA-013), financiado por el Programa Operativo FEDER 2014-2020 y por la Consejería de Economía y Conocimiento de la Junta de Andalucía.

En última instancia, nuestro trabajo pretende ser un reconocimiento sincero a los autores que con tanto esfuerzo y dedicación han logrado consolidar la *gran familia* de la lexicoestadística hispánica, como a menudo la consideraba el maestro López Morales.

El estudio estadístico del léxico de la lengua castellana

Las repercusiones de los avances de la lingüística de corpus en el desarrollo del estudio del léxico fueron constantes desde sus orígenes y particularmente significativas en la última década del siglo xx (Sinclair 1991). La utilización práctica inmediata más fácil y frecuente de un corpus es la realizada por expertos en lexicografía para determinar los artículos que debe contener un repertorio léxico, sus acepciones más importantes, las combinaciones semánticas y sintácticas que no pueden faltar, etc. (Alvar Ezquerra *et al.* 1994, pp. 10-11). En realidad, muy pronto se observó que los análisis estadísticos de los sistemas informáticos aplicados a corpus ofrecerían enormes posibilidades de cara al hallazgo de frecuencias relativas de las unidades léxicas para considerar o no, por ejemplo, su inclusión en diccionarios o manuales de enseñanza de lenguas. La lingüística de corpus facilitó el avance de la lexicografía computacional, gracias a que se relacionaron grandes bases de datos integradas por unidades léxicas con la reflexión teórica. A partir de este proceso relacional se potenciaron algunas hipótesis hasta entonces difícilmente demostrables por medio de la simple introspección erudita.

En la actualidad podemos ya hablar de una consolidada disciplina *lexicoestadística* a causa de la intervención generalizada de los modelos estadísticos en ámbitos de la descripción, el análisis y la construcción de modelos teóricos que se basan en corpus léxicos. En el ámbito hispánico, los estudios léxicos de la lengua castellana se suelen aglutinar en proyectos coordinados de gran amplitud, aunque también hay estudios particulares que usan sólidos índices estadísticos para elaborar listas de palabras con aplicaciones muy diversas.

Entre los primeros, han de citarse los *léxicos de la norma culta*, colección de índices de frecuencia con especificaciones de competencia léxica, del vocabulario conocido y utilizado por

el sociolecto alto de una comunidad dada. Uno de los trabajos más ambiciosos en este sentido es el "Proyecto de estudio del habla culta de las principales ciudades de habla hispana" —fundado en 1964 durante la celebración del II Simposio del Programa Interamericano de Lingüística y Enseñanza de Idiomas (PILEI) (Lope Blanch 1967)—, que pretende llevar a cabo un estudio de la fonología, la morfosintaxis y el léxico de la norma culta del español hablado en todas las ciudades que se han unido al macroproyecto. Los resultados del estudio sobre el léxico pueden consultarse en Lope Blanch (1977) y en los boletines informativos que publica la Asociación de Lingüística y Filología de América Latina (ALFAL), regente de tal proyecto, en cuyos diversos informes se han ido actualizando los datos pertinentes. Además, en cada una de las ciudades donde se han realizado las encuestas también se han publicado los resultados particulares: Bogotá, Buenos Aires, Caracas, Madrid, Sevilla, Granada, etcétera.

Otro importante proyecto de investigación que se desarrolla de manera oficial desde 1993 es el del "Español del mundo", presentado por primera vez en el X Congreso Internacional de la Asociación de Lingüística y Filología de América Latina, celebrado en la ciudad de Veracruz (México). Coordinado en sus orígenes por Hiroto Ueda y Toshihiro Takagaki, este ambicioso proyecto tiene como objetivo general conocer la situación actual del léxico español de todo el mundo mediante la elaboración de una red internacional de investigación. Se trata, en definitiva, de formar un archivo del léxico distintivo, con el propósito de obtener diversa y variada información que sea útil para la investigación. Los estudiosos interesados pueden, a partir de aquí, realizar sus trabajos dialectológicos, lexicográficos, etcétera.

En estrecha colaboración con este proyecto se encuentra el de "Difusión internacional del español por radio, televisión y prensa" (DIES-RTP), creado desde El Colegio de México por Raúl Ávila (1999). Ambos proyectos establecieron en su momento un convenio de colaboración y cooperación que permitió elaborar un archivo libre, abierto, relacional y flexible en continua renovación capaz de observar el ritmo frenético de los cambios lingüísticos tanto diacrónicos como diatópicos del mundo hispano.

Por cuestiones de espacio, el último de los trabajos colectivos que vamos a citar en este artículo es el "Proyecto panhispá-

nico de disponibilidad léxica" (PPHDL), cuya relación directa con el origen del índice de centralidad léxica recomienda una exposición algo más extensa. Coordinado por Humberto López Morales, el fin último del PPHDL es obtener un diccionario a partir de la elaboración de listas de disponibilidad léxica en diferentes lugares de habla hispana repartidos por todo el mundo. La importancia de este macroproyecto coordinado reside, por encima de cualquier otra consideración, en la propuesta de homogeneización y unificación de los criterios de diseño metodológico comunes, que permiten el intercambio de datos y el desarrollo de estudios comparados entre los resultados de los diferentes proyectos locales. Los distintos grupos adscritos al macroproyecto asumen directrices metodológicas que pueden sintetizarse en los siguientes puntos:

1) Se trabaja con estudiantes preuniversitarios, método que evita la contaminación "técnica" o específica de un área profesional determinada, a la vez que el estudio se centra en sujetos que se suponen suficientemente "maduros" desde el punto de vista léxico.

2) Los materiales se obtienen mediante pruebas asociativas a partir de centros de interés o núcleos temáticos sobre los que el informante debe aportar los elementos léxicos que considere relacionados.

3) Las listas son abiertas, ya que no se limita el número de palabras que puedan aportarse; en cambio, el informante debe escribir todas las posibilidades relacionadas con el asunto propuesto.

4) La única limitación en este sentido se refiere al tiempo de que disponen los individuos para elaborar sus listas; la condición es que no sobrepasen los dos minutos por centro de interés.

5) Una vez obtenidas las listas léxicas se editan los materiales según criterios consensuados de lematización; luego, las bases de datos se someten a un proceso de análisis estadístico que proporciona, en última instancia, las listas de disponibilidad léxica.

6) Puesto que la mayoría de los estudios cuenta con una base de datos sociológica complementaria donde se almacenan las características esenciales de los informantes estudiados, las correlaciones sociolingüísticas son frecuentes, y habituales los estudios de contraste entre las investigaciones particulares. Como es lógico, todo ello contribuye a enriquecer de manera muy significativa las posibilidades de los estudios de disponibilidad léxica.

En resumidas cuentas, todos los proyectos presentados hasta aquí —y otros de semejantes características y vocación panhispánica— han servido para recopilar extensas colecciones de datos léxicos. Una vez almacenados, suelen explotarse mediante la aplicación de procedimientos estadísticos comunes que ayudan a obtener índices matemáticos que permiten observar mejor la organización del léxico de nuestra lengua. En el siguiente apartado mostraremos los más conocidos, como preámbulo necesario de la presentación y explicación del índice de centralidad léxica.

