ISSN-e: 1697-0780

linred
. artículo
Lingüística en la Red

**Antonio Manjón-Cabeza Cruz**
Universidad de Granada

# *Determinación de asociaciones léxicas. Evaluación de tres métodos*

## *Determination of lexical associations. Evaluation of three methods*

Cómo citar:
Manjón-Cabeza Cruz, Antonio (2024): "Determinación de asociaciones léxicas. Evaluación de tres métodos", *Lingüística en la Red,* XXI, pp. 1-16. DOI: 10.37536/linred.2024.XXI.2467

**Resumen**

Se han propuesto varios procedimientos para descubrir y medir asociaciones léxicas en las encuestas de léxico disponible. En este trabajo evaluamos las ventajas e inconvenientes de tres métodos para descubrir asociaciones léxicas: Dispografo (Echeverría *et al.* 2008), índice de distancias ponderado (IDP) (Manjón-Cabeza 2008) e índice de contigüidad de vocablos (ICV) (Guerra Salas *et al.* 2015). Dispografo está limitado porque no tiene en cuenta la frecuencia relativa del vocablo ni normaliza los resultados a una escala. ICV tiene en cuenta la frecuencia de los vocablos, pero no normaliza a una escala. IDP tiene en cuenta tanto la frecuencia de los vocablos como la normalización a una escala.

**Palabras clave**

Sociolingüística; psicolingüística; disponibilidad léxica; redes semánticas; métodos; ventajas e inconvenientes.

**Abstract**

Several procedures have been proposed to discover and measure lexical associations in lexical availability surveys. In this paper we evaluate the advantages and disadvantages of three methods to discover lexical associations: Dispograph (Echeverría *et al.* 2008), weighted distance index (WDI) (Manjón-Cabeza 2008) and vocabulary contiguity index (VCI) (Guerra Salas *et al.* 2015). Dispograph is limited because it does not take into account the relative frequency of the word nor does it normalize the results to a scale. ICV takes into account the frequency of the vocabularies, but does not normalize to a scale. IDP takes into account both vowel frequency and normalization to a scale.

**Keywords**

Sociolinguistics; psycholinguistics; lexical availability; semantic network; methods; advantages and disadvantages.

Fecha de recepción: 20/2/2024 - Fecha de aceptación: 13/5/2024
DOI: 10.37536/linred.2024.XXI.2467

ISSN-e: 1697-0780

**Antonio Manjón-Cabeza Cruz**
Universidad de Granada



## 1. Introducción

Los estudios sobre léxico disponible constituyen un punto de confluencia entre la sociolingüística, la enseñanza de lenguas y la psicolingüística. Si nos centramos en el ámbito hispánico, el interés en cuestiones psicolingüísticas ha estado presente desde hace décadas.

Quizá sea Cañizal Arévalo (1991) la primera en proponer la existencia de redes semánticas asociadas al léxico disponible. Por su parte, Galloso Camacho (2003) describe lo que denomina cadenas de evocación, mientras Paredes García (2006) prefiere el término de asociaciones léxicas, Hernández Muñoz (2006) comienza a tratar distintos aspectos cognitivos de la disponibilidad léxica, Gómez Molina (2009) vuelve a redundar en la aplicación psicolingüística de las encuestas del léxico disponible, etc.

Se puede observar que, casi siempre, la vertiente psicolingüística de los estudios sobre léxico disponible está relacionada con la teoría de redes semánticas (Collins y Loftus 1975; Collins y Quillian 1969; Quillian 1968; Stayvers y Tenenbaum 2005). Los defensores de esta teoría entienden que las distintas relaciones que se establecen entre palabras son fundamentales en la organización del lexicón mental.

Las relaciones son, sobre todo, referenciales, es decir, determinadas por la realidad (*parchís* y *oca* presentarán fuerte relación por tratarse de juegos de mesa que, al menos en la tradición española, suelen ir unidos en un mismo tablero); pero también aparecen relaciones semánticas (relacionamos *blanco* y *negro* porque son antónimos), fónicas (dentro de los medios de locomoción *barco, avión, coche, camión* es muy probable que estén más cerca en nuestro diccionario mental la pareja *avión-camión* por presentar un esquema silábico y acentual semejante) y gramaticales (en la serie *profesor, alumno, escribir, explicar* agruparemos por un lado *profesor-alumno* porque son sustantivos y, por otro, *escribir-explicar* porque son verbos).

De esta manera, el almacenamiento de las palabras en nuestra memoria iría conformando una inmensa red donde los nudos serían las palabras, que estarían unidas por multitud de relaciones. En términos de teoría de redes las palabras serían nodos y las relaciones, aristas.

Las primeras pruebas alegadas por los defensores de la existencia de redes semánticas en el lexicón no tenían que ver con el léxico disponible, sino que solían ser pruebas de emparejamiento (Jenkins 1970), en las que ante una palabra estímulo se pide a los encuestados que digan la primera palabra que se les ocurra. Estos experimentos muestran que las palabras de un mismo campo semántico están cerca; por ejemplo, el 75% de los encuestados une *blanco* con *negro,* mientras que los encuestados que responden *mar* a una palabra estímulo como *queso* son anecdóticos.

Las pruebas de emparejamiento, sin embargo, fueron objeto de críticas y objeciones. Destacamos que se pregunta por una sola palabra, pero lo más probable es que las uniones sean múltiples y de distintas clases. Asimismo, como señala Aitchinson (1987: 73), "Can we build up a detailed mental map from these responses? Unfortunately not, in spite of the enormous amount of information available from word association experiments".

Parecía, pues, imposible establecer redes semánticas realistas por la casi inabarcable cantidad de datos que se debe manejar. Sin embargo, las listas con las que trabajan los investigadores sobre léxico disponible pueden ayudar a determinar parcelas de la red semántica mental, lo que nos puede permitir una aproximación adecuada a redes semánticas naturales.

Esta idea subyace a las distintas propuestas de determinación de asociaciones léxicas. En este trabajo nos proponemos evaluar tres de ellas: Dispografo (Echeverría *et al.* 2008), el índice de distancias ponderado (IDP) propuesto por Manjón-Cabeza (2008) y el índice de contigüidad de vocablos (ICV) de Guerra Salas *et al.* (2015).

Hay otras propuestas de análisis de datos de disponibilidad léxica, como Dispocen de Ávila *et al.* (2021), que no tratamos aquí porque, aunque ofrece un tratamiento de los datos innovador, no se dedica exactamente a determinar asociaciones léxicas. De hecho, Ávila (2022) usa Dispografo para el descubrimiento de redes léxicas.

Del mismo modo, LexPro, la reciente propuesta de Hernández *et al.* (2023), tiene un interesante apartado para la determinación de redes léxicas, pero es deudor de Dispografo porque trabaja con listas de adyacencias o contigüidades, aunque con varias mejoras con respecto a la propuesta de Echeverría *et al.* (2008). Lexpro, por ejemplo, permite analizar las relaciones más allá de la pareja de palabras contiguas, hasta cinco nodos de distancia. Esta herramienta deberá ser analizada pormenorizadamente en trabajos próximos porque es muy posible que Dispografo deje de usarse y los investigadores pasen a emplear habitualmente Lexpro.

