Boletín de Filología, Tomo LI
Número 1 (2016): 39-92

Medidas de disponibilidad léxica: comparabilidad y normalización

Francisco Javier Callealta Barroso[*]
Universidad de Alcalá, España

Diego Javier Gallego Gallego
Universidad Europea de Madrid, España

Resumen

Después de hacer una revisión pormenorizada de las fórmulas matemáticas utilizadas en los estudios de disponibilidad léxica, hemos encontrado algunos inconvenientes en los rangos de valores que pueden tomar las fórmulas propuestas originalmente que dificultan su comparabilidad al aplicarse en investigaciones que presenten diferentes tamaños de muestras. En concreto, nos referimos en primer lugar a la fórmula para el cálculo de la disponibilidad de López Chávez y Strassburger Frías (1987, 1991) que presenta problemas para medir la disponibilidad cuando las palabras se acercan al límite inferior –que es cero–, es decir, cuando son "muy poco disponibles". Por otro lado, en la fórmula propuesta por Ávila Muñoz y Sánchez Sáez (2010) que mide el grado de compatibilidad de los vocablos, a pesar de que un vocablo alcance el máximo grado de compatibilidad,

[*] Para correspondencia, dirigirse a: Francisco Javier Callealta Barroso (franciscoj.callealta@uah.es), Facultad de Ciencias Económicas y Empresariales, Universidad de Alcalá, Antiguo Colegio de Mínimos, Plaza de la Victoria, 2, 28802, Alcalá de Henares (Madrid), España; o Diego Javier Gallego Gallego (diegojavier.gallego@uem.es), Calle Holanda, 14, 28880, Meco (Madrid), España.

su valor nunca llega a ser 1, que es teóricamente el límite superior de referencia en dicha fórmula. Es por esto que en el presente trabajo, además de analizar en profundidad los casos descritos anteriormente, hemos propuesto diferentes opciones de fórmulas que, por un lado, permiten una mayor precisión matemática, y por otro, permiten la comparabilidad entre distintos estudios que utilicen muestras de diferente tamaño y características.

Palabras clave: disponibilidad léxica, léxico disponible, índice de disponibilidad léxica, índice de compatibilidad léxica, fórmulas matemáticas, palabras, vocablos.

Measures of lexical availability: comparability and standardization

Abstract

After a deep review of the mathematical formulas used in studies of lexical availability, we have found some inconvenient in the range of value in the original proposal formules which make difficult to compare researches which have different size of samples. Mainly, we refer to the formula for calculating lexical availability by Lopez Chavez and Strassburger Frías (1987, 1991) which shows problems to measure the rate of availability when words are close to the lower limit (0), that means, they are "very little available". On the other hand, the formula proposed by Ávila Muñoz and Sánchez Sáez (2010), which measures the degree of compatibility of term, although a word reaches the maximum compatibility, its value never reach 1, theoric upper reference limit in this formula. In this work we analyze the above described cases, and we propose different formulas which, allow greater mathematical precision, and also, allow comparability between different studies using samples of different sizes and characteristics.

Keywords: lexical availability, lexical available, availability lexical index, lexical compatibility index, mathematical formulas, words, terms.

Recibido: 29/12/15 Aceptado: 15/03/16

1. ANTECEDENTES

Las primeras investigaciones sobre disponibilidad léxica desarrolladas en Francia por el grupo de Gougenheim *et alii* (1956, 1964) consideraban la "frecuencia" de aparición de las palabras dentro de los listados como el único factor ponderador de la disponibilidad léxica. Teniendo en cuenta esta medida, las unidades léxicas se ordenaban por rangos, estableciendo una equivalencia entre el índice de disponibilidad[1] y la frecuencia alcanzada. Otros investigadores como Mackey (1971), Dimitrijévic (1969), López Morales (1973, 1978 y 1979) y Azurmendi Ayerbe, (1983), siguiendo el modelo francés, utilizaron el cómputo de la frecuencia para calcular el índice de disponibilidad.

Desde entonces, muchos investigadores se interesaron por encontrar una fórmula matemática que permitiera valorar no solo la frecuencia, sino también la posición que ocupaban las palabras en las listas a la hora de calcular tal índice. Es el caso de Lorán y López Morales (1983), quienes desarrollaron fórmulas para calcular el índice de disponibilidad de una palabra, en las cuales introducían un factor de ponderación decreciente con la posición que la palabra ocupaba en la lista y de tipo potencial, λ^{i-1}, siendo λ un valor inferior a la unidad (en los primeros trabajos era λ=0.90), que se aplicaría a la frecuencia alcanzada por la palabra en cada posición "i" (i=1,2,…). De esta manera, a la primera posición se le asigna un factor de ponderación igual a uno (ya que $\lambda^{1-1}=\lambda^{0}=1$), y a partir de ella, a cada posición "i" se le asocia el factor de ponderación λ^{i-1}, menor que uno y progresivamente decreciente a medida que se avanza en las posiciones "i" en que se pudieran haber mencionado las palabras. Concretamente, las fórmulas introducidas por estos autores, dependiendo de si las listas de palabras de los informantes tenían diferentes tamaños o no, eran las siguientes:

$$d(p)=\sum_{i=1}^{n} \lambda^{i-1} x_{pi} \quad \text{, siendo } x_{pi}=\frac{f_{pi}}{N_i} \qquad \text{vs.} \qquad d(p)=\sum_{i=1}^{n} \lambda^{i-1} \frac{f_{pi}}{N_1} \qquad [1]$$

en donde

n = máxima posición alcanzada por las palabras.

x_{pi} = frecuencia relativa de la palabra p en la posición "i".

[1] El índice de disponibilidad es un valor mediante el cual es posible ordenar las palabras según la frecuencia y el orden de aparición –el grado de "inmediatez"– con que los hablantes actualizan dichas palabras.

f_{pi} = frecuencia absoluta de la palabra *p* en la posición "i".
N_i = número de informantes cuyas listas contenían la posición "i".

Sin embargo, al aplicar la correspondiente fórmula en investigaciones posteriores que trabajaban con listas de distintos tamaños (Román 1985; Mena Osorio 1986; Justo Hernández 1986; Echeverría *et alii* 1987; Cañizal Arévalo 1987), se observó que podía ocurrir que, a partir de cierta posición, la fórmula asignase un coeficiente de mayor valor –de acuerdo a su posición– a palabras que podían ocupar posiciones inferiores en los listados de disponibilidad, adquiriendo así una importancia irregular y exagerada.

Para calcular el índice de disponibilidad léxica de un centro para un grupo de informantes, Butrón (1991) presenta en su trabajo *Nuevos índices de disponibilidad léxica*, cuatro fórmulas matemáticas que permitirían establecer diferentes órdenes entre los centros. La primera fórmula tiene en cuenta únicamente el rango o número total de palabras del grupo para ese centro; la segunda fórmula tiene en cuenta las habilidades de los informantes, más que del grupo, ya que calcula el promedio de los rangos de palabras de los informantes para dicho centro; la tercera fórmula permitiría conocer la frecuencia promedio de repetición de las diferentes palabras en un centro y, finalmente, dejando a un lado la variedad de las palabras, propone una cuarta fórmula que evalúa la suma de las disponibilidades de las palabras mencionadas en el centro.