Índices estadísticos tradicionales para el estudio del léxico

El índice de densidad (riqueza) léxica

La mayoría de los estudios interesados en medir la riqueza léxica individual recurre a un parámetro matemático llamado *índice de densidad léxica*. En esencia, este indicador se basa en promedios de recuentos y se utiliza, frecuentemente, con fines comparativos para medir la variedad léxica que contienen los textos. Este índice, en su concepción más elemental, se calcula mediante un sencillo procedimiento que consiste en dividir el corpus objeto de estudio en trozos iguales y en calcular el número de tipos léxicos distintos o vocablos en cada trozo; el índice de densidad léxica es el promedio de todos estos recuentos. Los textos sobre los que se aplica tal índice suelen manipularse artificialmente para conseguir que todos sean del mismo tamaño y favorecer así los procedimientos comparativos. En cualquier caso, este índice nos facilita una perspectiva de la diversidad léxica que contienen los textos, por lo que se ha considerado tradicionalmente como un excelente indicador de su naturaleza cualitativa. Por esta razón, se ha empleado de manera recurrente en el ámbito de la lingüística aplicada para medir la habilidad lingüística de aprendices tanto de segundas lenguas como de la lengua materna.

El procedimiento para el cálculo de la densidad léxica ha gozado de gran aceptación desde las primeras propuestas de Guiraud en 1960. El empleado por Baldi en 1972 difería ligeramente del utilizado por Raúl Ávila (1988), ante todo en el tipo de palabra considerado para realizar el recuento. En el contex-

to hispánico, también se refiere a este índice de densidad léxica López Morales en 1983. Por lo demás, Ávila Muñoz (2016) revisó recientemente el concepto y propuso un modelo para el cálculo de la estimación léxica de los hablantes compatible con el concepto tradicional de densidad léxica, usado frecuentemente para medir la riqueza léxica de textos escritos. A diferencia de los procedimientos tradicionales, la propuesta permitía acceder al tamaño virtual del vocabulario individual sin necesidad de recortar de un modo más o menos artificial el corpus de datos, al tiempo que se adaptaba con facilidad a las características propias de la dinámica discursiva conversacional.

El índice de frecuencia: teoría y principios

La frecuencia de uso de las unidades léxicas ha demostrado ser una de las variables de mayor peso en el procesamiento del lenguaje, sea desde una perspectiva de producción o de descodificación (Alameda y Cuetos 1995). Ellis (1985) puso de manifiesto la dificultad e incluso imposibilidad de producción lingüística cuando el emisor debe componer mensajes con unidades que le resultan desconocidas. Asimismo, Vega *et al.* (1990) demostraron que existe una proporción inversa, en el sentido de que cuanto menor sea la frecuencia de utilización de las palabras de determinado texto, mayor será la dificultad de comprensión para el receptor.

Hay diccionarios de frecuencias léxicas de las principales lenguas occidentales que se han aplicado de manera muy diversa en virtud de las nuevas tecnologías. Un índice de frecuencias del léxico de una lengua está formado por un listado de palabras al que se adjunta determinada frecuencia de aparición en un corpus considerado. La procedencia de esas palabras y el tratamiento estadístico aplicado variarán en función de los intereses que origine el trabajo. Se suele tener en cuenta el destino que quiere darse a la lista de frecuencias, aunque normalmente ésta se obtendrá por medio de la aplicación de fórmulas de frecuencia, dispersión y uso, referencias estadísticas que facilitan la elaboración de los vocabularios básicos, es decir, las aproximadamente 5 000 palabras de mayor uso en la comunidad estudiada.

Los diccionarios de frecuencias más desarrollados, o sea, aquellos en los que se emplean fórmulas estadísticas pondera-

doras como la desviación típica, no se limitan a presentar simples colecciones de palabras, sino a mostrar una selección de las que mediante los métodos estadísticos empleados resultan ser más estables y básicas de la muestra en cuestión. Por medio de procedimientos metodológicos muy concretos se obtiene un léxico que posee determinada estabilidad estadística, es decir, el léxico al que los hablantes recurren más a menudo para construir sus mensajes, independientemente del tema del discurso.

Índices de frecuencia, uso y dispersión. Los primeros diccionarios encargados de medir la frecuencia de aparición de los distintos elementos léxicos en los textos manejaban el concepto de *frecuencia* como base para realizar sus cálculos. En el caso de los diccionarios de frecuencia más antiguos del español, Rodríguez Bou (1952) utilizó la frecuencia ponderada, mientras que Keniston (1941), primero, y García Hoz (1953), después, se basaron en el *rango* como factor de cálculo básico.

La mera frecuencia, sin embargo, sea considerada de modo absoluto o relativo, no resulta una medida particularmente significativa, ya que podemos encontrar dos vocablos con la misma frecuencia, pero con distribución irregular entre los géneros establecidos, lo que significa que el vocablo está sujeto a ciertas variables tipológicas; en cambio, una distribución regular entre los géneros denota que la palabra es independiente de esas variables, y, por tanto, su peso y utilidad en el idioma debe ser mayor. Por esta razón, se buscaron cálculos cuyos índices estabilizaran la frecuencia en el texto considerado, y se llegó a los índices de *dispersión compleja* y *uso* ya utilizados en la serie *The Romance languages and their structures* (Juilland 1973), y en el considerado primer diccionario de frecuencias léxicas de la lengua española (Juilland & Chang-Rodríguez 1964).

Los listados obtenidos luego de utilizar el simple cómputo de frecuencia de aparición no pueden controlar las irregularidades que surgen de modo circunstancial en la elección de muestras escogidas al azar. Evidentemente, podría ocurrir que, de modo casual, se seleccionaran varios textos monográficos sobre un mismo tema, con lo cual el vocabulario correspondiente daría la sensación de aparecer con gran frecuencia; es por ello que este factor se ponderó rápidamente. El *índice de dispersión compleja* mide la distribución de la frecuencia de un modo muy sencillo: parte de la distribución del universo léxico considerado en tipos o mundos delimitados por el contenido o por las

condiciones formales de recogida de los textos. La función de la dispersión es determinar la estabilidad de la frecuencia en las categorías establecidas, lo que favorece que la tipología considerada se convierta en representativa y básica en la ponderación.

La fórmula definitiva propuesta por Juilland en 1973, tras diversas correcciones, se basa en el promedio de los desvíos que tiene la frecuencia de cada término en cada subfrecuencia respecto a la frecuencia teórica del término, con lo que se obtiene el coeficiente de variación de cada palabra en la muestra escogida, independientemente de la frecuencia. Después, otro cálculo introduce el factor del número de categorías consideradas, lo que permite obtener un índice que oscila entre 0 —dispersión pésima— y 1 —dispersión óptima. Variantes de esta fórmula han venido aplicándose a los distintos recuentos léxicos a partir de ese momento, a pesar de los varios errores en la presentación del tratamiento estadístico que han advertido algunos autores desde Müller (1965, p. 35).

Una vez conocido si el lexema se reparte o no uniformemente, es decir, si los hablantes —con independencia del registro utilizado, de la variable sociolingüística que lo identifica o del tema del discurso lingüístico— hacen uso uniforme o no de aquél, sólo resta obtener el listado de palabras que se ha de incluir en el diccionario básico de la lengua. Para ello contamos con un nuevo índice, el *factor de uso léxico*, cuya fórmula pondera al alza los repartos equitativos de frecuencias en detrimento de aquellas palabras que tienen mayor frecuencia, pero distribución desigual, de tal manera que el índice de uso se acercará más a la frecuencia a medida que se regularice la repartición de las unidades en las categorías escogidas (Müller 1977, pp. 68-76). La fórmula de uso léxico es el resultado de multiplicar la frecuencia por la dispersión.