## 2. Datos y métodos

Parece evidente que para evaluar los tres métodos hay que partir de datos léxicos comunes para los tres. Esta es una de las principales aportaciones de este estudio porque, según lo que conocemos, nadie se ha propuesto evaluar los tres métodos aplicándolos al mismo corpus. Nos basaremos en datos de Manjón-Cabeza (2010). En aquella ocasión –y en esta– nos serviremos de datos de una encuesta de léxico disponible de 142 individuos de Toledo (España) sobre el centro de interés de juegos y diversiones. Debemos resaltar que haber escogido estos datos u otros no es muy relevante para este estudio porque no estamos haciendo estudio del léxico, sino el estudio de tres métodos. Trabajaremos, pues, con un total de 457 lexías y, en todos los casos, para mantener la homogeneidad del tratamiento de los tres métodos, las asociaciones serán no dirigidas, es decir, que no discriminaremos el orden de aparición de las contigüidades –*bigrams* o binomios en términos de Hernández *et al.* (2023)–, de modo que daremos el mismo valor a, por ejemplo, *futbol-baloncesto* que a *baloncesto-fútbol*.

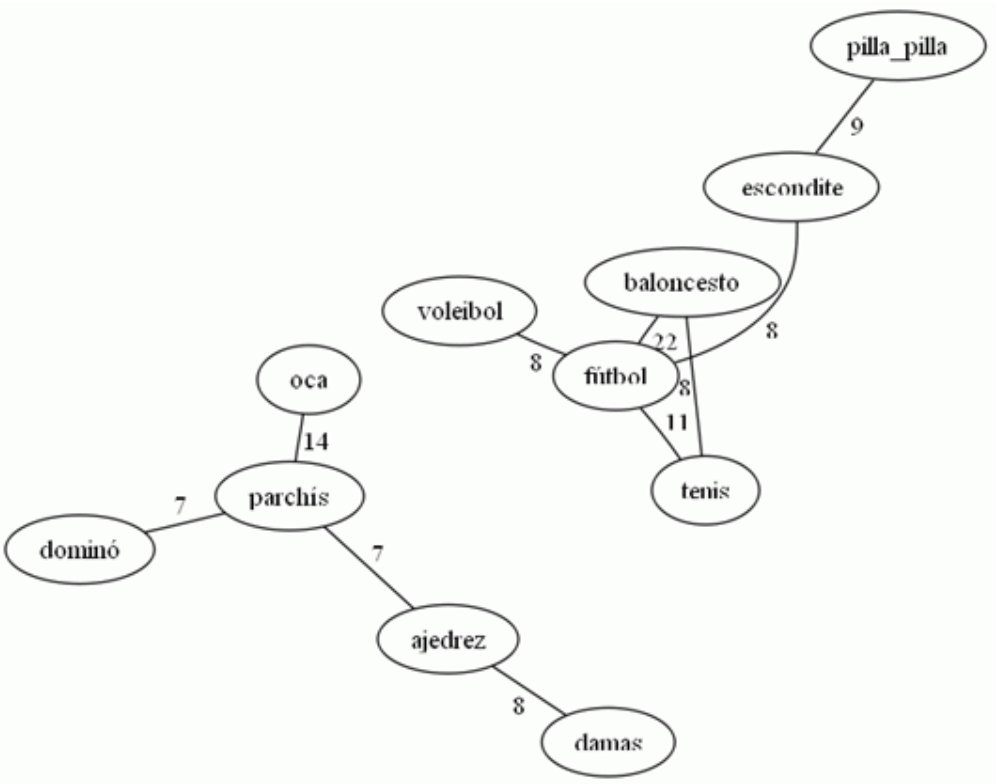
### *2.1 Dispografo*

En 2008 Echeverría y colaboradores publicaron un artículo donde hacían explícita una propuesta para el descubrimiento de redes léxicas ligadas a encuestas de léxico disponible. En realidad, el programa había sido presentado con anterioridad, en la reunión del proyecto panhispánico celebrado en San Millán de la Cogolla (La Rioja, ES) en 2005, de modo que, aunque otro de los métodos evaluados –Manjón-Cabeza (2008)– tenga la misma fecha de publicación, la propuesta de Echeverría es anterior. El programa utilizado se denomina Dispografo.

El carácter pionero y su facilidad de uso ha propiciado que los estudios de asociaciones léxicas mediante Dispografo hayan sido, con mucho, los más abundantes en el mundo hispánico, tanto en América, como en Europa. Sin ánimo de exhaustividad podemos referirnos a trabajos de Blanco *et al.* (2020), Ferreira y Echeverría (2010), Gómez Devís (2019 y 2021), Gómez Devís y Cepeda (2022), Gómez Devís y Gómez Molina (2022), Henríquez *et al.* (2016), Hernández *et al.* (2014), Hernández y Tomé (2017), López González (2014 y 2016), Mahecha y Mateus (2020), Sánchez-Saus (2019 y 2022), Santos Díaz (2017), y un largo etcétera.

Si sometemos nuestros datos a Dispografo obtenemos, en primer lugar, una red prácticamente inmanejable (como en cualquier método que trabaje con redes) porque aparecen todos los vocablos (nodos) con todas las aristas posibles. Es necesario podar por el peso de las aristas. Está claro que cuanto mayor sea el peso elegido más se reducirá la red. Elegimos la opción de una poda cercana al 25% del valor máximo. En este caso si podamos a 6, como se muestra en la figura 1, supone tener en cuenta un límite de 27,27% del máximo (22).

Figura 1. Dispografo. Aristas podadas a 6 y nodos podados a 1




Fuente: elaboración propia con Dispografo

En la figura 1 observamos que son 11 los vocablos integrados en la red y que hay tres conglomerados claros: deportes (*voleibol*, *fútbol*, *tenis*, *baloncesto*), juegos de mesa (*parchís*, *oca*, *dominó*, *ajedrez*, *damas*) y juegos infantiles (*escondite*, *pilla-pilla*).

## *2.2 Índice de distancias ponderado (IDP)*

En 2008 Manjón-Cabeza publicó un artículo donde se proponía una manera distinta de medir las asociaciones léxicas, tanto para intentar salvar las debilidades que presenta Dispografo, como para comparar encuestas de léxico disponible con distinto número de participantes.

El principal cambio consiste en que no se trabaja con contigüidades absolutas, sino con un índice de distancias, que es un valor que se determina mediante un cálculo. Esto implica la elaboración de una matriz de distancias en la que se toman en cuenta solo las palabras más disponibles. De modo que se siguen los siguientes pasos:

a) Selección de las palabras más disponibles. Para esta ocasión se trabaja con las 37 primeras palabras porque suponen un 50% de la frecuencia acumulada y son, por orden de disponibilidad: *fútbol, parchís, escondite, baloncesto, cartas, comba, cine, ajedrez, tenis, dominó, leer, pilla-pilla, bailar, oca, mus, corro de la patata, natación, play-station, música, televisión, canicas, balonmano, damas, bicicleta, pasear, petanca, ordenador, tute, monopoli, solitario, chapas, voleibol, teatro, escondite inglés, viajar*, *tres en raya* y *cantar*.

b) Medición de las contigüidades inmediatas y mediatas, asignando un valor exponencial decreciente: si la palabra está contigua se le asigna un valor de 100; si hay una distancia de una palabra intermedia el valor es de 33.3; y si la distancia es de dos palabras intermedias, el valor es 11.1. Más distancias no se tienen en cuenta porque trabajamos con el valor máximo de siete unidades en la memoria de trabajo.

c) Se suman las distancias entre palabras y se establece la media ponderada de las distancias, de modo que el sumatorio de las distancias entre dos palabras habrá de ser dividido por el número de apariciones de la palabra con menor ocurrencia.