Adicionalmente, interesada por disponer de una metodología que permita calcular el grado de disponibilidad léxica que tiene un informante dentro de un grupo, Butrón (1991) propone de nuevo diferentes fórmulas con las cuales se podría obtener el deseado índice de disponibilidad léxica individual. Estas fórmulas tienen en cuenta diversos factores, como la suma de las disponibilidades de todas las palabras ofrecidas por un informante para un centro, el número de palabras mencionadas en un centro, el tamaño promedio de los centros, etc. Como su propia autora señala, el propósito de todas estas nuevas fórmulas es ofrecer el mayor número de posibilidades para calcular el índice de disponibilidad de un informante o de un grupo de informantes, y advierte que pueden resultar más o menos apropiadas dependiendo de las características de cada investigación.

A pesar del gran abanico de posibilidades que ofrece Butrón, estas nuevas fórmulas no logran medir de una manera exacta el grado de ordenación de las palabras que en un mismo centro de interés alcanzan las respuestas de los informantes; o utilizando las palabras de López Morales, estos nuevos intentos "lograron controlar bastante el desajuste, pero no eliminarlo" (1999: 18). La razón para que esto suceda puede residir en que todas esas fórmulas pivotan sobre la medición correcta de las disponibilidades de los centros de

interés para los informantes, y esta, a su vez, sobre la correcta medición de la disponibilidad de las palabras que los informantes emplean en cada centro. Sin embargo, como se ha dicho anteriormente, las alternativas presentadas hasta entonces para medir dicha disponibilidad de las palabras presentaban serios inconvenientes.

1.1. Índice de la Disponibilidad Léxica de una Palabra (IDLP)

No fue sino hasta la presentación de la fórmula de López Chávez y Strassburger Frías (1987, 1991) que se superaron muchas de las dificultades de las fórmulas anteriores, permitiendo medir más adecuadamente el índice de disponibilidad de una unidad léxica desde el punto de vista lingüístico. La formulación de este nuevo Índice de Disponibilidad Léxica de una Palabra (en adelante IDLP) es la siguiente:

$$D(P_j) = \sum_{i=1}^{n} e^{-2.3\left(\frac{i-1}{n-1}\right)} \frac{f_{ji}}{I_1} \qquad [2]$$

en donde

- n = máxima posición alcanzada en el centro de interés en esta encuesta.
- i = número de la posición en cuestión.
- j = índice de la palabra tratada.
- e = número de Euler, o constante de Napier (2,718281828459045…)
- f_{ji} = frecuencia absoluta de la palabra j en la posición i.
- I_1 = número de informantes que participaron en la encuesta.
- $D(Pj)$ = disponibilidad de la palabra j.

Esta nueva fórmula desarrollada por López Chávez y Strassburger Frías (1987, 1991) proporciona una medición más fiable de la disponibilidad de una palabra de un centro, a partir de la suma de diferentes términos que se construyen como producto de factores construidos a partir de:

a) La frecuencia absoluta con que fue dicha la palabra en cada posición;
b) El número de informantes que participan en la encuesta;
c) El número *n,* o máxima posición alcanzada por las palabras en la encuesta en el centro de interés analizado;
d) Las posiciones en que fue dicha la palabra por los informantes que la mencionaron.

Como sus propios autores señalan, el verdadero factor que se emplea para ponderar la posición de mención de una palabra en el cálculo de su disponibilidad es el número *e* elevado al exponente *-2.3* que aparece en la fórmula. Como hemos mencionado más arriba, las fórmulas anteriores utilizaban una factor de ponderación lambda (λ) elevado a la potencia "i-1" (siendo *i* la posición que alcanza la frecuencia de la palabra) que funcionaba razonablemente bien con listados de igual tamaño pero que podía perder el poder discriminador a partir de cierta posición cuando las listas eran de tamaños diferentes. El nuevo factor de ponderación propuesto por López Chávez y Strassburger Frías (1991: 93) permite combinar sin distorsión la frecuencia y la posición de emisión de cada vocablo, manteniendo una ponderación variable entre 1 y 0.1, independientemente del número de informantes que participen en la prueba, de la extensión que presente los listados producidos por cada informante, del número de informantes que lleguen a cada posición y de la frecuencia de aparición del vocablo.

1.2. Índice de Disponibilidad Léxica Individual (IDLI)

Además del índice de disponibilidad léxica de una palabra (IDLP), López Chávez y Strassburger Frías (1991) presentan dos nuevos índices que pretenden calcular el grado de aportación de un individuo dentro de los listados generales de disponibilidad. El primero de ellos es el Índice de Disponibilidad Léxica Individual (en adelante IDLI), y parte de varios supuestos: 1) Si un informante X produce un listado mayor que el del informante Y, y si además el listado de Y está contenido en el de X y ambos listados responden a un mismo ordenamiento, entonces el IDLI del informante X deberá ser mayor que el IDLI del informante Y; 2) Si los informantes X y Z producen un listado de palabras de idéntica longitud pero las palabras aportadas por X tienen un índice de disponibilidad mayor que las del informante Z, entonces el IDLI del informante X será mayor que el IDLI de Z; 3) Si el informante X produce un listado de palabras que incluye las mismas palabras producidas por el informante S pero el listado de X tiene una mayor correlación con el listado general del léxico disponible que el listado de S, entonces el IDLI del informante X deberá ser mayor el IDLI del informante S. Para calcular este índice (IDLI), López Chávez y Strassburger Frías (1991) utilizan una fórmula análoga a la empleada para calcular el IDLP, con algunas variaciones que permitan reflejar el grado de participación de un individuo dentro de los listados generales de léxico disponible y cumplir así las condiciones que hemos numerado anteriormente. La formulación del índice de disponibilidad léxica individual (IDLI) es la siguiente:

$$D(S_i) = \frac{1}{k} \sum_{j=1}^{n} d\left(P_{ij}\right) \cdot e^{-2.3\left(\frac{j-1}{n-1}\right)} \quad [3]$$

en donde

$D(S_i)$ = disponibilidad del sujeto i
$d(P_{ij})$ = disponibilidad de la palabra respondida por el sujeto i en la posición j
n = máximo número de respuestas del centro de interés en cuestión
k = constante para ajustar las calificaciones

Esta nueva fórmula, que puede leerse como una suma ponderada de los índices de disponibilidad léxica de las palabras empleadas por un sujeto en un centro de interés dividida por una constante (k) que permite ajustar las calificaciones, no tiene en cuenta los diferentes centros de interés considerados en cada investigación, sino que se refiere a un centro determinado. Así pues, para calcular la disponibilidad léxica de un informante teniendo en cuenta los diferentes centros, se efectúa este cálculo para cada centro de interés y se suman sus resultados. Sin embargo, este método puede llevar a apreciaciones erróneas en el grado de participación de cada informante, ya que a mayor número de centros de interés, mayor será el IDLI[2].

1.3. Índice de la Competencia de la Disponibilidad Léxica del Individuo (ICDLI)

Por otro lado, López Chávez y Strassburger Frías (1991) proponen otra fórmula para medir el Índice de la Competencia Léxica Individual (en adelante ICDLI). Este nuevo índice permitiría conocer la competencia que un individuo tiene del léxico disponible de un grupo o comunidad. Partiendo de la constatación de que tanto las palabras como los individuos presentan índices de disponibilidad diferentes en los listados, López Chávez y Strassburger Frías (1991) se plantean que la actuación de cada informante refleja una competencia diferente. Igualmente se preguntan si

[2] En un intento anterior para calcular el índice de disponibilidad léxica del informante, Butrón (1987) había propuesto una ponderación de cada uno de los centros teniendo en cuenta el promedio de palabras mencionadas y el número de palabras de cada centro.

cuando un informante produce únicamente palabras con un alto índice de disponibilidad tiene necesariamente mayor competencia léxica disponible, o si por el contrario, el informante que solo produce palabras con bajo índice de disponibilidad tiene por oposición una menor competencia que el primero.