La disponibilidad léxica

También hay otros tipos de análisis léxicos que buscan inventariar el vocabulario más usual en determinados contextos comunicativos. Tal es el caso de los *índices de disponibilidad* que, en cierta manera, vienen a suplir las deficiencias que se observan en las listas de vocabulario frecuente. Al examinar estos listados, podemos observar que ciertas palabras muy comunes y conocidas pueden faltar o alcanzar índices de frecuencia muy

bajos, y por ello quedar descartadas o relegadas del repertorio básico. Esta conclusión estadística puede en determinadas ocasiones conducirnos a error. Gougenheim (1967) llegó a la conclusión de que el hecho de que en el *Dictionnaire fondamental* (Gougenheim 1958) no aparezcan palabras de sobra usadas en ambientes francófonos, como *métro*, *lettre* o *timbre*, se debía fundamentalmente a cuestiones metodológicas que no estaban relacionadas con el desconocimiento de estos términos entre la comunidad de hablantes. Michéa (1953) inauguró la tradición de la elaboración de los léxicos disponibles: hasta su incursión en el mundo estadístico, la frecuencia era el único factor que se manejaba para establecer el orden de los vocablos en las listas.

Para compensar estas irregularidades, los *léxicos disponibles* pueden resultar el complemento adecuado por ser el reflejo del caudal léxico usado en una situación comunicativa concreta. Este tipo de listado explota el concepto de *situación frecuente* respecto al manejado en los listados que tienen por propósito registrar el *léxico frecuente* en cualquier situación. En realidad, el léxico de disponibilidad encuentra su sentido en la máxima de que ciertas palabras muy usadas en determinada lengua están estrechamente relacionadas con la aparición o no de temas concretos. De hecho, hay una baja probabilidad, a menos que surja un tema constreñido al ámbito teatral, de que se mencionen palabras como *camerino* o *telón* en situaciones comunicativas corrientes. Del mismo modo, como indica López Morales (1983, p. 213), es "poco probable… que salgan en nuestra expresión palabras como *carta* y *sello* si no nos referimos específicamente a asuntos del correo".

Para la recolección del vocabulario disponible, aunque no frecuente, se parte de la elaboración de unas pruebas asociativas que giran en torno a unos estímulos o *centros de interés*; éstos son muy variados y pueden abarcar temas tan dispares como el cuerpo humano, nuevas tecnologías, la prensa, la agricultura, sindicatos, finanzas, medioambiente, vestuario, los transportes, el circo, las profesiones, la ganadería, etc. Alrededor de estos núcleos temáticos surge determinado vocabulario relacionado que es el inventario del que el hablante realmente hace uso si una conversación, en momento dado, discurre por los cauces necesarios. El número total de términos que engrosan estos listados de léxico disponible está limitado, normalmente, bien por el tiempo de reacción de los hablantes, bien por el cierre

de listas tras determinada cantidad de palabras. Se supone que son más disponibles las palabras que acuden antes a la memoria, o sea, aquellas que aparecen en los primeros lugares de las listas. El léxico disponible forma parte del lexicón mental de los hablantes, pero no suele actualizarse en los intercambios lingüísticos, a menos que haya una especialización temática.

Las diversas aproximaciones cuantitativas orientadas a realizar clasificaciones del léxico disponible han tratado de poner de manifiesto que, al tocar determinados temas, las palabras que primero acuden a nuestra mente son más disponibles que aquellas otras que no hacen su aparición de forma inmediata. El índice de disponibilidad es, a grandes rasgos, un índice numérico que trata de poner en relación criterios de frecuencia y orden.

La necesidad de ordenar los términos más disponibles y de ofrecer listas basadas en cuantificaciones hizo que los precursores franceses de los estudios de disponibilidad léxica se propusieran fórmulas útiles para tales fines que poco a poco fueron mejorándose. No obstante, los estudios que se limitan a la utilización de fórmulas matemáticas para la obtención de listados de palabras resultan, desde un punto de vista estrictamente científico, algo simples. Como es lógico, luego de observar estas limitaciones, muchos autores han aprovechado las posibilidades que ofrecen los diversos programas diseñados para el cálculo de la disponibilidad (por ejemplo, *Lexidisp*) con el fin de ir un poco más lejos y de establecer correlaciones entre variables léxicas y sociales. Sin embargo, las restricciones que imponen estos programas al número de variables sociales que permiten manejar y, sobre todo, el hecho de que el índice tradicional de disponibilidad léxica se refiera a una propiedad de las palabras, y no de los hablantes, hacen que estos intentos estadísticos no pasen de ser meras descripciones de los acontecimientos que pretenden estudiarse. La excepción que confirma la regla se encuentra en el intento de López Chávez y Strassburger (1991) de crear un índice de disponibilidad léxica individual a partir del grado de participación del hablante en el resultado total de la muestra. Sin embargo, la propuesta está basada en herramientas matemáticas no totalmente adecuadas a los objetivos perseguidos, que llevan a los autores a enunciar una serie de condiciones en la parte final de su trabajo que, incluso, invalida el propio modelo propuesto y justifica el desarrollo de otros más avanzados que respon-

dan al concepto que desean representar. En cualquier caso, no está de más recordar que desde la inauguración del "Proyecto panhispánico de estudio sobre la disponibilidad léxica" (PPHDL), este tipo de estudios ha incorporado determinadas innovaciones que buscan encontrar el mejor sistema posible de tratamiento y explotación de los datos (Ávila Muñoz y Villena Ponsoda 2010; Callealta y Gallego 2016).

Hacia nuevas vías de estudio: el índice de centralidad léxica

El índice de disponibilidad léxica utiliza un ajuste matemático preciso, aunque en los estudios que lo emplean se advierte la ausencia de una exposición sólida de la naturaleza del modelo subyacente que lo sustenta. La falta de esta aproximación teórica ha propiciado la concatenación de sucesivos ajustes para acometer diferentes problemas que han ido apareciendo mediante su aplicación como, por ejemplo, fijar un límite a partir del cual el índice de disponibilidad de los vocablos sea lo suficientemente alto que permita incluir un elemento en un diccionario general de disponibilidad léxica (Samper 1999, pp. 554-555; Bartol 2001, pp. 227-230; Carcedo 2001, p. 62). Esta circunstancia ha llevado, en demasiadas ocasiones, a tomar decisiones subjetivas y difíciles de justificar que, como mostraremos a continuación con un ejemplo de caso práctico, quedan superadas con la utilización del índice de centralidad léxica.

De hecho, tal índice tiene su origen en la búsqueda de una respuesta a esta necesidad de aproximación teórica que explique el modelo subyacente en el que se basan los estudios de disponibilidad léxica (Ávila Muñoz y Villena Ponsoda 2010). Dicha búsqueda ha permitido construir no sólo el índice de centralidad, sino también otros elementos complementarios a partir de la interpretación de los datos obtenidos. El índice de centralidad ofrece así la posibilidad de incorporar estas evaluaciones en un marco teórico que permite desarrollos posteriores por medio de un modelo que proporciona resultados coherentes, en el sentido de que son semejantes a los ofrecidos por el índice de disponibilidad; con ello, excepto en casuísticas muy específicas y poco naturales, se ha logrado obtener resultados comparables, ya que se pretende representar la misma información.