La fórmula puede ser representada como:

$$IDP\,(AB) = \frac{\sum_{i=1}^{J} d\,(AB)}{NB}$$

donde:
IDP = índice de distancia ponderada.
d = distancia.
A y B son las dos palabras entre las que se mide la distancia, siendo B la de menor aparición.
$N_B$ = número de apariciones de la palabra menos frecuente.
J corresponde al número de informantes del grupo.

Este método, a diferencia de Dispografo, ha sido poco usado. Junto con Manjón-Cabeza (2008, 2009 y 2010), se puede citar a Podhajská (2023a y 2023b).

Se ha comentado que el resultado es necesariamente una matriz de datos. Con nuestras 37 palabras más disponibles esa matriz consta de 648 casillas que, en caso de haber optado por la direccionalidad en el tratamiento de los datos, habría sido del doble: 1296. A modo de ejemplo, ofrecemos la matriz resultante para los diez vocablos más disponibles en la tabla 1.

Tabla 1. Matriz de IDP para los diez vocablos más disponibles

| | fútbol | parchís | escondite | baloncesto | cartas | comba | cine | ajedrez | tenis |
|---|---|---|---|---|---|---|---|---|---|
| fútbol | | | | | | | | | |
| parchís | 4.44 | | | | | | | | |
| escondite | 16.87 | 6.17 | | | | | | | |
| baloncesto | 52.89 | 3.14 | 6.52 | | | | | | |
| cartas | 7.58 | 15.44 | 13 | 9.75 | | | | | |
| comba | 5.56 | 3.98 | 14.81 | 8.83 | 4.27 | | | | |
| cine | 10.03 | 4.66 | 4.66 | 2.86 | 3.58 | 3.58 | | | |
| ajedrez | 6.81 | 32.61 | 1.43 | 5.02 | 19.71 | 5.01 | 3.23 | | |
| tenis | 44.44 | 5.37 | 1.79 | 30.1 | 4.3 | 2.51 | 4.66 | 0.72 | |
| dominó | 1.78 | 33.33 | 9.33 | 8 | 15.55 | 4.44 | 1.33 | 12 | 4 |

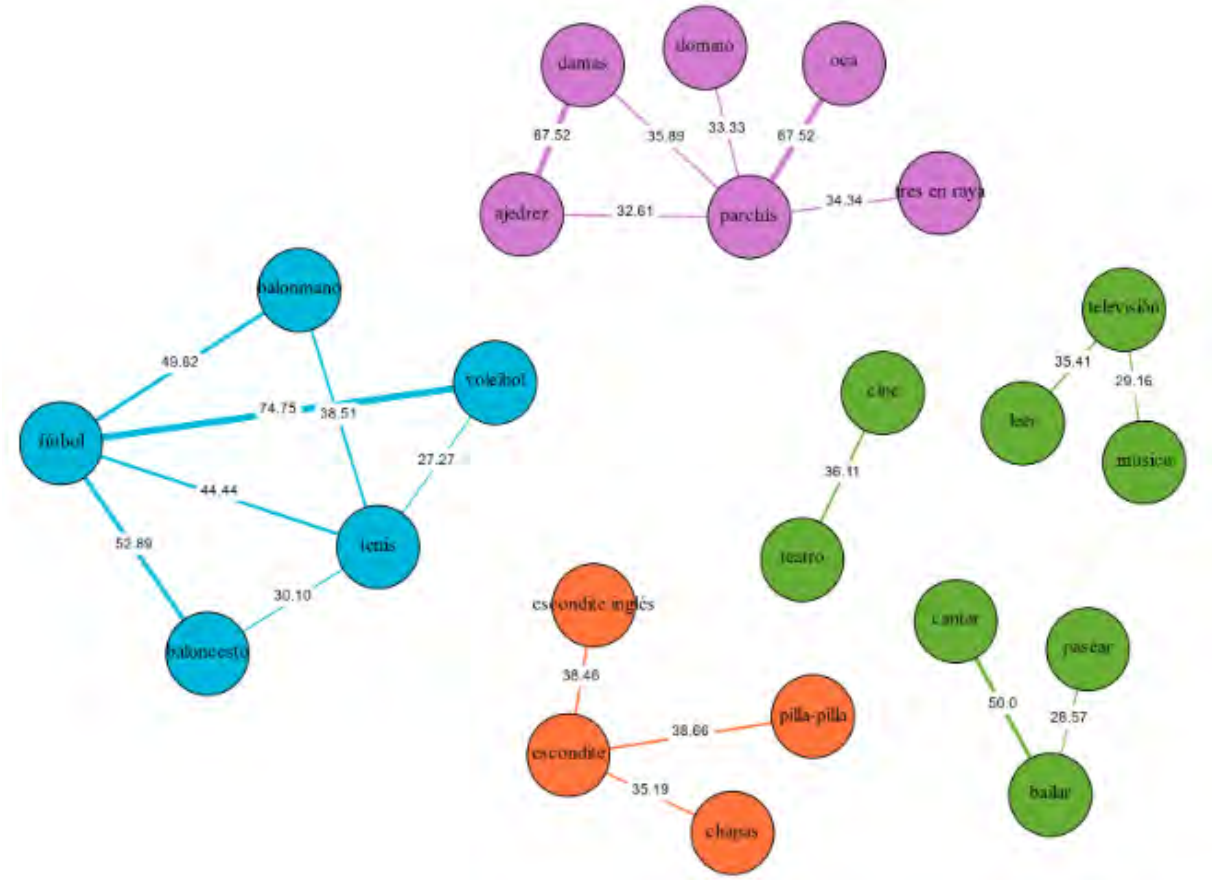
Fuente: elaboración propia

Debe observarse que el IDP oscilará entre 100 (en el caso de que dos palabras se dijeran siempre inmediatamente contiguas en todos los casos posibles) y 0 (en el caso de que dos palabras nunca fueran actualizadas contiguas inmediata o mediatamente).

A diferencia de Dispografo, para la representación de la red semántica se necesita un programa de gestión de grafos. Hemos usado en este caso Gephi para generar la red (Bastian *et al.* 2009).

Para la comparación con Dispografo, habrá que ofrecer una red podada al 27.27% (de manera similar a la Figura 1). Aquí cabrían dos opciones: tomar como valor máximo el máximo teórico (100) o tomar como valor máximo el valor máximo hallado en la matriz: 74.75 entre *fútbol* y *voleibol*. Esta duda se plantea porque el IDP está ponderado mientras que los resultados por contigüidad no lo están. Optamos por la primera opción, porque es más restrictiva, en el sentido de que la red será más limitada con un valor mínimo de 27.27 que con un valor mínimo de 20.38 (que es el 27.27% de 74.75, valor máximo hallado en la muestra). De esta manera obtenemos la figura 2.

Figura 2. IDP. Aristas podadas a 27.27




Fuente: elaboración propia con Gephi

Debe observarse en la figura 2 que no aparecen nodos aislados, al igual que en la figura 1, pero en este caso es una opción de representación, para no sobrecargar la figura, pero hay que dejar constancia que del grafo se han eliminado los nodos aislados porque tienen relaciones menores que 27.27 en el IDP.