Para calcular el ICDLI se ordena primero el listado de palabras en orden decreciente de acuerdo al IDLP y se divide en grupos o rangos atendiendo a la "proporción acumulada"; luego se agrupa la producción de cada informante según los grupos o rangos de palabras del paso 1, así como el IDLP y la frecuencia absoluta. A continuación se aplica una fórmula de dispersión para obtener un factor de dispersión para cada informante, según el IDLI y la frecuencia. Finalmente, se multiplica el IDLI por la dispersión y se obtiene el Índice de la Competencia Léxica Individual (ICDLI).

1.4. Índice de Compatibilidad Léxica de un Vocablo (COMP)

Después de realizar una revisión del modelo clásico utilizado para el cálculo del índice de disponibilidad léxica, Ávila Muñoz y Sánchez Sáez (2010) plantean algunas limitaciones presentes en las fórmulas utilizadas tanto por López Chávez y Strassburger Frías (1987, 1991) como en la de sus predecesores, al señalar que dichos intentos de análisis estadístico se refieren a las propiedades de las palabras y no a la de los hablantes, por lo que según estos autores, "[…] no pasen de ser meras descripciones de los acontecimientos que se pretenden estudiar" (Ávila Muñoz y Sánchez Sáez, 2010: 53)[3]. Igualmente señalan que desde sus comienzos, los estudios de disponibilidad léxica han dado mayor prioridad a la cuantificación de las palabras producidas por los informantes para la obtención de los listados de disponibilidad –a través de la creación de fórmulas matemáticas basadas principalmente en el factor frecuencia–, que a la búsqueda y creación de modelos teóricos que permitan una representación de la realidad que se quiere cuantificar.

Interesados precisamente en desarrollar un modelo que permita una mayor aproximación a la realidad representada en los estudios de disponibilidad

[3] Como excepción a los modelos de análisis para el cálculo del índice de disponibilidad que solo tienen en cuenta las propiedades de las palabras, Ávila Muñoz y Sánchez Sáez (2010) señalan el intento de López Chavéz y Strassburger Frías (1987, 1991) para calcular la participación que tiene cada individuo en la obtención de los resultados generales de disponibilidad a través de la creación del ICDLI al que ya nos hemos referido más arriba en el apartado (cf. § 1.3).

léxica, Ávila Muñoz y Sánchez Sáez (2010) proponen la utilización de un nuevo enfoque matemático –alternativo al enfoque clásico utilizado por López Chávez y Strassburger Frías (1987, 1991)–, que permita describir con mayor exactitud los procesos de asociación de las palabras producidas por los informantes durante las pruebas de disponibilidad. Una de las condiciones de esta nueva propuesta o enfoque es que sea compatible con los resultados obtenidos a través de los modelos matemáticos anteriores siempre y cuando dichos resultados sean válidos y aceptables, ya que presentan similitudes con los datos obtenidos a partir del nuevo enfoque, que además ofrece soluciones a algunos problemas e incoherencias que no son posibles con las herramientas y métodos de cálculo utilizados hasta el momento.

Partiendo de la teoría de los prototipos (*Prototypes theory:* Wittgenstein, 1953; Rosch, 1978; Lakoff, 1987), Ávila Muñoz y Sánchez Sáez (2010) establecen que todos los centros de interés giran alrededor de un prototipo (o núcleo temático) determinado por el concepto que representa cada centro de interés. Según esta propuesta, las palabras mencionadas por los informantes en cada centro de interés constituyen una red léxica y están determinadas por su relación con el núcleo prototípico de ese centro: mientras más cercanas al núcleo, mayor será su grado de "accesibilidad" y a medida que se alejan, menor accesibilidad. Desde este enfoque, el concepto de accesibilidad de una palabra es equivalente al concepto de disponibilidad, y al igual que este último, es un intento por reflejar la relación existente entre las palabras y la red léxica o campo semántico al que pertenecen.

Una vez que han establecido que los centros de interés están determinados por un núcleo prototípico y que las palabras producidas en dicho centro pueden ser más o menos "accesibles" (disponibles), Ávila Muñoz y Sánchez Sáez (2010) recurren al concepto matemático de *conjunto difuso* para intentar interpretar los resultados obtenidos a partir de las pruebas de disponibilidad. Como sus propios autores señalan, los conjuntos difusos "son una generalización de la teoría de los conjuntos en la que, en lugar de la pertenencia de los elementos, se considera la compatibilidad de estos con el concepto representado por el conjunto" (Ávila Muñoz y Sánchez Sáez, 2010: 60).

A diferencia de la teoría clásica de los conjuntos, que establece la pertenencia o no de un elemento dentro de un conjunto, el concepto de los conjuntos difusos permite utilizar el grado de compatibilidad[4] entre

[4] Ávila Muñoz y Sánchez Sáez (2010) establecen una correspondencia entre el concepto de "compatibilidad" y el concepto de "accesibilidad", que como hemos señalado más atrás, es equivalente al concepto de "disponibilidad".

los diferentes elementos que conforman el conjunto y el conjunto que representan. Sin embargo, la importancia de este nuevo enfoque según sus autores, no está en el proceso de reformulación de los modelos de análisis utilizados en los estudios de disponibilidad, sino en contar con una herramienta matemática perteneciente a un marco teórico establecido y contrastado como es el de los conjuntos difusos.

El concepto de compatibilidad o accesibilidad de los vocablos propuesto por Ávila Muñoz y Sánchez Sáez (2010) permite medir la compatibilidad o "centralización" de los vocablos respecto al conjunto de datos que constituyen el núcleo prototípico (o central) de un centro de interés. Para calcular el Índice de Compatibilidad Léxica de un Vocablo (en adelante COMP), Ávila Muñoz y Sánchez Sáez se valen de una herramienta propia de la teoría de los conjuntos difusos, FEV (*Fuzzy Expected Value*), que permite determinar el valor de compatibilidad característico del conjunto difuso. Esta herramienta permite establecer los límites de caracterización de los valores de pertenencia de los elementos dentro de un conjunto e identificar a partir de parámetros objetivos de corte, aquellos elementos que son muy característicos o poco característicos respecto al conjunto.

Para determinar el grado de compatibilidad de un término, se crea primero un conjunto difuso que represente la participación de cada individuo dentro de los listados de disponibilidad léxica, partiendo de una valoración inicial de cada vocablo mediante la fórmula:

$$t = \frac{k}{n} \qquad [4]$$

donde *n* es la posición que ocupa en la lista cada término y *k* es una constante para cada problema. Una vez establecido el conjunto difuso para cada individuo, se construye el modelo colectivo tomando como referencia el valor de nula disponibilidad para todos los términos. Para ello se incorpora al modelo colectivo, el valor de cada término proporcionado en el espectro individual, utilizando una ley que premie su representatividad si los valores proporcionados por los individuos tienen altos índices de disponibilidad.

$$a + b - a \cdot b$$

La reiteración de esta ley, empleada para el cálculo de la compatibilidad a partir de las premisas y explicaciones dadas por Ávila Muñoz y Sánchez Sáez (2010), nos conduce a la siguiente formulación final que hemos empleado para el cálculo del Índice de Compatibilidad Léxica de un Vocablo (COMP):

$$COMP_k = 1 - \prod_{i=1}^{n} \left(1 - \left(\frac{k}{i}\right)\right)^{F_i} \qquad [5]$$

donde *n* es el número de la máxima posición que puede alcanzar un vocablo en las menciones que hacen de él los informantes; *i* representa a cada una de las posibles posiciones de mención del vocablo; F_i, la frecuencia absoluta de aparición del vocablo en la posición i-ésima, y *k* es la constante introducida por los autores en la valoración inicial de un vocablo que aparece en una posición ($t=k/i$).