Del léxico disponible al léxico accesible. Una transformación teórica necesaria

El concepto de *disponibilidad léxica* está basado en una sencilla consideración: existe un vocabulario compartido por los miembros de una comunidad de habla asociado a determinados prototipos cognitivos muy frecuentes (Ávila Muñoz y Villena Ponsoda 2010). Como hemos mostrado en el apartado anterior, el índice de disponibilidad es un atributo propio de las palabras que, en nuestro caso, hemos adaptado para poder emplearlo como característica comunitaria. Esta transformación se articula según las siguientes premisas:

1) El estímulo original que pone en marcha el proceso asociativo de palabras es un punto de acceso a una red de elementos léxicos que el sujeto relaciona entre sí y con el concepto sugerido por el estímulo inicial conocido tradicionalmente como centro de interés.

2) La estructura de esa red es subjetiva e inherente al individuo, por lo que, estrictamente, no puede extrapolarse a otros sujetos.

3) Por diversas razones, entre ellas sociales y culturales, los individuos construyen estructuras léxicas semejantes. Esa metaestructura sociocultural compartida es lo que consideramos el léxico asociado a un prototipo cognitivo comunitario. Es decir, el prototipo comunitario compartido no existe por sí mismo, sino como realización del conjunto de hablantes. Si se cambia el conjunto de hablantes, es posible, e incluso probable, que cambie la metaestructura compartida.

4) Cuando se solicita el listado de disponibilidad a un hablante, éste accede a los elementos léxicos más cercanos al concepto propuesto mediante un estímulo cognitivo (centro de interés). Posteriormente, recorre su red léxica hacia elementos más alejados de ese punto de entrada a medida que consume los elementos más accesibles. Es factible que vuelva a reintroducirse en la red en el momento en que considera que se ha alejado demasiado y encuentra un nuevo punto de entrada (Ávila Muñoz y Sánchez Sáez 2014).

5) En los listados de léxico disponible individuales son relevantes tanto la cantidad de vocabulario aportado como la velocidad de acceso. Evidentemente, el número de palabras que puede enunciar un sujeto es relevante para determinar su diversidad léxica, pero no exclusivo (Villena Ponsoda *et al.* en prensa).

6) El orden de aparición de los términos es importante respecto a la accesibilidad, aunque, como hemos señalado, pueden producirse puntos de reentrada (reinicialización de la prueba en un mismo proceso) que limitarían esa importancia.

7) Se considera que la configuración de un prototipo cognitivo tiene un núcleo, accesible a todos los hablantes, y una periferia, a la cual accederán los individuos en función de su mayor diversidad léxica. La distinción entre núcleo y periferia es gradual y no discreta.

Este último punto da lugar al modelo que explica la naturaleza del índice de centralidad léxica. En el estudio del modelo hay limitaciones inherentes a la forma y estructura de las pruebas. Aunque consideramos que la estructura del léxico es una red multiconectada, la determinación de esta red requeriría experiencias que superan las capacidades de cualquier experimento. Se recurre por lo tanto a una simplificación del modelo en la que se considera que un prototipo compartido es una red de términos con distintos grados de accesibilidad, obviando conscientemente la naturaleza de esas conexiones, y orientando la estructura del estudio hacia la determinación de la facilidad de acceso a los términos. Así, se considera relevante la posición de los términos, pero se ignora —de forma artificial, aunque necesaria— su secuencia de aparición.

Desde nuestro punto de vista, la aproximación teórica que mejor traslada el modelo que subyace a los estudios de disponibilidad hacia un marco matemático consolidado que permita la construcción del nuevo índice propuesto es la teoría de los conjuntos difusos (Zadeh 1965; Zimmermann 2001) y, en particular, la teoría de la posibilidad (Zadeh 1978), que proporciona herramientas de caracterización con un comportamiento consolidado y contrastado. A grandes rasgos, los conjuntos difusos son una generalización de la teoría de conjuntos en la que, en lugar de la pertenencia absoluta de elementos a categorías, se considera la compatibilidad de éstos con el concepto representado por el conjunto. Según la teoría clásica, la pertenencia o no de un elemento a un (sub)conjunto viene dada por una función específica que otorga el valor de 1 a los elementos que pertenecen al conjunto; aquellos elementos que no pertenecen a él obtienen el valor de 0. Sin embargo, en la teoría de los conjuntos difusos los elementos pueden obtener cualquier valor entre 0 y 1 según la compatibilidad de cada uno de ellos con la clase representada por el conjunto. En los casos extremos, 0

y 1 equivalen semánticamente a la teoría clásica de conjuntos, pero la teoría de los conjuntos difusos nos permite establecer diferentes niveles de compatibilidad entre un elemento y el conjunto respecto al que se mide esa compatibilidad, en lugar de la dicotomía pertenencia-no pertenencia propia de la teoría clásica. Este concepto de *compatibilidad* se corresponde, en nuestro caso, con el concepto que identificamos como *centralidad*.

Una de las herramientas que pone a nuestro alcance la teoría de los conjuntos difusos es la determinación del valor de compatibilidad característico del conjunto difuso, FEV (*Fuzzy expected value*), o su variación WFEV (*Weighted fuzzy expected value*). Con su empleo podemos fijar, por ejemplo, un límite de caracterización de los valores de pertenencia y, además, establecer parámetros para identificar elementos "muy característicos" o "poco característicos" en el conjunto analizado. Se consigue así, en definitiva, proponer una marca objetiva de corte en los niveles superiores e inferiores del conjunto difuso que no dependa directamente de la percepción subjetiva del investigador. Esta marca objetiva pone en relación directa el grado de compatibilidad de los elementos con una valoración del conjunto de los elementos seleccionados, factor este último que nos ha permitido estudiar el proceso mediante parámetros.

Para determinar el espectro de un centro de interés partimos de la construcción de un conjunto difuso que represente la realización llevada a cabo por cada individuo a partir de las listas de palabras que ha producido. Para ello, se recurre a una valoración de cada término (t) en cada lista individual en función de dos parámetros: la posición que ocupa en la lista (n) y una constante para cada uno de los conjuntos de datos analizados (k) cuyo valor por defecto es 1:

$$t = \frac{k}{n}$$

Una vez que se han determinado los espectros individuales, se procede a la construcción del modelo colectivo. Para ello, se toma como referencia el valor de nula disponibilidad para todos los términos. De esta manera, incorporamos el valor proporcionado en el espectro individual para cada término, y lo acumulamos en el modelo colectivo siguiendo una ley que premie su representatividad si el término aparece en la realización de ese

hablante con altos índices de disponibilidad. Una ley que opera de esta forma es la probabilística:

$$a + b – a \cdot b;$$

es decir, si la disponibilidad de un término, en cualquier momento del proceso, fuera de 0.3, y para un hablante ese término tuviera una disponibilidad personal de 0.2, el nuevo valor del término sería:

$$0.3 + 0.2 - 0.3 \times 0.2 = 0.5 - 0.06 = 0.44;$$

si, por el contrario, el término no aparece en el nuevo hablante, la disponibilidad no mejoraría, al sumar cero.

El mayor inconveniente de este tipo de valoración es que nunca decrece: o no varía (si se le añade un valor nulo) o crece, aunque sea poco (si el valor que se añade no es nulo). Así, es fácil que los valores se acerquen tanto a 1 que se hagan indistinguibles. El parámetro *k* que se tomó en el paso inicial tiene como propósito conseguir que las valoraciones de los hablantes sean tales que se puedan acumular, según la ley probabilística, y que las valoraciones cuantitativas sean fácilmente interpretables. La determinación de este valor *k* responde a la pregunta de en qué situación de ocurrencia se puede considerar que un término se sitúa en el núcleo del centro de interés de los hablantes. Esto nos ha permitido establecer el valor constante de *k* para cada uno de los conjuntos difusos creados, lo cual hace posible plantear una situación que puede justificarse de forma relativamente poco subjetiva y admite determinar de forma unívoca su interpretación.