En la red mostrada en la figura 2 se integran 23 nodos. Los conglomerados son más complejos que en la figura 1, aunque se pueden discernir con claridad los grupos de deportes (a la izquierda), juegos infantiles (abajo), juegos de mesa (arriba) y actividades, representadas por tres subconglomerados (a la derecha). Este último grupo de palabras no aparece en la figura 1, elaborada con Dispografo.

### *2.3 Índice de contiguidad de vocablos (ICV)*

En 2015, Guerra, Gómez y Basterrechea presentaron otra manera de cuantificar las asociaciones léxicas: el índice de contigüidad de vocablos (ICV). Al igual que Manjón-Cabeza (2008) pretendían resolver algunos problemas planteados por Dispografo. Conviene anotar que los autores no conocían la propuesta de Manjón-Cabeza (2008), de modo que la cronología científica es solo relativa y la aportación, por tanto, es independiente.

Proponen una fórmula logarítmica donde se priman las palabras de mayor aparición y se penalizan las palabras de aparición esporádica, lo que contribuye a paliar la anomalía que suponen los datos de escasa ocurrencia.

La fórmula propuesta es la siguiente:

$$\mathrm{ICV}\ (\mathrm{A,B}) - \mathrm{n}\ (\mathrm{A} \cap \mathrm{B})\ \log 2 \frac{\mathrm{n}\ (\mathrm{A} \cap \mathrm{B})}{\mathrm{f\ A\ f\ B\ N}}$$

donde:

$n_{(A\ y\ B)}$ es el número de veces que las palabras A y B aparecen contiguas en las encuestas.

$f_A$ es la frecuencia relativa de la palabra A, y $f_B$ es la frecuencia relativa de la palabra B.

N es el número de comparaciones totales entre las palabras de la lista.

Este cálculo también ha sido poco transitado. Junto con Guerra *et al.* (2015), solo conocemos su aplicación en Paredes *et al.* (2022).

Al no trabajar con contigüidades, los resultados han de ser necesariamente plasmados en una matriz de datos. Los autores no ofrecen ejemplos de matrices, pero nos hemos permitido calcular, al igual que en el caso del IDP, la matriz para las 37 palabras más disponibles de nuestra muestra. Por razones de espacio, ofrecemos en la tabla 2 la matriz resultante para los diez vocablos más disponibles.

Tabla 2. Matriz de IC para los diez vocablos más disponibles

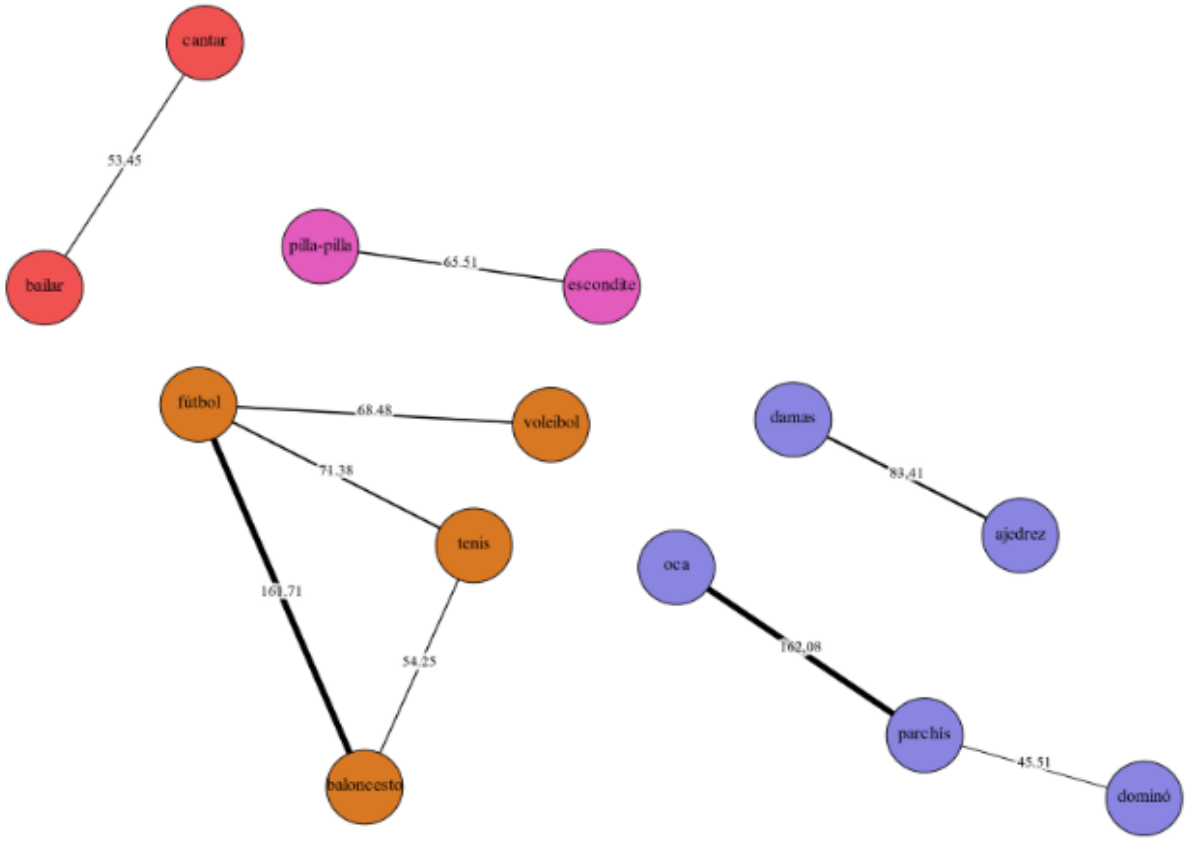
| | fútbol | parchís | escondite | baloncesto | cartas | comba | cine | ajedrez | tenis |
|---|---|---|---|---|---|---|---|---|---|
| fútbol | | | | | | | | | |
| parchís | -2.08 | | | | | | | | |
| escondite | 31.75 | 0 | | | | | | | |
| baloncesto | 161.71 | 0 | -0.81 | | | | | | |
| cartas | -1.23 | 20.51 | 14.04 | 15.9 | | | | | |
| comba | 1.81 | 0 | 14.62 | 9.86 | 0 | | | | |
| cine | 3.14 | 0.26 | 0.31 | 0 | 1.11 | 1.25 | | | |
| ajedrez | -0.42 | 41.16 | 0 | 0.78 | 28.78 | 1.25 | 1.92 | | |
| tenis | 71.38 | 0.26 | 0 | 54.25 | 1.11 | 0 | 1.92 | 0 | |
| dominó | 0 | 45.51 | 5.87 | 6.8 | 7.46 | 0 | 0 | 17.13 | 2.54 |

Fuente: elaboración propia

Para la comparación con Dispografo y con el IDP, hemos de ofrecer una red podada al 27.27%. En este caso se plantea el problema de la existencia de valores negativos, de modo que el valor máximo no es 162.08 (*oca-parchís*) sino la diferencia entre ese valor y el mínimo que es de -2,08 (*fútbol-parchís*), es decir, 164.16. De este modo, el 27.27% de 164.16 es 44.77, que debe ser el valor que tomemos como referencia en la poda de aristas, si queremos comparar con Dispografo y con IDP.