1.5. Índice de Descentralización Léxica del Informante (IDD)

Dentro de la propuesta matemática alternativa al enfoque clásico o tradicional, Ávila Muñoz y Sánchez Sáez (2010) presentan una nueva fórmula para calcular la capacidad léxica de cada informante, que se integra y complementa al modelo colectivo para el cálculo de la compatibilidad (COMP). La nueva propuesta busca medir la capacidad léxica individual tomando en consideración la "descentralización" de los términos producidos por cada informante, para producir el Índice de Descentralización Léxica del Informante (en adelante IDD), que es lo contrario al índice de compatibilidad. Según los autores, una vez que se activa el mecanismo de asociación léxica a partir del prototipo de un centro de interés, el informante actualiza aquellas palabras más cercanas al concepto propuesto por el centro (aquellas palabras más cercanas al núcleo prototípico tendrían un alto índice de compatibilidad). A medida que se aleja de aquellas palabras "comunes" al núcleo (que también podemos llamar "centrales"), aparecen palabras menos disponibles, es decir, menos "compatibles" con el núcleo propuesto y por tanto "descentralizadas", lo que demostraría una mayor capacidad léxica del informante.

Partiendo de la idea de que el léxico mencionado en un centro de interés tiende a tener una estructura similar para un grupo de personas, ya sea por razones sociales o culturales compartidas, Ávila Muñoz y Sánchez Sáez (2010) plantean como hipótesis de partida que si un informante actualiza un vocabulario más descentralizado o específico[5], aquel demostraría una mayor capacidad léxica en la medida que su léxico se aleja del núcleo o prototipo:

[5] Al igual que el Índice de compatibilidad (COMP) establecía una correspondencia entre los conceptos de "accesibilidad" y "centralización" de los vocablos, el Índice de descentralización (IDD) establece una similitud entre los conceptos de "especificidad" o "descentralización" del vocabulario.

> Puesto que una de nuestras hipótesis de trabajo considera que los usuarios comparten un vocabulario de base que forma el centro del espectro del léxico de cada centro de interés, la aparición en las listas individuales de esos términos centrales debería ser poco relevante a la hora de establecer la capacidad léxica individual. Pero las listas de disponibilidad también contienen términos no centrales o no prototípicos. La aparición de estos términos indica una capacidad léxica mayor, ya que son vocablos de uso más restringido y, por tanto, se consideran de más difícil acceso para el conjunto de hablantes. En consecuencia, estos elementos léxicos deberían tener una mayor relevancia en la determinación de la capacidad léxica individual (2010: 69).

En definitiva, el Índice de Descentralización Léxica del Informante (IDD) puede explicarse como un valor que mide la descentralización o "especificidad" de la producción léxica de cada informante respecto al conjunto de datos "compatibles" o prototípicos de un centro de interés. Nuevamente, hemos reproducido la fórmula para calcular el IDD a partir de las premisas y explicaciones consideradas por Ávila Muñoz y Sánchez Sáez (2010: 76-81). La fórmula sería la siguiente:

$$IDD_j = 1 - \prod_{i=1}^{P_j} \left(1 - k\left(1 - COMP_{palabra\_i}\right)\right) \qquad [6]$$

donde k es una constante determinada experimentalmente y para la que sus autores proponen en su trabajo el valor de 0,2; P_j es el número de palabras mencionadas por el informante j-ésimo, y $COMP_{palabra\_i}$ es el índice de compatibilidad calculado para el vocablo, empleado en la palabra i-ésima de este informante j-ésimo.

2. SOBRE LA COMPARABILIDAD DE ESTOS ÍNDICES CUANDO SE APLICAN EN DIFERENTES ESTUDIOS

Sería deseable que estos indicadores presentados anteriormente, cuando se calculan para los correspondientes informantes o vocablos mencionados por algún colectivo de informantes en un determinado estudio, produjeran niveles de medida que pudieran compararse a los análogamente obtenidos en otros estudios similares realizados sobre otros colectivos. Así podríamos dar una respuesta fiable a comparaciones del tipo de si tal vocablo es más disponible o menos entre los informantes de un colectivo que en el otro; si

tal informante contribuye más o menos que otro a la disponibilidad léxica de su colectivo de pertenencia; o si tal colectivo posee una mayor o menor disponibilidad (competencia) léxica comparativamente con el otro.

La clave para poder responder de forma fiable a este tipo de preguntas reside en que las medidas que se obtuvieran mediante dichos indicadores estuvieran referidas a unas mismas escalas comunes de referencia: a unas escalas de medida comunes. De esta forma, y a modo de ejemplo con relación a la medición de la disponibilidad léxica de un cierto vocablo en cierto colectivo de informantes, si comparamos los resultados de dos estudios similares y hubiéramos obtenido que la medida de su disponibilidad en el primer estudio fuese de 0,011 y de 0,022 en el segundo, podríamos concluir sin ninguna duda que "dicho vocablo es menos disponible en el colectivo de informantes del primer estudio que en el del segundo". Sin embargo, las conclusiones extraídas de la comparación directa de valores cuando las escalas de referencia no son comunes pueden resultar engañosas, cuando no absolutamente erróneas.

Efectivamente, imaginemos que consideramos unos hipotéticos vocablos de máxima disponibilidad (el que es mencionado por todos los informantes en primer lugar) y de mínima disponibilidad posible (el que tiende a no aparecer en las listas, que en la muestra podría aproximarse a un vocablo que pudiera haber sido mencionado en última posición por el informante que más palabras mencionó y que no hubiera mencionado nadie más): cualquier otro vocablo debería tener una disponibilidad medida entre la de estos dos vocablos, de forma que si calculáramos las disponibilidades de estos dos hipotéticos vocablos, los valores obtenidos podrían servirnos de referencia para valorar el grado de disponibilidad de los demás vocablos. Y, así, si dichos valores de referencia para estos dos hipotéticos vocablos fueran en todo caso los mismos (por ejemplo, los valores 1 y 0 respectivamente, que suelen ser un estándar para indicadores normalizados), diríamos que la disponibilidad de ambos vocablos parece ser de pequeña magnitud, algo menor en el primer caso que en el segundo.

Sin embargo, el problema de los anteriores índices estriba en que pueden producir valores de referencia diferentes, cuando se aplican a los estudios que se desean comparar. ¿Y si los valores de referencia fueran respectivamente 1 y 0,005 en el primer estudio, y 1 y 0,022 en el segundo? En este caso, el vocablo considerado presentaría la disponibilidad mínima en el segundo estudio, mientras que parece ser algo más disponible en el primero, ya que su disponibilidad es mayor que la menor posible.

Para analizar el comportamiento de los indicadores presentados anteriormente, desde la perspectiva de la comparabilidad de los resultados que proporcionan cuando se aplican a diversos estudios, hemos construido

artificialmente tres muestras ideales a partir de los datos recogidos para un determinado centro en un estudio real[6] y procedido a comparar sus resultados, que analizaremos en los apartados siguientes.