En definitiva, el índice de centralidad léxica se ha elaborado como un conjunto difuso con valores de compatibilidad para cada uno de los elementos que lo componen (entre 0 y 1). Estos valores de compatibilidad se identifican con el concepto de *centralidad léxica*. La principal ventaja del modelo propuesto es el establecimiento de un marco de trabajo en el que se pueden aplicar herramientas matemáticas que no estaban a nuestro alcance en los trabajos anteriores de disponibilidad léxica. Si bien los resultados serán similares, a partir de ahora pueden justificarse[1].

[1] Los investigadores interesados disponen de un programa electrónico de acceso gratuito y manejo intuitivo construido en el entorno *R*. Por

Del índice de disponibilidad al índice de centralidad léxica. Ejemplo de caso práctico

Como hemos visto, el índice de disponibilidad es un excelente indicador del grado de prototipicidad que los vocablos poseen dentro de cada uno de los centros de interés, pues trasciende la simple naturaleza de los vocablos y se relaciona con la categorización conceptual colectiva (*Prototypes theory*: Wittgenstein 1953; Rosch 1978; Lakoff 1987). Este simple cambio de perspectiva nos permite transformar el índice de disponibilidad (atributo propio de las palabras) en el índice de centralidad léxica (atributo comunitario). Tal y como vamos a demostrar a continuación, aunque la transformación es simple, las consecuencias teórico-metodológicas son importantes.

Parece claro que la disponibilidad de un vocablo en un centro de interés responde, esencialmente, al concepto de *accesibilidad*. De hecho, se podría interpretar que, *grosso modo*, las listas producidas por cada informante corresponden, en su forma, a una representación del acceso a su léxico particular. Evidentemente, obtener la estructura de esta red léxica para cada sujeto es una tarea imposible, ya que se supone determinada por multitud de factores biográficos incontrolables, sin tener en cuenta el hecho de que es una estructura dinámica en continuo cambio provocado por la interacción del hablante con el entorno. Sin embargo, a partir de las realizaciones particulares, el índice de disponibilidad puede permitirnos estimar cuantitativamente la estructura de la accesibilidad al léxico para una población en un centro de interés concreto. Como hemos mostrado en el apartado anterior, la cuantificación de esta accesibilidad es posible y representa la medida del concepto de disponibilidad de cada término en ese centro de interés, una vez que se ha integrado la información proporcionada por todos y cada uno de los individuos que forman la población.

La asociación entre las palabras y la accesibilidad que presenta cada una de ellas determina la representación de la estructura del léxico en un centro de interés. En esa representación los vocablos más próximos al prototipo del núcleo temático

medio de esta herramienta se pueden calcular los índices de disponibilidad y centralidad, así como los conjuntos de corte correspondientes a cada una de las listas mediante las herramientas señaladas (Ávila Muñoz *et al.* 2021). Véase https://www.youtube.com/watch?v=IU5VfUvG4Ag [consultado el 23 de junio de 2022].

mostrarán mayor valor de accesibilidad (o, lo que es lo mismo, mayor índice de disponibilidad).

La Tabla 1 muestra los vocablos encontrados en el centro de interés *Pandemia* (ID) > 0.1 y empleados por el 30% de la muestra (N = 220) que formó parte del estudio de las percepciones compartidas en una población de estudiantes universitarios españoles (véase Ávila Muñoz *et al.* 2020, fuente de todas las tablas y dibujos de este trabajo). Se pretendía acceder a la percepción de la realidad de la población estudiada durante el período de confinamiento severo vivido en España durante los meses de marzo-abril de 2020 a consecuencia de la crisis sanitaria generada por el COVID-19.

Tabla 1

Vocablos con índice de disponibilidad (ID) > 0.1 y empleados por el 30% de la muestra

Pandemia		
Vocablo	*Disponibilidad*	*% Aparición*
1. Terror	0.29463	62.500
2. Muerte	0.27140	50.000
3. Enfermedad	0.24871	48.321
4. Virus	0.20661	42.500
5. Confinamiento	0.19760	40.000
6. Coronavirus	0.17098	40.000
7. Crisis	0.13434	36.389
8. COVID-19	0.10912	33.611

Como se observa, los vocablos empleados por un mayor porcentaje de la población estudiada son los que alcanzan mayor índice de disponibilidad. Ello nos lleva a confirmar que los vocablos con mayor índice de disponibilidad de cada centro de interés formarían la categorización conceptual colectiva del estímulo en cuestión, mientras que, a medida que el índice de disponibilidad de los vocablos disminuye y, por tanto, es empleado por menor número de hablantes, los elementos léxicos se van alejando del núcleo prototípico del centro de interés. La Tabla 2 presenta los veinte vocablos que logran menor índice de disponibilidad en el mismo centro de interés y que, según nuestra hipótesis, están más alejados del núcleo de categorización colectiva del centro de interés.

Tabla 2

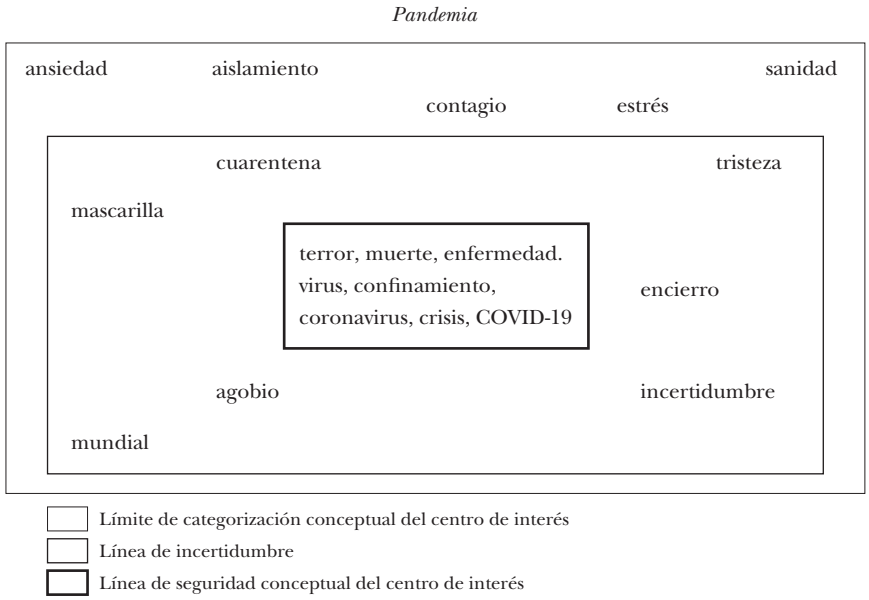
Vocablos con índice de disponibilidad (ID) > 0,1 y empleados por el 30 % de la muestra

Pandemia		
Vocablo	*Disponibilidad*	*% Aparición*
1. Gel hidroalcohólico	0.00150	0.454
2. Multa	0.00150	0.454
3. Empleado	0.00150	0.454
4. Ducha	0.00137	0.454
5. Deporte	0.00137	0.454
6. Investigación	0.00137	0.454
7. Lavarse las manos	0.00126	0.454
8. Pasear al perro	0.00126	0.454
9. Sin feria	0.00116	0.454
10. ERTE	0.00116	0.454
11.Temperatura	0.00106	0.454
12. Paciente	0.00098	0.454
13. Tos	0.00090	0.454
14. Fallecimiento	0.00082	0.454
15. Abandono	0.00077	0.454
16. Porcentaje	0.00076	0.454
17. Fase	0.00069	0.454
18. Estado	0.00064	0.454
19. Núcleo	0.00056	0.454
20. Foco	0.00049	0.454