Tras el tratamiento de los datos del ICV con el programa Gephi obtenemos la figura 3.

Figura 3. IC. Aristas podadas a 44.77




Fuente: elaboración propia con Gephi

En la figura 3, al igual que en la figura 2, hemos optado por eliminar los nodos no integrados de la matriz porque son muchísimos (24) y dificultan la visión del gráfico. Se integran, por tanto, 13 nodos en la red y se pueden distinguir conglomerados muy sencillos de actividades (*bailar-cantar*), juegos infantiles (*pilla-pilla-escondite*), deportes, con cuatro representantes y juegos de mesa, abajo a la derecha, a su vez subdivididos en dos conglomerados: *ajedrez-damas*, por un lado, y *oca, parchís, dominó*, por otro.

## 3. Discusión

Los métodos presentados muestran fortalezas y debilidades. Las principales fortalezas de Dispografo creemos que son:

a) Usa una interfaz sencilla y amigable.

b) Permite el empleo directo de los datos originales de las encuestas de léxico disponible y sus filtros.

c) Posibilita elegir direccionalidad o no de las relaciones, podas de aristas, nodos, etc.

d) Los resultados se ofrecen tanto en tablas de adyacencias o contigüidades como en grafos sencillos de fácil interpretación.

Sin embargo, no todo son virtudes en este método de descubrimiento de asociaciones léxicas. Sorprende en buena medida que haya sido utilizado acríticamente en la mayor parte de estudios sobre asociaciones léxicas basadas en las encuestas de léxico disponible. Las principales debilidades parecen estar en:

a) Como cualquier método que se base exclusivamente en el conteo de contigüidades, no pondera la probabilidad de asociación de palabras, lo que desvirtúa los resultados. De este modo, la probabilidad de que aparezca la unión entre *fútbol* y *parchís* es muy alta porque son palabras actualizadas por muchos individuos –72 y 55 casos, respectivamente, en nuestra muestra–, mientras que la probabilidad de que aparezca la unión entre *televisión* y *leer* es mucho menor porque aparecen en menos individuos de la muestra –16 y 26 apariciones, respectivamente–. Parece necesario tener siempre en cuenta la aparición de la palabra de menor ocurrencia, que es la que determina la probabilidad de asociación. Obsérvese que en la figura 1 la unión entre *fútbol* y *baloncesto* es la máxima (22), pero la palabra que menos se actualiza de esa pareja es *baloncesto* que es anotada por 46 encuestados. Esto quiere decir que *fútbol* y *baloncesto* son unidas por los encuestados en el 47.83% de los casos posibles, mientras que la unión entre *parchís* y *oca*, que en la figura 1 es de 14, podría haberse hecho teóricamente en 27 ocasiones (número de encuestados que actualizan la palabra de menor aparición de la pareja, *oca*), por lo que han sido unidas por los encuestados en el 51.85% de los casos posibles. Sin embargo, los resultados de Dispografo apuntan a que es mucho mayor la unión de *fútbol-baloncesto* (22) que la de *parchís-oca* (14) y no parece ser así.

b) Dispografo no normaliza a una escala, es decir, en cada muestra habrá unos datos numéricos distintos que dependerán básicamente del número de informantes. En el caso que nos ocupa la escala oscila entre 0 –en el caso de un par de palabras que nunca aparecen contiguas, como *televisión* y *escondite*– a 22. Si la muestra hubiera sido otra, la escala hubiera sido otra y los conglomerados resultantes distintos. Esta falta de normalización impide comparar encuestas con distinto número de informantes, aunque en no pocos trabajos que usan Dispografo se ha ignorado este inconveniente.

c) Solo se tienen en cuenta las contigüidades absolutas, es decir, las que no están mediadas por otras palabras y, aunque es un modo de proceder defendible, no parece casar bien con el poder de almacenamiento de la memoria a corto plazo, que suele arrojar una media de siete unidades. Esto contribuye también a que las palabras integradas en la red, en el caso que nos ocupa, sean solo 11 (figura 1), frente a las 33 de IDP (figura 2).

d) La representación gráfica de Dispografo no permite que aparezcan palabras muy disponibles, pero que no presenten relaciones con otras. Si en Dispografo podamos las aristas y no actuamos sobre los nodos aparecerán en el grafo todas las palabras (recordemos que son 457 palabras distintas), lo que resulta en un gráfico enmarañado e inmanejable. Se hace necesario podar los nodos a 1, es decir, hacer que solo aparezcan las palabras que, al menos, tengan una conexión, como hemos hecho en la figura 1, de modo que palabras sin conexión no aparecen, independientemente de que sean altamente disponibles o no.

IDP resuelve dos debilidades principales de Dispografo: pondera la aparición de las palabras y proporciona datos normalizados a una escala.

La normalización permite la comparación entre resultados de muestras distintas, con distinto número de individuos, por ejemplo, entre jóvenes y adultos, o entre individuos con distinto nivel de escolarización, como hace, por ejemplo, Manjón-Cabeza (2010). Con los otros métodos, si la muestra no es idéntica, los resultados no se pueden comparar.

El IDP presenta dos debilidades, una salvable y otra insalvable. Una dificultad, creemos que menor, es que no dispone de interfaz gráfica propia, aunque sí se proporciona un script de cálculo (Manjón-Cabeza y Manjón-Cabeza 2021), basado en el entorno Python.

La mayor dificultad estriba en que se hace necesario trabajar solo con las palabras más disponibles. Esto es así porque los datos obtenidos con las encuestas de léxico disponible son intrínsecamente anómalos, es decir, se van volviendo anómalos según bajamos en el índice de disponibilidad y el papel de individuo se va acentuando, como ha señalado Paredes (2022). De modo que, si trabajáramos con todos los vocablos, pudiera ocurrir que una asociación de dos palabras que solo aparecieran una vez tuviera un IDP máximo al aplicarle la fórmula.

El ICV de Guerra *et al.* (2015) también surge para corregir las deficiencias de Dispografo, aunque no se tuvieron presentes las correcciones del IDP, con el que comparte alguno de los propósitos. Entre sus virtudes destaca que tiene en cuenta la frecuencia relativa de aparición de los vocablos y que su fórmula minimiza el valor de los índices de las palabras con baja aparición.

Al igual que en el caso anterior, los resultados deben ser gestionados por una matriz de datos que se trata posteriormente con programas de gestión de grafos. Como los datos están en forma de matriz, se pueden representar palabras altamente disponibles pero que no se integran en la red, aunque no lo hacemos en la figura 3 porque constituyen un número demasiado elevado.

A pesar del avance que supone el ICV no logra solucionar uno de los principales problemas de Dispografo, porque los datos no se normalizan a una escala, de modo que siguen sin poderse comparar encuestas con diferente número de informantes. En este aspecto, aunque la propuesta es posterior a Manjón-Cabeza (2008), no la supera.

También se habrá observado, a la vista de la figura 3, que hay relativamente pocos nodos integrados en la red. Esto es así porque al usar una fórmula logarítmica los extremos se disparan y la amplitud de valores es muy grande. Como debemos comparar tres métodos, al optar por representar solo las aristas que presenten un peso igual o superior al 27.27% en todos los casos, el valor resultante en ICV, dada la amplitud de rango, es muy grande (44.77) y deja fuera de la representación muchas asociaciones. Además, precisamente por ser una

escala logarítmica, el significado del porcentaje, inherentemente lineal, no está claro. De todas formas, resulta llamativo cómo, a pesar de usar una fórmula logarítmica especial, el resultado gráfico de ICV, si se comparan las figuras 1 y 3, es muy similar al de Dispografo.