La primera muestra (que llamaremos muestra 10i) trata de simular el caso de un estudio con pocos informantes. Está constituida por diez informantes que han producido un total de 109 vocablos, y en la que hemos añadido artificialmente un vocablo de máxima disponibilidad (en adelante, "Voc1MaxDisp"), que habría sido mencionado por todos los informantes en primera posición. Análogamente, hemos incluido un segundo vocablo de máxima disponibilidad (en adelante, "Voc2MaxDisp") que habría sido mencionado también por todos los informantes siempre en la segunda posición. Además, hemos incluido artificialmente un tercer vocablo de mínima disponibilidad (en adelante, "VocMinDisp") que se ha añadido al final de la lista de vocablos mencionados por el informante que más vocablos mencionó (a quien en adelante nos referiremos como el "informante más locuaz"), quien originariamente había mencionado 32 vocablos. Por tanto, este vocablo VocMinDisp se añadió en la posición 35 del informante más locuaz, ya que en sus dos primeras posiciones ahora estarían los vocablos Voc1MaxDisp y Voc2MaxDisp de máxima disponibilidad, seguidos de las 32 palabras que realmente mencionó.

La segunda muestra (que llamaremos muestra 10x10i) trata de representar una muestra con suficiente número de informantes. Se construye replicando 10 veces la muestra 10i antes descrita, pero dejando un único informante más locuaz de las correspondientes diez replicaciones. De esta forma, la muestra constaría de 100 informantes que mencionan en las posiciones 1ª y 2ª respectivamente los vocablos Voc1MaxDisp y Voc2MaxDisp de máxima disponibilidad, así como un tercer vocablo de mínima disponibilidad artificialmente construido al final de la lista de vocablos mencionados en la posición 35 (VocMinDisp) por uno solo de los 10 informantes replicados que originariamente mencionaron 32 palabras. Además, en la muestra 10x10i se mencionan los mismos vocablos y en las mismas posiciones que en la muestra 10i, pero con una frecuencia de aparición en cada posición 10 veces mayor debido a la replicación (salvo en el caso del vocablo de mínima disponibilidad VocMinDisp que sigue mencionándose una sola vez por un único informante, el "más locuaz").

[6] Estudio realizado en el desarrollo de la Tesis Doctoral "Léxico disponible de estudiantes de español como lengua extranjera en la Comunidad de Madrid", defendida por Diego Javier Gallego Gallego en la Universidad de Alcalá en 2014.

Finalmente, hemos confeccionado una tercera muestra (que llamaremos muestra 10x10i50p) para analizar el efecto que sobre los indicadores podría tener una longitud máxima de las correspondientes listas de palabras en los estudios comparados. Nótese que, aun velando por la homogeneidad en la recogida de datos de las pruebas de disponibilidad (mismos enunciados de centros, mismo tiempo de respuesta, etc.), es relativamente fácil encontrarnos con esta circunstancia. La propia competencia léxica de los informantes (que son generalmente distintos en cada estudio comparado) influye sobre el número de palabras que mencionan en los listados; en este caso, la mayor o menor longitud de las listas de palabras mencionadas por los informantes puede ofrecer información importante para observar la mayor o menor disponibilidad (competencia) léxica de un colectivo de informantes frente al otro. Por otra parte, el tamaño de las muestras (número de informantes) suele hacer disminuir o crecer en cierto grado, de forma natural, la longitud máxima de las listas de palabras observadas (al considerar más informantes, aumenta la probabilidad de contar efectivamente con alguno más locuaz). Además, incluso en el caso de que las listas de palabras proporcionadas por la práctica totalidad de los informantes de los estudios comparados mostraran características de disponibilidad similares, podría aparecer en una de las muestras comparadas un informante muy diferente del resto que menciona muchas más palabras que los demás informantes. En este caso, ¿debería el comportamiento de este único informante atípico modificar sustancialmente la medición del grado de disponibilidad léxica de los vocablos en el grupo? Para tratar de analizar estos aspectos hemos construido la muestra 10x10i50p, que es la misma muestra 10x10i a la que hemos añadido artificialmente más palabras a partir del último vocablo que mencionó originariamente el informante "más locuaz" (en las posiciones 35 a 49) y colocando ahora el VocMinDisp en la posición 50 de su lista. Al igual que en las dos muestras anteriores, hemos mantenido los dos vocablos de contraste (Voc1MaxDisp y Voc2MaxDisp) en las posiciones 1ª y 2ª de todas las listas.

Una vez construidas las tres muestras descritas anteriormente[7], hemos procedido a calcular para ellas los índices antes presentados para luego pasar

[7] El principal objetivo de este trabajo ha sido por un lado, analizar los efectos que sobre las mediciones clásicas de disponibilidad pueden producir particularmente las variaciones del número de informantes y de la máxima longitud de las listas de vocablos, cuando el resto de factores permanece inalterado, y por otro, proponer nuevas alternativas que palíen aquellos efectos. En consecuencia, a los métodos propuestos debemos exigirles que, cuando se apliquen a muestras diferentes procedentes de una misma realidad léxica, proporcionen las mismas mediciones de disponibilidad en todas ellas y no se vean afectados ni por el diferente número de informantes ni por la longitud de la lista más larga de vocablos. Es por ello que hemos

a analizar sus resultados, siempre bajo la perspectiva de su comparabilidad entre estudios, que presentamos a continuación.

3. COMPARABILIDAD DEL ÍNDICE DE LA DISPONIBILIDAD LÉXICA DE UNA PALABRA (IDLP) DE LÓPEZ CHÁVEZ Y STRASSBURGER FRÍAS

En la Tabla 1 se presentan las principales disponibilidades calculadas para las tres muestras construidas mediante el IDLP propuesto por López Chávez y Strassburger Frías (1987, 1991).

diseñado estas tres muestras tratando de reproducir en todas ellas una misma realidad léxica básica (mismo nivel léxico de los informantes y mismas disponibilidades de los vocablos que emplean), diferenciándose estas muestras exclusivamente en aquellos factores que producen los efectos particularmente mencionados antes (número de informantes y máxima longitud de las listas de vocablos). Las tres muestras de prueba así generadas permiten reproducir niveles similares de disponibilidad para todos los vocablos e informantes considerados en las muestras construidas, con las lógicas excepciones del vocablo de mínima disponibilidad (mencionado solo por el informante más locuaz en la última posición) y del informante más locuaz (que es el único que menciona este vocablo). Ciertamente estas muestras han sido construidas artificialmente, pero si hubiéramos empleado muestras más realistas, éstas se diferenciarían también en otros aspectos además de los dos que queremos controlar (mínimamente se diferenciarían también en la propia aleatoriedad que induce en muestreo), no dejando comprobar tan nítidamente el cumplimiento de las propiedades exigidas a los índices. En este sentido hay que tener en cuenta que las fórmulas que se proponen en este trabajo son generalizaciones de las formulaciones clásicas, quedando éstas como casos particulares de aquellas. En consecuencia, es de esperar que nuestras fórmulas hereden la esencia cualitativa que para la medición de los aspectos puramente léxicos presentan las fórmulas originales, de forma que no se prevén grandes problemas para la validación de estos nuevos índices en situaciones reales. En todo caso, para la aplicación práctica de estas nuevas fórmulas a situaciones reales será necesario decidir sobre las combinaciones paramétricas plausibles que pudieran ser consideradas "estándares" para la comparabilidad universal, para lo que previamente sería necesario validar dichas fórmulas concretas en tales situaciones reales. Ambos requerimientos podrían ser abordados, en primera instancia, mediante análisis de muestras generadas con técnicas de remuestreo o *bootstrapping* controlado a partir de algún caso amplio y real de referencia; lo que lógicamente requiere nuevos esfuerzos de investigación que van más allá de los objetivos fundamentalmente teóricos planteados inicialmente en este trabajo.