En función de los valores alcanzados por cada vocablo en sus correspondientes centros de interés, el índice de disponibilidad léxica ha permitido hasta ahora establecer, de manera aproximada, diferentes líneas de seguridad de categorización conceptual que abarcarían, en un primer estadio, los vocablos más próximos al núcleo prototípico del centro de interés estudiado (Cuadro 1). A medida que decrece el índice de disponibilidad y, por tanto, los vocablos se alejan del centro del estímulo, los elementos léxicos comienzan a acercarse a lo que podemos llamar "línea de incertidumbre", límite a partir del cual el sujeto podría empezar a dudar de si la información que está aportando corresponde exactamente al estímulo propuesto; así, pues, tal límite puede considerarse la frontera a partir de la que el hablante duda acerca de si el término anotado es suficientemente compatible o no con el prototipo (Cuadro 1).

Cuadro 1

Representación aproximada de la categorización conceptual colectiva basada en el índice de disponibilidad



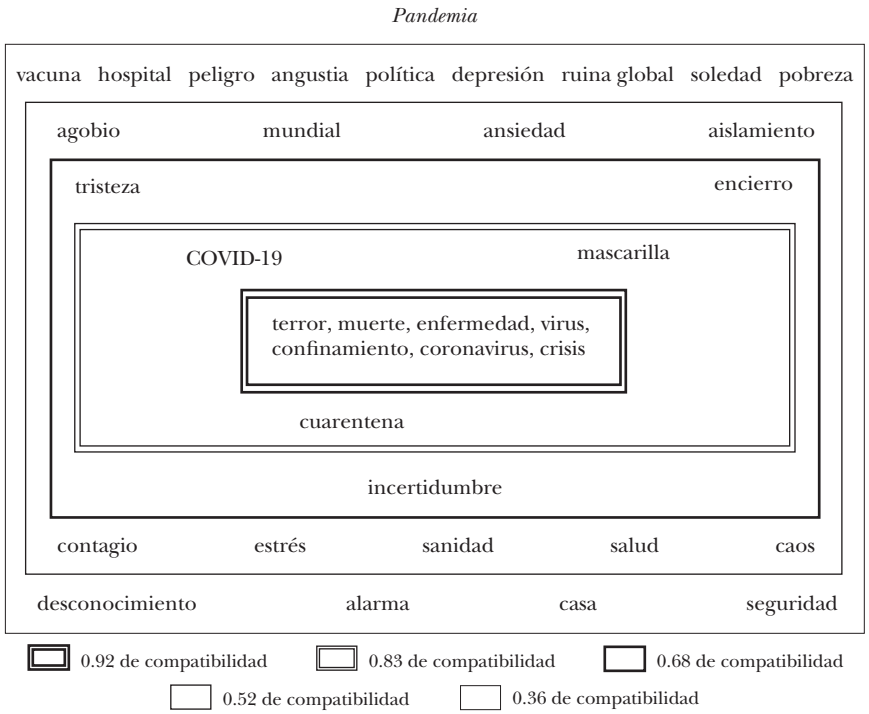
La cuestión está en la dificultad que los estudios de disponibilidad léxica encuentran a la hora de señalar las diferentes líneas de seguridad de categorizaciones conceptuales, pues, como se ha comentado más arriba, los criterios usados han sido siempre subjetivos o, cuando menos, poco justificados. Sin embargo, cuando aplicamos a los datos de disponibilidad el modelo teórico que ha servido para la determinación del índice de centralidad léxica, obtenemos un conjunto difuso que permite precisar mucho mejor cuál es la composición de las zonas próximas al núcleo prototípico para, de esta manera, observar con mayor detalle la categorización conceptual compartida que atañe al estímulo inicial.

El modelo elaborado considera que, de los 301 vocablos que componen el núcleo temático, sólo 57 (19% del total) pueden considerarse en el rango de "muy compatible" con el conjunto al que representan. Este nivel de corte fue establecido de manera automática en un índice de compatibilidad de 0.28 por la herramienta *Fuzzy expected value* (FEV). Como se mues-

tra en el Cuadro 2, de estos 57 vocablos, 7 forman el núcleo de máxima prototipicidad (2.3% del conjunto con un nivel de compatibilidad del 0.92): *terror*, *muerte*, *enfermedad*, *virus*, *confinamiento*, *coronavirus*, *crisis*. A partir de aquí aparecen otros niveles de corte diferentes que corresponden a las distintas zonas que, aun siendo muy compatibles con el estímulo de referencia, se van alejando progresivamente del núcleo central. Como es lógico, el Cuadro 2 podría extenderse y abarcar los diferentes límites de seguridad hasta alcanzar la zona de incertidumbre límite del estímulo inicial. En cualquier caso, los vocablos que están representados en el Cuadro 2, al alcanzar el mayor índice de centralidad léxica, resultan ser los más compatibles con el enunciado del centro de interés o, lo que es lo mismo, suponen la categorización conceptual colectiva del estímulo propuesto (conjunto de referencia). Con este procedimiento de selección léxica evitamos trabajar con fenómenos particula-

Cuadro 2

Representación de la categorización conceptual colectiva del prototipo



res del habla individual a partir de herramientas matemáticas fiables y objetivas que nos permiten centrarnos únicamente en hechos de norma compartida.

Aplicaciones del índice de centralidad

Son numerosas las posibilidades de aplicación práctica que nos ofrece el empleo del índice de centralidad. Uno de los ámbitos en los que se puede obtener mayor beneficio de este nuevo índice es el de la enseñanza de la lengua, sea materna o extranjera, pues el hecho de establecer conjuntos de corte basados en el concepto de *prototipicidad* asociado a categorías cognitivas compartidas nos ofrece la posibilidad de delimitar objetivamente el vocabulario más accesible para una muestra de hablantes nativos y, a partir de ahí, de establecer el léxico que se debería enseñar a los estudiantes según sus niveles de adquisición (Ávila Muñoz 2016a). Sin embargo, el empleo del índice de centralidad léxica puede realizarse desde disciplinas adyacentes, que encontrarán en los resultados de su aplicación una fuente de datos fiable y relativamente fácil de estudiar mediante la cual se puedan llevar a cabo diferentes experimentos aplicados.

En primer lugar, no cabe duda de que cuando un hablante produce una lista de léxico disponible, realiza una tarea cognitiva compleja en la que intervienen determinados procesos psicológicos que, en gran medida, condicionan la mencionada lista. Por tanto, no es de extrañar que la psicolingüística pueda aprovechar los resultados complementarios ofrecidos por el índice de centralidad para acceder a la descripción, análisis e interpretación de prototipos comunitarios compartidos que podrían resultar, incluso, más naturales que las pruebas psicolingüísticas tradicionales (Ávila Muñoz y Sánchez Sáez 2014; Ávila Muñoz *et al.* 2020).

En segundo lugar, gracias a este índice, la sociolingüística puede abrir líneas de análisis que salvarán fácilmente los inconvenientes teóricos que dificultan el estudio cuantitativo del léxico (López Morales 1999, p. 25; Escoriza Morera 1999, 2002, 2003 y 2004).