Se pueden señalar otras debilidades. Dos de ellas saltan a la vista si se consultan y comparan las tablas 1 y 2, con ejemplos de matrices con datos de IDP e ICV, respectivamente. En la tabla 2, con los datos de ICV, hay bastantes ceros, mientras que no aparecen en la tabla 1 con datos de IDP. Esto es así porque el ICV, al igual que Dispografo, solo tiene en cuenta las contigüidades absolutas y no el resto de las palabras cercanas. Por otra parte, se puede observar en la tabla 2 que existen números negativos; por ejemplo, la relación que se establece entre *parchís* y *fútbol* es de -2.08. Esto constituye una paradoja porque si dos vocablos no aparecen juntos nunca, el resultado es 0, mientras que si aparecen juntos pocas veces el resultado es un número negativo, es decir, menor que cero. La complicación logarítmica se puede justificar si el extra de complejidad resulta en un tratamiento matemático más correcto, pero no parece ser así.

Asimismo, al igual que el IDP, no dispone de interfaz gráfica propia ni conocemos script de ayuda al cálculo del ICV.

Otra manera de comparar los datos es ofrecer los valores resultantes de las parejas que resultan más unidas con los tres métodos. Es lo que hacemos en la tabla 3 donde aparecen las dieciséis primeras parejas.

Tabla 3. Valores de las 16 primeras parejas según los tres métodos

| DISPOGRAFO | | | IDP | | | ICV | | |
|---|---|---|---|---|---|---|---|---|
| 1 | fútbol-baloncesto | 22 | 1 | voleibol-fútbol | 74.75 | 1 | oca-parchís | 162.08 |
| 2 | oca-parchís | 14 | 2 | oca-parchís | 67.52 | 2 | fútbol-baloncesto | 161.71 |
| 3 | fútbol-tenis | 11 | 3 | ajedrez-damas | 67.51 | 3 | ajedrez-damas | 83.42 |
| 4 | escondite-pilla-p. | 9 | 4 | fútbol-baloncesto | 52.89 | 4 | fútbol-tenis | 71.38 |
| 5 | ajedrez-damas | 8 | 5 | bailar-cantar | 50.00 | 5 | voleibol-fútbol | 68.48 |
| 6 | voleibol-fútbol | 8 | 6 | balonmano-fútbol | 49.62 | 6 | escondite-pilla-p | 65.51 |
| 7 | BALONCES.-TENIS | 8 | 7 | fútbol-tenis | 44.44 | 7 | BALONCESTO-TENIS | 54.25 |
| 8 | ESCOND.-FÚTBOL | 8 | 8 | escondite-pilla-p | 38.66 | 8 | cantar-bailar | 53.45 |
| 9 | parchís-dominó | 7 | 9 | tenis-balonmano | 38.51 | 9 | parchís-dominó | 45.51 |
| 10 | parchís-ajedrez | 7 | 10 | **escond.-escond.ing.** | 38.46 | 10 | televisión-leer | 44.35 |
| 11 | balonmano-fútbol | 6 | 11 | *cine-teatro* | 36.11 | 11 | balonmano-tenis | 43.29 |
| 12 | **cartas-ajedrez** | 5 | 12 | **parchís-damas** | 35.89 | 12 | ajedrez-parchís | 41.16 |
| 13 | bailar-cantar | 5 | 13 | televisión-leer | 35.41 | 13 | balonmano-fútbol | 41.00 |
| 14 | televisión-leer | 5 | 14 | **escondite-chapas** | 35.19 | 14 | *cine-teatro* | 34.64 |
| 15 | balonmano-tenis | 5 | 15 | **parchís-tres en r.** | 34.34 | 15 | ESCONDITE-FÚTBOL | 31.75 |
| 16 | **cartas-parchís** | 5 | 16 | parchís-dominó | 33.33 | 16 | **mus-dominó** | 31.32 |

Fuente: elaboración propia

En la tabla 3 aparecen, para cada método de medición de distancias, de izquierda a derecha, el rango de cada pareja léxica, la pareja en cuestión y el índice que proporciona cada metodología. En caso de que la pareja léxica sea privativa de una metodología aparece en negritas, mientras que si es compartida entre Dispografo e ICV se resalta en versalitas. Si es compartida por ICV e IDP se resalta en cursiva. No hay casos, entre estas dieciséis parejas más unidas, de pareja compartida por Dispografo e IDP y no por ICV.

El primer comentario es que se comparten la mayoría de las parejas: doce; aunque hay diferencias en el orden. También es comentable las amplitudes de los valores en Dispografo e ICV en comparación con IDP ya que, como se ha señalado, ni Dispografo ni ICV normalizan a una escala. Esto tiene una consecuencia evidente en el trabajo con grafos y es que las parejas que se integran en una red son mucho menores en Dispografo y en ICV. Con el corte que hemos establecido en el 27.27% el valor en Dispografo es 6, de modo que la última pareja que se integra en la red es la que tiene el rango 11 (*balonmano-fútbol*) –véase la tabla 3–; el valor en ICV es 44.77, de modo que la última pareja integrada en la red ocupa el rango 9 (*parchís-dominó*); mientras que en IDP el valor de corte establecido es 27.27, por lo que la última pareja dentro de ese valor es *tenis-voleibol*, que ocupa el rango 21, fuera de los límites presentados en la tabla 3.

En lo que toca a las diferencias de orden, creemos que podemos resumir las tendencias generales de los resultados ofrecidos en la tabla 3 comentando algunas asociaciones relacionadas con los deportes.

Una tendencia resaltable es que las parejas en las que interviene *fútbol* van perdiendo fuerza de unión de Dispografo a ICV y, finalmente, a IDP. Esto se ve por ejemplo en la pareja *fútbol-baloncesto* que pasa del rango 1 en Dispografo, al 2 en ICV y al 4 en IDP. O *fútbol-tenis* con los rangos 3, 4 y 7. Lo mismo sucede con *escondite-fútbol* que de ocupar el rango 8 en Dispografo, pasa al 15 en ICV y a no estar presente entre las dieciséis primeras parejas de IDP. Hay excepciones, como la pareja *balonmano-fútbol* y, sobre todo, la pareja *voleibol-fútbol*, que sube según el orden de rango, de 6 a 5 y 1.

La explicación de estos datos es sencilla: *fútbol* es la palabra que más aparece entre los encuestados, de modo que el método que no tiene en cuenta la frecuencia de las palabras tiende a dar más rango a las parejas donde está presente *fútbol*, ICV ofrece un tratamiento intermedio e IDP es el que más pondera la aparición. Esto último también explica que la primera pareja en IDP sea *voleibol-fútbol*: *fútbol* aparece 70 veces y voleibol 11 y nada menos que en ocho ocasiones aparecen inmediatamente contiguas y en otras dos mediatamente contiguas.

Otro binomio interesante, dentro del ámbito de los deportes, lo proporciona la pareja *baloncesto-tenis*, que aparece en séptimo lugar tanto en Dispografo como en ICV, pero que no está entre las 16 primeras de IDP. La explicación vuelve a ser similar a las anteriores: IDP toma en cuenta la probabilidad de coaparición. Resulta que baloncesto es dicho por 46 informantes y tenis por 31, con lo que podrían haberse relacionado en 31 ocasiones, pero solo lo hacen en 15 casos y, además, solo en ocho de manera inmediata, por eso queda fuera de las 16 primeras parejas. Explicaciones parecidas afectan a los distintos valores de otras muchas parejas presentes en la tabla 3.