Tabla 1: IDLP en las tres muestras comparadas

	IDLP		
	10i	**10x10i**	**10x10i50p**
Voc1MaxDisp	1,000000	1,000000	1,000000
Voc2MaxDisp	0,934590	0,934590	0,954146
calle(s)	0,591980	0,591980	0,647608
… …	**… …**	**… …**	**… …**
piscina	0,011478	0,011478	0,022268
VocMinDisp(pos35)	**0,010026**	**0,0010026**	
VocMinDisp(pos50)			**0,0010026**

En primer lugar, podemos observar que la medida de disponibilidad obtenida por el vocablo de máxima disponibilidad (Voc1MaxDisp) en las tres muestras alcanza adecuadamente el valor 1. Además, y con excepción del vocablo de mínima disponibilidad (VocMinDisp), la valoración de la disponibilidad de los restantes vocablos originales (de "calle" a "piscina", en orden decreciente de disponibilidad) es idéntica en las dos primeras muestras, cosa además lógica considerando que la segunda muestra es simplemente una replicación de la primera. Por otra parte, respecto al vocablo de mínima disponibilidad (VocMinDisp), también parece razonable que su disponibilidad en la segunda muestra sea exactamente la décima parte de la medida obtenida en la primera muestra, ya que en la primera muestra lo menciona un informante de los diez, mientras que en la segunda lo hace solo uno de los 100 que compone la muestra.

Sin embargo, si nos fijamos en la tercera muestra, en donde la única diferencia con la segunda es exclusivamente que el informante "más locuaz" menciona 50 vocablos en lugar de 35 (el resto de menciones permanece exactamente igual en número y posición que en la segunda muestra), observamos que, con excepción del primer vocablo de máxima disponibilidad (Voc1MaxDisp) que toma el valor 1, las medidas de disponibilidad calculadas para todos los demás vocablos se ven alteradas al alza, siendo la disponibilidad del vocablo de mínima disponibilidad (VocMinDisp) exactamente la misma que en la segunda muestra, independientemente de que se encuentra mencionada en una posición bastante posterior.

Para comprender los motivos de este comportamiento del índice, debemos analizar la expresión [2]. Por una parte, como indican López Chávez y Strassburger Frías (1987, 1991), el peso que aplica la fórmula a la última posición "n" de mención a cada vocablo en cada muestra, es siempre el

mismo (0.10026), lo que produce que el valor mínimo efectivo que puede presentar este índice (en el vocablo de mínima disponibilidad, VocMinDisp) vale siempre, en cualquier estudio:

$$D(VocMinDisp) = e^{-2.3\left(\frac{n-1}{n-1}\right)}\frac{1}{I_1} = \frac{e^{-2.3}}{I_1} = \frac{0.100258843}{I_1} \approx \frac{0.10026}{I_1}$$

dependiendo este valor solo del nº de informantes del estudio, independientemente de cómo de temprana o tardía sea la posición en que se mencione. Esto explica los valores mostrados en la Tabla 1 para las disponibilidades calculadas de VocMinDisp, donde el número de informantes (I_1) es respectivamente de 10, 100 y 100 en cada una de las tres muestras.

Por tanto, la referencia inferior de la escala de medida de este índice no tiende a cero a medida que crece la posición de mención del vocablo de mínima disponibilidad, sino que está determinada exclusivamente por la constante "2.3" que aparece en el exponente de la expresión, y por el tamaño de la muestra empleada (número de informantes del estudio), independientemente de la posición en que haya sido mencionado dicho vocablo.

A modo ilustrativo, en la Tabla 2 se muestran las disponibilidades que calcularía el índice propuesto por López Chávez y Strassburger Frías (1987, 1991) para el vocablo de mínima disponibilidad cuando es aplicado en estudios con diferente número de informantes, independientemente de la posición de mención en que apareciera dicho vocablo en los diferentes estudios:

Tabla 2: Disponibilidad mínima de los vocablos en las tres muestras

IDLP	Número de Informantes				
	10	20	50	100	200
(VocMinDisp)	0,010026	0,005013	0,002005	0,001003	0,000501

Desde un punto de vista práctico y para tamaños de muestras habituales (en torno a los 50 o más informantes), estos valores se encuentran relativamente cercanos a cero, y las diferencias que se observan entre ellos se deben exclusivamente al hecho de que el vocablo de mínima disponibilidad es proporcionalmente menos frecuente en términos relativos cuanto mayor es el número de informantes. De hecho, si el vocablo de mínima disponibilidad se diera con la misma frecuencia relativa en todos los estudios, por ejemplo, con frecuencia 1 cuando el número de informantes es 10; con frecuencia 2,

si es 20; 5 cuando es 50; 10 cuando es 100 y 20 cuando es 200 (frecuencias relativas del 10% en todos los casos), su disponibilidad calculada sería en todos los casos la misma: 0,010026.

A diferencia de la propuesta de Lorán y López Morales (1983), López Chávez y Strassburger Frías introducen en su fórmula la longitud "n" de la lista más larga, buscando posiblemente una escala homogénea que permitiera comparar en términos relativos las disponibilidades de vocablos procedentes de distintos centros dentro de un mismo estudio (una única muestra). Como el número de vocablos potencialmente disponibles en cada centro podría diferir sustancialmente, la formulación de López Chávez y Strassburger Frías (1987, 1991) proporciona una cierta vía de normalización para los distintos centros de la medida de disponibilidad, de forma que la disponibilidad medida en cualquiera de los considerados en un mismo estudio siempre estaría entre los valores 1 (máxima disponibilidad) y $e^{-2,3}/I_1$ (mínima disponibilidad). Y como en un mismo estudio los informantes son siempre los mismos, su número I_1 es siempre constante y, en consecuencia, los valores de referencia de mínima y máxima disponibilidad serían siempre los mismos para todos los centros. De esta manera, a partir de las disponibilidades de los vocablos calculadas para cada centro, pueden abordar más sencillamente la construcción de su índice de disponibilidad léxica individual para cada centro (IDLI en el centro) como media ponderada de las disponibilidades de los vocablos que emplea, y a partir de estos, calcular como promedio simple el índice de disponibilidad léxica individual (IDLI Total).

Sin embargo, si buscamos poder comparar los resultados alcanzados por los vocablos en un mismo centro con los obtenidos en diversos estudios, el hecho de que el nivel de disponibilidad asignado por la fórmula original al vocablo de mínima disponibilidad sea siempre el mismo, independientemente de la posición en que se mencione, puede suponer una limitación importante, al poder ser muy diferentes las longitudes de las correspondientes listas más largas de palabras observadas en los diferentes estudios por motivos esencialmente imputables al distinto nivel léxico de los informantes de las muestras comparadas (diferentes disponibilidades léxicas de los grupos comparados); hecho este que queda enmascarado con este índice al no tener en cuenta las diferentes longitudes máximas de las listas. Así pues, si bien el tamaño de las muestras (nº de informantes) de los estudios comparados no parece ser problemático para la comparabilidad de este índice con estudios diferentes, sí que lo puede ser este hecho.