En tercer lugar, las diferencias culturales que se observan entre las comunidades de habla estudiadas quedan recogidas en las listas que se obtienen en cada sintopía. Al aplicar el índice de centralidad léxica, la etnolingüística encontrará una rica fuen-

te de datos sobre determinados aspectos culturales de los grupos por medio de estos listados sometidos a análisis.

En cuarto lugar, con la aplicación de este nuevo índice se podrá fomentar la tradición consistente en realizar comparaciones diatópicas de los materiales proporcionados por los estudios de disponibilidad léxica. La dialectología comparada, por tanto, es otra de las disciplinas que pueden beneficiarse del nuevo índice para organizar las fuentes de datos cuantitativamente numerosos y cualitativamente diversos. Desde que López Chávez (1992 y 1993) desarrolló índices matemáticos que permiten la comparación de las listas de disponibilidad obtenidas en diferentes sintopías, todos los estudios que las han empleado muestran que junto a los vocablos comunes hay otros exclusivos. Con el empleo del índice de centralidad léxica, podremos obtener resultados aún más concluyentes sobre el grado de aproximación o compatibilidad entre los diferentes dialectos y sus percepciones categoriales.

En quinto lugar, se podrá determinar el léxico relevante para un centro de interés de manera objetiva. Como hemos visto más arriba, hasta ahora ha sido habitual la discusión entre diferentes opciones que no ofrecen otra justificación que planteamientos subjetivos, los cuales, aunque relevantes por sí mismos, no dejan de ser discutibles por la escasa formalización en la que se basan. Gracias a la herramienta *Fuzzy expected value* (FEV) podemos determinar el valor característico de pertenencia en un conjunto difuso respecto a determinada medida que en nuestro caso será el tamaño relativo de los conjuntos de términos que superan un nivel de compatibilidad (conjuntos de corte). Este FEV encuentra un valor de pertenencia que establece un equilibrio entre el número de términos que lo supera y el propio valor. Además, si ponderamos en diferente grado la relevancia del tamaño del conjunto de corte, se pueden establecer grados de restricción en su interior que nos permitirían construir conceptos como *muy representativo* o *poco representativo*.

Por último, en sexto lugar, el índice de centralidad nos permitirá acceder a la diversidad (riqueza) léxica de los individuos de una manera novedosa y original. En la construcción del nuevo índice hemos usado un modelo de representación del lenguaje donde el lexicón es una característica elaborada por los propios hablantes. Así, el núcleo del léxico de un prototipo determinado (en nuestro caso, presentado en forma de estímulo cognitivo que genera listas de palabras) estaría dispo-

nible para todos los sujetos. Por lo tanto, a cualquier individuo se le supone el acceso al léxico que forma parte del núcleo del prototipo, con lo que la caracterización de su diversidad léxica vendrá determinada, particularmente, por aquella parte del léxico que es más específica y menos característica entre el resto de la población. Según este planteamiento, las diferentes listas de disponibilidad léxica individuales pueden considerarse complementarias entre sí; por ello, nuestra propuesta abre nuevas vías de análisis cuantitativo que, hasta el momento, no se habían explorado: la hipótesis es que si un individuo actualiza palabras con un bajo índice de centralidad debe de tener más diversidad léxica. Parece lógico suponer que, si las palabras que proporciona son menos centrales (disponibles), debe de acceder más fácilmente a aquellas otras compartidas por todos, y a otras menos generalizadas, con lo que su diversidad léxica tendría que considerarse mayor (Villena Ponsoda *et al.* en prensa).

Conclusiones

La historia de la lexicoestadística hispánica se caracteriza por la ilusión de investigadores empeñados en construir redes de cooperación panhispánica con el propósito de poner en marcha macroproyectos que permitan describir, analizar y entender los mecanismos que subyacen al empleo del léxico de nuestra lengua. Todos ellos han utilizado índices estadísticos, sometidos luego a revisión, para lograr el perfeccionamiento de los análisis y la creación de herramientas apropiadas con que alcanzar los objetivos comunes. Con la perspectiva que nos ofrece el paso del tiempo, se puede hablar de una tradición consolidada que, a la vista de la pluralidad de resultados, tiene un futuro muy prometedor.

Con miras a contribuir al desarrollo de esta tradición, hemos revisado los índices mayoritariamente empleados, en particular el de disponibilidad léxica, para generar un nuevo indicador muy versátil que nos ha permitido idear un marco teórico firme que lo justifica y pergeñar nuevas vías de investigación gracias a la inclusión de herramientas hasta ahora inéditas en este ámbito. El índice de centralidad léxica se construye a partir de la elaboración de un conjunto difuso donde el valor de cada término viene determinado por su compatibilidad con el resto de términos que componen las listas y por el promedio de esas com-

patibilidades entre el número total de listas. Su naturaleza nos facilita la creación de prototipos cognitivos comunitarios con posibilidades reales de aplicación desde diversos ámbitos de estudio: la psicolingüística, la sociolingüística, la dialectología social, la etnolingüística o la enseñanza de lenguas son disciplinas que se pueden beneficiar de este nuevo índice, el cual, por lo demás, vale como atributo individual para acercarnos al cálculo de la diversidad (riqueza) léxica de los sujetos.

Referencias

Alameda, José Ramón y Fernando Cuetos 1995. *Diccionario de frecuencias de las unidades lingüísticas del castellano*, Universidad de Oviedo, Oviedo.

Alvar Ezquerra, Manuel, María José Blanco Rodríguez y Fernando Pérez Lagos 1994. "Diseño de un corpus español en el marco de un corpus europeo", en *Estudios para un corpus del español*. Eds. Manuel Alvar Ezquerra y Juan Andrés Villena Ponsoda, Universidad de Málaga, Málaga, pp. 9-31.

Ávila, Raúl 1988. "Lengua hablada y estrato social. Un acercamiento lexicoestadístico", *Nueva Revista de Filología Hispánica*, 36, pp. 131-148; doi: 10.24201/nrfh.v36i1.668.

Ávila, Raúl 1999. "Difusión Internacional del Español por Radio, Televisión y Prensa", en *Actas del XI Congreso Internacional de la Asociación de Lingüística y Filología de América Latina*. Coords. José Antonio Samper Padilla y Magnolia Troya Déniz, Universidad, Las Palmas de Gran Canaria, t. 3, pp. 2507-2518.

Ávila Muñoz, Antonio Manuel 2016. "Can speakers' virtual lexical richness be calculated?: Individual and social determining factors", *Spanish in Context*, 13, pp. 285-307; doi: 10.1075/sic.13.2.06avi.

Ávila Muñoz, Antonio Manuel 2016a. "El léxico disponible y la enseñanza del español. Propuesta de selección léxica basada en la teoría de los conjuntos difusos", *Journal of Spanish Language Teaching*, 3, pp. 31-43; doi: 10.1080/23247797.2016.1163038.

Ávila Muñoz, Antonio Manuel y José María Sánchez Sáez 2014. "Fuzzy sets and prototype theory: Representational model of cognitive community structures based on lexical availability trials", *Review of Cognitive Linguistics*, 12, pp. 133-159; doi: 10.1075/rcl.12.1.05avi.

Ávila Muñoz, Antonio Manuel, Inmaculada Clotilde Santos Díaz y Ester Trigo Ibáñez 2020. "Análisis léxico-cognitivo de la influencia de los medios de comunicación en las percepciones de universitarios españoles ante la COVID-19", *Círculo de Lingüística Aplicada a la Comunicación*, 84, pp. 85-95; doi: 10.5209/clac.70701.