Esta discusión sobre fortalezas y debilidades de los métodos evaluados se puede resumir en la tabla 4.

Tabla 4. Resumen de las características principales de los tres métodos

| | Dispografo | IDP | ICV |
|---|---|---|---|
| Gráfico integrado | Sí | No | No |
| Ponderación | No | Sí | Sí |
| Normalización | No | No | Sí |
| Linear | Sí | No | Sí |

Fuente: elaboración propia

## 4. Conclusiones

Tras el análisis de los resultados de los tres métodos evaluados tomando como base las mismas encuestas de léxico disponible, constatamos que todos los métodos tienen virtudes y defectos. Buena parte de las debilidades se debe al carácter intrínsecamente anómalo de la aparición de palabras poco frecuentes en algunos informantes en las encuestas de disponibilidad léxica.

Dispografo tuvo la virtud de abrir un campo fecundo y sigue teniendo la fortaleza de la interfaz amigable y la facilidad de manejo. Para un primer análisis exploratorio de los datos sigue siendo útil, siempre que comprendamos que el peso otorgado a las aristas es engañoso porque no tiene en cuenta la probabilidad de aparición de las palabras. Esta falta de ponderación y la falta de normalización a una escala impiden su uso para comparar muestras distintas o hacer un análisis estadístico riguroso. También debemos anotar en el haber de Dispografo que haya servido de punto de partida para el reciente programa LexPro, que aporta diferentes mejoras con respecto a su antecesor y que necesitará un análisis comparativo particular en trabajos futuros.

Aunque tenga debilidades, sobre todo el hecho de que no podemos trabajar directamente con todos los datos de las encuestas de léxico disponible sino solo con las palabras más disponibles, seguimos creyendo que el IDP de Manjón-Cabeza (2008) presenta una cuantificación de las asociaciones léxicas más cercanas a la realidad y una operatividad mayor. No solo tiene en cuenta la disponibilidad del léxico en la memoria a corto plazo, sino que, además, es fundamental ponderar la probabilidad de la unión entre palabras para no primar en exceso las uniones de las palabras con más frecuencia. Por otra parte, si queremos comparar entre muestras o submuestras con distinto número de informantes es obligatorio normalizar a una escala y eso solo lo hace de manera directa y lineal el IDP.

El índice de contigüidad (ICV) soluciona alguno de los problemas que plantea Dispografo, ya que tiene en cuenta la frecuencia de aparición de las palabras, y puede ayudar a tratar muestras con pocos datos. Al emplear una escala logarítmica, los valores de las distintas asociaciones son muy dispares, lo que implica que las palabras integradas en la red sean relativamente pocas si se hace necesaria, como es habitual, una poda según el peso de las aristas.

El ICV tampoco normaliza a una escala, lo que imposibilita su uso directo para comparar datos de distintas muestras. No obstante, hay que reconocer que con los datos básicos que proporciona la aplicación de la fórmula de IC podríamos normalizar a una escala común a varias muestras, en un paso posterior, aunque no de manera simple. Eso posibilitaría comparaciones entre distintas encuestas de léxico disponible.

En esta ocasión hemos evaluado métodos de determinación de asociaciones léxicas con binomios no dirigidos. Queda la duda de si la consideración de la dirección podría arrojarnos casos de direccionalidad intrínseca cuyo tratamiento estadístico seguramente deba implicar nuevos planteamientos. Es evidente que se trata de un aspecto de la disponibilidad léxica que debe exigir un tratamiento particular en posteriores trabajos.

**Antonio Manjón-Cabeza Cruz**
ORCID: 0000-0002-2112-3793
amanjoncabeza@ugr.es
Universidad de Granada

## Referencias bibliográficas

Aitchinson, Jean (1987): *Words in the mind. An introduction to the mental léxicon*, Oxford (Uk) Y Cambridge (USA): Blackwell.

Ávila Muñoz, Antonio Manuel (2022): "Algunas percepciones categoriales compartidas por preuniversitarios andaluces sobre la crisis del coronavirus y sus consecuencias. ¿Deberíamos preocuparnos? Un acercamiento desde la disponibilidade léxica y la centralidad léxica", *Tejuelo*, 35(3), pp. 17-42. DOI: https://doi.org/10.17398/1988-8430.35.3.17j

Ávila Muñoz, Antonio Manuel; Sánchez Sáez, José María; Odishelidze, Nana (2021): "Dispocen. Mucho más que un programa para el cálculo de la disponibilidad léxica", *ELUA*, 35, pp. 9-36. DOI: https://doi.org/10.14198/elua2021.35.1

Bastian, Mathieu; Heymann, Sebastien; Jacomy, Mathieu (2009): "Gephi: an open source software for exploring and manipulating networks", *ICWSM,* 8, pp. 361-362. DOI: https://doi.org/10.1609/icwsm.v3i1.13937

Blanco Correa, Oscar Elías; Salcedo Lagos, Pedro; Kotz Grabole, Grabiela Emilce (2020): "Análisis del léxico de las emociones: una aproximación desde la disponibilidad léxica y la teoría de los grafos léxicos", *Lingüística y Literatura*, 78, pp. 55-83. DOI: https://doi.org/10.17533/udea.lyl.n78a03

Cañizal Arévalo, Alva Valentina (1991): "Redes semánticas y disponibilidad léxica en el español de escolares mexicanos", César Hernández Alonso y otros (eds.), *El español de América*, II, Valladolid: Junta de Castilla y León, pp. 631-641.

Collins, Allan; Loftus, Elizabeth (1975): "A spreading-activation theory of semantic processing", *Psychological Review*, 82, pp. 407–428. DOI: https://doi.org/10.1037/0033-295X.82.6.407

Collins, Allan; Quillian, Ross (1969): "Retrieval time from semantic memory", *Journal of Verbal Learning and Verbal Behaviour*, 8, pp. 240-247. DOI: https://doi.org/10.1016/s0022-5371(69)80069-1

Echeverría, Max S.; Vargas, Roberto; Urzúa, Paula; Ferreira, Roberto (2008): "DispoGrafo: una nueva herramienta computacional para el análisis de relaciones semánticas en el léxico disponible", *RLA, Revista de Lingüística Teórica y Aplicada*, 46(1), pp. 81-91. DOI: http://dx.doi.org/10.4067/S0718-48832008000100005

Ferreira, Roberto; Echeverría, Max S. (2010): "Redes semánticas en el léxico disponible de inglés L1 e inglés LE", *Onomázein*, 21, pp. 133-153.

Galloso Camacho, María Victoria (2003): "El léxico disponible de los estudiantes preuniversitarios en Salamanca", *Salamanca: Revista de Estudios*, 50, pp. 201-224.