Por otra parte, las formulaciones de López Chávez y Strassburger Frías (1987, 1991) para medir la disponibilidad léxica de las palabras y la disponibilidad léxica de un informante en un centro introducen mecanismos para penalizar gradualmente las menciones menos inmediatas de los vocablos

(posiciones de aparición tardía en las listas de los informantes), de forma que una menor inmediatez de aparición implica una menor disponibilidad. Sin embargo, la intensidad con que dichos indicadores consideran (ponderan o penalizan) las distintas posiciones en que los vocablos son mencionados pueden variar también entre los estudios comparados, ya que ésta también depende de las respectivas longitudes máximas de sus listas de palabras, "n".

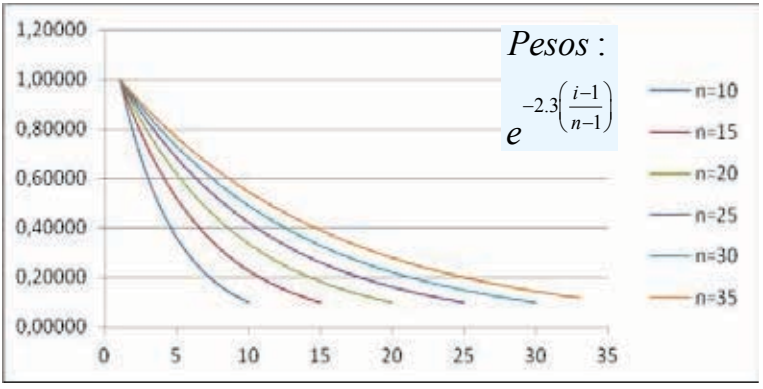
Como podemos observar en la expresión del índice, la ponderación que la fórmula aplica sobre la frecuencia relativa de aparición del vocablo considerado en cada posición (i-ésima), es

$$e^{-2.3\left(\frac{i-1}{n-1}\right)} \qquad [7]$$

y depende del número máximo de palabras (n) dado en la lista más larga del estudio. De esta forma, cualquier diferencia en las respectivas longitudes de las listas más largas dadas por los informantes en los estudios comparados producirá variaciones también en las ponderaciones que aplicará la fórmula sobre las menciones de los vocablos en cada posición; en consecuencia, se altera también la valoración general de la disponibilidad del vocablo que, en cuestión, calcula la fórmula. Como ya hemos mencionado anteriormente, la diferente competencia léxica de los colectivos de informantes comparados, el diferente tamaño de las muestras empleadas en los estudios, o la simple aparición en alguno de los listados de algún informante atípicamente locuaz, entre otras circunstancias, pueden propiciar fácilmente la aparición de diferentes longitudes "n" en los estudios cuyos resultados deseemos comparar.

En el Gráfico 1 hemos representado el decaimiento de las ponderaciones que emplea la fórmula de López Chávez y Strassburger (1987, 1991) para valorar las menciones que se realizan en cada posición de mención (i-ésima) del vocablo cuya disponibilidad se desea calcular, para distintas longitudes (n) de la lista de palabras más larga dada por los informantes. Vemos, efectivamente, cómo todas las curvas decaen de forma exponencial, tanto más lentamente cuanto mayor es el valor de "n" en la muestra analizada, para terminar en todo caso tomando el valor 0,10026 justo en la posición "n".

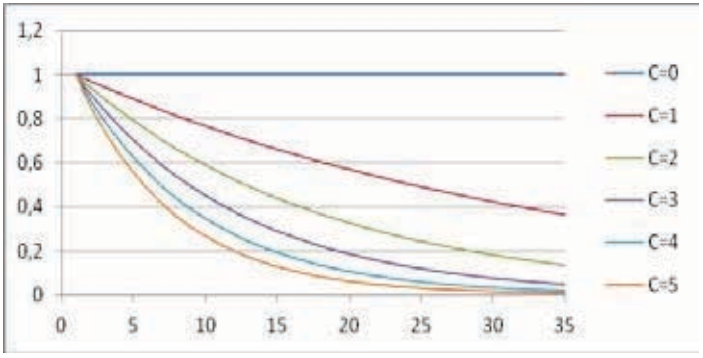
Gráfico 1: Caída de ponderaciones con la posición (eje horizontal) para varias longitudes, 'n', de la lista más larga (curvas)



Es este el hecho, el simple incremento de "n" (que incluso podría producirse un tanto accidentalmente por un comportamiento atípico de un único informante), el que conduce a sobrevalorar la disponibilidad de todos los vocablos en la tercera muestra, según deja ver la Tabla 1. Este efecto, en nuestra opinión, aún se hace más patente si consideramos el segundo vocablo de máxima disponibilidad (Voc2MaxDisp), el que mencionan todos los informantes inmediatamente después del primero, que igualmente fue mencionado por todos. Obviamente fue menos inmediato que el primero y es lógico que presente menor disponibilidad, pero una vez que todos los informantes han mencionado el primer vocablo en la primera posición, ningún otro vocablo podría ser más disponible que este segundo, en ninguna situación, ya que vuelve a ser mencionado por todos en la posición inmediatamente posterior. Pensamos que esta situación hipotética debería conducir a un mismo nivel de disponibilidad en cualquier estudio, lo que no ocurre con el índice analizado.

Por otra parte, López Chávez y Strassburger Frías (1987, 1991) introducen en las ponderaciones de su índice de disponibilidad léxica, de forma experimental, el exponente "2.3"; que como hemos visto, conduce a que las ponderaciones caigan siempre al valor 0,10026, justamente en la posición "n". En este sentido, en el Gráfico 2 se representa cómo varía el grado de decaimiento de aquellas ponderaciones cuando aplicamos distintos valores de *C* en lugar del valor 2.3 que aparece en el exponente de la fórmula, habiendo fijado para ello el parámetro n=35, a modo de ejemplo (valor representativo de la posición máxima alcanzada en las dos primeras muestras que estamos tomando de referencia en este trabajo).

Gráfico 2: Caída de ponderaciones con la posición (eje horizontal) para varios valores de "C" (curvas)



$$e^{-C\left(\frac{i-1}{n-1}\right)}$$

Encontramos que la caída se hace más brusca a medida que aumenta el valor de *C*. Cuanto mayor es *C*, el poder discriminativo de los pesos se concentra más en las primeras posiciones, llegando a ser casi irrelevante en las últimas. A la inversa, la caída de la curva es más suave, o poco marcada, cuando *C* toma valores más pequeños, en cuyo caso las ponderaciones obtenidas decrecen casi linealmente, proporcionando poco valor discriminador y relativamente altos pesos en las últimas posiciones[8]. Un valor intermedio de *C* produce pues una caída intermedia, con una progresividad decreciente del poder discriminador de las ponderaciones, que llegarían a ser pequeñas, pero no insignificantes, en las últimas posiciones. Ello justificaría haber tomado para *C* un valor intermedio, como podría ser el 2,3 (que, como hemos dicho, conduce a una ponderación aproximada de 0,1 en posición "n" más remota de las menciones).

3.1. Índice Estandarizado de Disponibilidad Léxica de una Palabra

Con la intención de paliar los inconvenientes reseñados que desde la perspectiva de su comparabilidad entre estudios presenta el índice de disponibilidad léxica propuesto por López Chávez y Strassburger Frías

[8] Obsérvese que la situación de indiferencia en la que el índice no distinguiría entre posiciones de mención se obtiene para el valor *C=0*, en cuyo caso el peso sería igual a uno en cualquier posición.