Ávila Muñoz, Antonio Manuel, José María Sánchez Sáez y Nana Odishelidze 2021. "DispoCen. Mucho más que un programa para el cálculo de la disponibilidad léxica", *Estudios de Lingüística. Universidad de Alicante*, 35, pp. 9-36; doi: 10.14198/ELUA2021.35.1.

Ávila Muñoz, Antonio Manuel y Juan Andrés Villena Ponsoda (eds.) 2010. *Variación social del léxico disponible en la ciudad de Málaga*, Sarriá, Málaga.

Baldi, Pierre Luigi 1972. "Fattori sociali dell'abilitá lingüistica nella produzione scritta di bambini di 9-10 anni", *Studi Italiani di Lingüistica Teorica e Applicata*, 1, pp. 335-416.

Bartol, José Antonio 2001. "Reflexiones sobre la disponibilidad léxica", en *Nuevas aportaciones al estudio de la lengua española*. Coord. José Antonio Bartol, Luso-Española de Ediciones, Salamanca, pp. 221-235.

Callealta Barroso, Francisco y Diego Gallego Gallego 2016. "Medidas de disponibilidad léxica: comparabilidad y normalización", *Boletín de Filología*, 51, 1, pp. 39-92**;** doi: 10.4067/S0718-93032016000100002**.**

Carcedo González, Alberto 2001. *Léxico disponible de Asturias*, Departamento de Español de la Universidad de Turku, Turku.

Ellis, Andrew 1985. "The production of spoken words. A cognitive neuropsychological perspective", en *Progress in the psychology of language*. Ed. Andrew Ellis, Lawrence Erlbaum Associates, Hillsdale, NJ, t. 2, pp. 107-145.

Escoriza Morera, Luis 1999. "El concepto de variación lingüística", en *Lingüística para el siglo xxi: III Congreso organizado por el Departamento de Lengua Española*, Universidad de Salamanca, Salamanca, pp. 533-540.

Escoriza Morera, Luis 2002. "Posibilidades teóricas en el establecimiento de variantes léxicas", en *Actas del IV Congreso de Lingüística General*, Universidad de Cádiz, Cádiz, pp. 877-886.

Escoriza Morera, Luis 2003. *Perspectivas de análisis en el ámbito de la variación lingüística*, Universidad de Cádiz, Cádiz.

Escoriza Morera, Luis 2004. "Posibilidades de aplicación del concepto de variación lingüística al nivel léxico en el ámbito de la sociolingüística", en *Actas del V Congreso de Lingüística General*, Universidad de León, León, pp. 829-835.

García Hoz, Víctor 1953. *Vocabulario usual, vocabulario común y fundamental. Determinación y análisis de sus factores*, Consejo Superior de Investigaciones Científicas, Madrid.

Gougenheim, Georges 1958. *Dictionnaire fondamental de la langue française*, Didier, Paris.

Gougenheim, Georges 1967. "La statistique du vocabulaire et son application dans l'enseignement des langues", *Les Langues Modernes*, 61, pp. 137-144.

Guiraud, Pierre 1960. *Problèmes et méthodes de la statistique linguistique*, Reidel Pub., Dordrecht.

Juilland, Alphonse 1973. *The Romance languages and their structures. First series*, De Gruyter, Berlin.

Juilland, Alphonse & Eugenio Chang-Rodríguez 1964. *Frequency dictionary of Spanish words*, De Gruyter, La Haya.

Keniston, Hayward 1941. *A standard list of Spanish words and idioms*, Heath and Co., New York.

Lakoff, George 1987. *Women, fire and dangerous things: What categories reveal about the mind*, The University of Chicago Press, Chicago-London.

Lope Blanch, Juan Miguel 1967. "Para el conocimiento del habla hispanoamericana", en *II Simposio de Bloomington. Actas, informes, comunicaciones*, Instituto Caro y Cuervo, Bogotá, pp. 255-264.

Lope Blanch, Juan Miguel 1977. *Estudios sobre el español hablado en las principales ciudades de América*, Universidad Nacional Autónoma de México, México.

López Chávez, Juan 1992. "Alcances panhispánicos del léxico disponible", *Lingüística*, 4, pp. 26-124.

López Chávez, Juan 1993 "Léxico fundamental panhispánico: realidad o utopía", *Actas del III Congreso Internacional sobre el Español de América*, Universidad Católica de Chile, Santiago de Chile, t. 2, pp. 1006-1014.

López Chávez, Juan y Carlos Strassburger Frías 1991. "Un modelo para el cálculo de disponibilidad léxica individual", en *La enseñanza de la lengua materna. Actas del II Seminario Internacional sobre Aportes de la Lingüística a la Enseñanza de la Lengua Materna*, Universidad de Puerto Rico, Río Piedras, pp. 99-112.

López Morales, Humberto 1999. *Léxico disponible de Puerto Rico*, Arco/ Libros, Madrid.

López Morales, Humberto (coord.) 1983. *Introducción a la lingüística actual*, Playor, Madrid.

Michéa, René 1953. "Mots fréquents et mots disponibles; un aspect nouveau de la statistique du langage", *Les Langues Modernes*, 47, pp. 338-344.

Müller, Charles 1965. "Fréquence, dispersion et usage", *Cahiers de Lexicologie*, 7, pp. 33-42.

Müller, Charles 1977. *Principes et méthodes de statistique lexicale*, Hachette, Paris.

Rodríguez Bou, Ismael 1952. *Recuento del vocabulario español*, Consejo Superior de Enseñanza de la Universidad de Puerto Rico, Río Piedras.

Rosch, Eleanor 1978. "Principles of categorization", en *Cognition and categorization*. Eds. Eleanor Rosch & Barbara Lloyd, Lawrence Erlbaum, Hillsdale, NJ, pp. 27-48.

Samper, José Antonio 1999. "Léxico disponible y variación dialectal: datos de Puerto Rico y Gran Canaria", en *Estudios de lingüística hispánica. Homenaje a María Vaquero*. Eds. Amparo Morales, Eduardo Forastieri, Julia Cardona y Humberto López Morales, Universidad de Puerto Rico, San Juan de Puerto Rico, pp. 550-573.

Sinclair, John 1991. "The automatic analysis of corpora", en *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991*. Ed. Jan Svartvik, De Gruyter Mouton, Berlin-New York, pp. 379-400.

Vega, Manuel de, Manuel Carreiras, Manuel Gutiérrez Calvo y María L. Alonso-Quecuty 1990. *Lectura y comprensión. Una perspectiva cognitiva*, Alianza Editorial, Madrid.

Villena Ponsoda, Juan Andrés, Antonio Manuel Ávila Muñoz & José María Sánchez Sáez (en prensa). "Individual lexical breadth and its associated measures. A contribution to the calculation of individual lexical richness", en *New perspectives on Spanish lexical development*. Eds. Laura Pérez Marqués & Irene Checa, De Gruyter, Berlin-New York.

Wittgenstein, Ludwig 1953. *Philosophical investigations*, McMillan, New York.

Zadeh, Lofti Asker 1965. "Fuzzy sets", *Information and Control*, 8, pp. 338-353.
Zadeh, Lofti Asker 1978. "Fuzzy sets as a basis for a theory of possibility", *Fuzzy Sets and Systems*, 1, pp. 3-28.
Zimmermann, Hans-Jurgen 2001. *Fuzzy set theory and its applications*, Kluwer Academic Publishers, Boston, MA.