Gómez Devís, María Begoña (2019): "A propósito de las redes semánticas en el léxico disponible de escolares de primero de Educación Primaria", *Ogigia. Revista Electrónica de Estudios Hispánicos*, 25, pp. 165-183. DOI: https://doi.org/10.24197/ogigia.25.2019

Gómez Devís, María Begoña (2021): "Disponibilidad léxica en niños de 6 años. Alcance y proyección didáctica del corpus léxico infantil", *Cultura, Lenguaje y Representación*, 25, pp. 169-181. DOI: https://doi.org/10.6035/CLR.2021.25.10

Gómez Devís, María Begoña; Cepeda Guerra, Milko (2022): "Bases para la enseñanza del léxico: Mecanismos de asociación y configuración de redes en el léxico disponible infantil", *Tejuelo,* 35(3), pp. 105-134. DOI: https://doi.org/10.17398/1988-8430.35.3.105

Gómez-Devís, María Begoña; Gómez Molina, Juan Ramón (2022): "La investigación en disponibilidad léxica y su proyección en el ámbito de la didáctica de la lengua", *Tejuelo,* 35(3), pp. 1-15. DOI: https://tejuelo.unex.es/article/view/4369

Gómez Molina, Juan Ramón (2009): "Una aplicación psicolingüística de la disponibilidad léxica: la categoría nocional 'animales'", Montserrat Veyrat Rigat; Enric Serra Alegre (eds.), *La Lingüística como reto epistemológico y como acción social. Estudios dedicados al profesor Ángel López con ocasión de su sexagésimo aniversario,* Madrid: Arco-Libros, pp. 1047-1060.

Guerra Salas, Luis; Gómez Sánchez, María Elena; Basterrechea Salido, Martín (2015): "Cuantificación y representación de las asociaciones léxicas en las listas de disponibilidad: el índice de contigüidad de los vocablos", *Lingüística Española Actual,* 37(2), pp. 265-277.

Henríquez Guarín, María Clara; Mahecha Mahecha, Viviana; Matéus Ferro, Gerald Eduardo (2016): "Análisis de los mecanismos cognitivos del léxico disponible del cuerpo humano a través de grafos", *Lingüística y Literatura,* 69, pp. 229-251. DOI: https://doi.org/10.17533/udea.lyl.n69a10

Hernández Muñoz, Natividad (2006): *Hacia una teoría cognitiva integrada de la disponibilidad léxica: el léxico disponible de los estudiantes castellano-manchegos*, Salamanca: Univ. de Salamanca.

Hernández Muñoz, Natividad; Izura González, Cristina; Tomé Cornejo, Carmela (2014): "Cognitive Factors of Lexical Availability in a Second Language", Rosa María Jiménez Catalán (ed.), *Lexical Availability in English and Spanish as a Second Language*, Dordrecht: Springer, pp. 169-186. DOI: https://doi.org/10.1007/978-94-007-7158-1_10

Hernández Muñoz, Natividad; Tomé Cornejo, Carmela (2017): "Léxico disponible en primera y segunda lengua: bases cognitivas", Florencio del Barrio de la Rosa (coord.), *Palabras Vocabulario Léxico. La lexicología aplicada a la didáctica y a la diacronía,* Venecia: Edizioni Ca Foscari, pp. 99-122.

Hernández Muñoz, Natividad; Tomé Cornejo, Carmela; López García, Miguel; Bartol Hernández, José Antonio (2023): *Manual de LexPro*. <https://dispogram.usal.es/> (Consulta: 17/01/ 2024).

Jenkins, James J. (1970): "The 1952 Minnesota word association norms", Leo Postman y Geoffrey Keppel (eds.), *Norms of word associations*, New York: Academic Press, pp. 1-38. DOI: https://doi.org/10.1016/B978-0-12-563050-4.50004-2

López González, Antonio María (2014): "La estructura interna del léxico disponible en español como lengua extranjera (ELE) de los preuniversitarios polacos", *Studia Romanica Posnaniensia*, 41(1), pp. 45-61. DOI: 10.7169/strop2014.411.004

López González, Antonio María (2016): "Las asociaciones léxicas en el léxico disponible en lengua materna y en lengua extranjera". *Tonos Digital: Revista Electrónica de Estudios Filológicos,* 31, pp. 1-26. DOI: http://hdl.handle.net/10201/50323

Mahecha Mahecha, Viviana; Mateus Ferro, Gerald (2020): "Tipología de mecanismos cognitivos y lingüísticos que caracterizan el léxico disponible", *Círculo de lingüística aplicada a la comunicación*, 82, pp. 165-178. DOI: https://doi.org/10.5209/clac.68971

Manjón-Cabeza Cruz, Antonio (2008): "Redes semánticas naturales en escolares de 5 a 16 años: los colores", *Docencia e Investigación*, 33, pp. 127-146.

Manjón-Cabeza Cruz, Antonio (2009): "Léxico disponible de los juegos y diversiones en Toledo", *Docencia e Investigación*, 34, pp. 127-144.

Manjón-Cabeza Cruz, Antonio (2010): "Aproximación a la organización semántica del léxico sobre juegos y diversiones", *ELUA,* 24, pp. 199-224. https://doi.org/10.14198/ELUA2010.24.08

Manjón-Cabeza Córdoba, Antonio; Manjón-Cabeza Cruz, Antonio (2021): "PSILIN. Script to measure word distance for sociolinguistic studies", *Zenodo.* Recuperado de: https://zenodo.org/record/5783307

Paredes García, Florentino (2006): "Aportes de la disponibilidad léxica a la psicolingüística: una aproximación desde el léxico del color", *Lingüística*, 18, pp. 19-55. DOI: http://hdl.handle.net/10017/41079

Paredes García, Florentino (2022): "Qué nos dice la disponibilidad léxica acerca de los individuos: el índice de descentralización léxica", Carmen Díaz Ayalón (coord.), *Studia philologica: in honorem José Antonio Samper,* Madrid: Academia Canaria de la Lengua y Arco/Libros-La Muralla, pp. 793-812.

Paredes García, Florentino; Guerra Salas, Luis; Gómez Sánchez, María Elena (2022): *Léxico disponible de los jóvenes preuniversitarios de la Comunidad de Madrid*, Alcalá de Henares: Universidad de Alcalá.

Podhaská, Daniela (2023a): *Acomodación léxica de la comunidad mexicana en Granada (España)*, Univ. De Olomouc/Univ. De Granada.

Podhaská, Daniela (2023b): "Redes semánticas entre los sinónimos geolectales en el centro de interés de la Ropa" (en prensa).

Quillian, M. Ross (1968): "Semantic Memory", Marvin Lee Minsky (ed.), *Semantic Information Processing,* Cambridge (Mass.): MIT Press., pp. 216-270.

Sánchez-Saus Laserna, Marta (2019): *Centros de interés y capacidad asociativa de las palabras*, Sevilla: Editorial Universidad de Sevilla.

Sánchez-Saus Laserna, Marta (2022): "Redes semánticas, léxico disponible y didáctica del vocabulario en ELE: un análisis por niveles de español", *Tejuelo*, 35(3), pp. 167-204. DOI: https://doi.org/10.17398/1988-8430.35.3.167

Santos Díaz, Inmaculada C. (2017): "Organización de las palabras en la mente en lengua materna y lengua extranjera", *Pragmalingüística,* 25, pp. 603-617. DOI: https://revistas.uca.es/index.php/pragma/article/view/2406

Stayvers, Mark; Tenenbaum, Joshua B. (2005): "The large-scale structure of semantic networks: statistical analisys and a model of semantic growth", *Cognitive Science,* 29(1), pp. 41-78. DOI: https://bit.ly/2Z4knR2.