(1987, 1991),y de acuerdo con el análisis realizado, proponemos una sencilla modificación de este, consistente en parametrizar el índice de forma que, en cualquier estudio, podamos considerar una misma curva de ponderaciones acordada previamente por los investigadores, independientemente de las diferentes longitudes máximas de las listas de palabras "n" observadas en los mismos.

La formulación de este nuevo Índice Estandarizado de Disponibilidad Léxica de un Vocablo que se propone (en adelante, IDLP_st) es la siguiente:

$$D^{st}{}_{w,k,m}\left(V_j\right) = \sum_{i=1}^{n} w^{\left(\frac{i-1}{k-1}\right)^m} \frac{f_{ji}}{I_1} \qquad [8]$$

siendo:

- V_j = vocablo cuya disponibilidad se va a medir, siendo *j* el índice que identifica a V_j en la lista de vocablos del centro de interés
- i = indicador de las posiciones en que puede mencionarse el vocablo V_j
- n = máxima posición alcanzada por los vocablos en el centro de interés
- I_1 = número de informantes del centro que participaron en el estudio
- f_{ji} = número de menciones del vocablo V_j del centro en la posición *i-ésima*
- k = indicador de la posición en la que se desea que el peso valga "*w*"
- w = nivel del peso, comprendido entre 0 y 1, deseado en la posición "*k*"
- m = constante que modula la intensidad con que se aplican los pesos

Esta formulación generaliza las de Lorán y López Morales (1983), y López Chávez y Strassburger Frías (1987, 1991), quedando sus formulaciones como casos particulares de esta. De hecho, la fórmula de Lorán y López Morales sería la que proporcionaría el índice para los parámetros $w=\lambda^{k-1}$ y $m=1$:

$$d\left(V_j\right) = \sum_{i=1}^{n} \lambda^{i-1} \frac{f_{ji}}{I_1} = D^{st}{}_{\lambda^{k-1},k,1}\left(V_j\right) \qquad [9]$$

mientras que la de López Chávez y Strassburger Frías sería la que proporcionaría el índice para los parámetros w=e$^{-2,3}$, k=n y m=1:

$$D(V_j) = \sum_{i=1}^{n} e^{-2.3\left(\frac{i-1}{n-1}\right)} \frac{f_{ji}}{I_1} = D^{st}{}_{e^{-2,3},n,1}\left(V_j\right) \qquad [10]$$

Con esta nueva formulación, el vocablo de referencia de máxima disponibilidad (Voc1MaxDisp) sigue dando una disponibilidad de 1, mientras que el vocablo de referencia de mínima disponibilidad (VocMinDisp) obtiene una disponibilidad de valor:

$$D^{st}{}_{w,k,m}\left(\text{VocMinDisp}\right) = \frac{w^{\left(\frac{n-1}{k-1}\right)^m}}{I_1} \xrightarrow[n\to\infty]{} 0 \qquad [11]$$

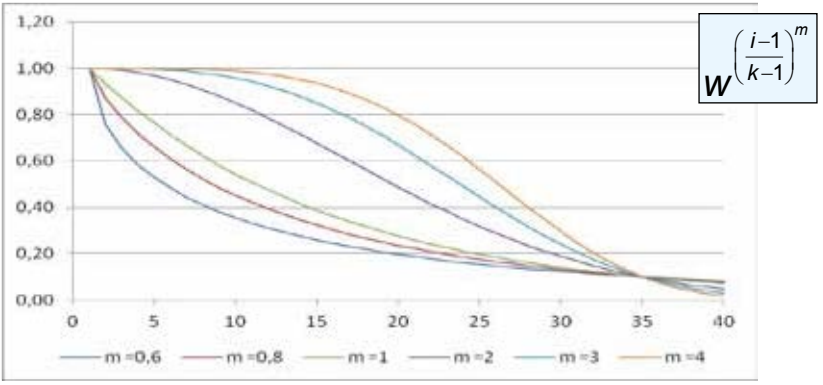
Que ahora (al ser 0<w<1) tiende a cero a medida que se retarda la posición (n) en que se menciona. Además, se consigue también que las ponderaciones aplicadas a la frecuencia relativa de aparición del vocablo en cada posición,

$$w^{\left(\frac{i-1}{k-1}\right)^m} \qquad [12]$$

solo dependan de los parámetros "w", "k" y "m", y de la propia posición (i-ésima) considerada. En consecuencia, fijados unos valores convenientes para estos parámetros, las medidas de disponibilidad resultantes serán comparables con las de cualquier estudio que aplique a sus resultados este índice con la misma parametrización.

La función de pesos asociada con este índice toma pues el valor de 1 en la primera posición, y decae progresivamente hasta alcanzar el valor fijado para el parámetro "w" al llegar a la posición prefijada "k", siendo el parámetro "m" el que determina la intensidad de este decaimiento de las ponderaciones a medida que avanzamos sobre las posiciones de mención. Como ilustración de este comportamiento, en el Gráfico 3 se puede apreciar, a modo de ejemplo, cómo varía el grado de decaimiento de dichas ponderaciones cuando aplicamos distintos valores del *parámetro de modulación (m)*, para el caso en que parametrizamos el índice para que esta curva tome una ponderación de valor 0,1 (parámetro w=0,1) en la posición de mención número 35 (parámetro k=35).

Gráfico 3: Caída de ponderaciones con la posición (eje horizontal) para varios valores de "m" (curvas)



En ella vemos que cuando m<=1, el decaimiento de las curvas (las tres de abajo) se asemeja en la forma a la que presentaba la función de pesos asociada a la fórmula de López Chávez y Strassburger Frías (1987, 1991). Sin embargo, en aquel caso, la curva estaba determinada de forma única decayendo al ritmo que imponía el exponente "2,3" que aparece fijo en su expresión, hasta tomar el valor preestablecido de 0,10026 en la posición de mención que se correspondía con la mayor longitud de las listas de palabras mencionadas por los informantes. Ahora, mediante los tres parámetros de que dispone el nuevo índice propuesto, dicha curva puede flexibilizarse para que baje hasta un nivel deseado (*w*) en una posición determinada (*k*) y acentuando más o menos el decaimiento de los pesos de las primeras posiciones frente a los de las últimas mediante el nuevo parámetro de modulación (*m*) introducido en la fórmula.

Como puede verse también en las tres curvas superiores, cuando este parámetro de modulación (*m*) es superior a uno, la diferencia entre los pesos de las primeras posiciones (tantas más cuanto mayor sea el parámetro "*m*") puede hacerse muy pequeña, permitiendo valorar por un lado de forma muy parecida las primeras menciones y en consecuencia, de forma prácticamente independiente de su posición de aparición, y por otro, permite no penalizar en exceso la posición de mención para los vocablos mencionados en los primeros lugares.

En todo caso, y desde la perspectiva de la comparabilidad de resultados, deberá ser la comunidad científica la que decida sobre el conjunto de parámetros más adecuado a adoptar como "estándares", de forma que el

índice muestre lo mejor posible los diversos niveles de disponibilidad en los estudios.

Prefijados los valores de los parámetros "*w*" y "*k*", la siguiente fórmula muestra cómo puede determinarse de forma general el valor de "*m*" que consigue que la ponderación con que actúa la fórmula en una posición determinada "k^*", baje hasta un nivel "w^*":

$$m = \frac{\ln\left(\frac{\ln\left(w^*\right)}{\ln\left(w\right)}\right)}{\ln\left(\frac{k^*-1}{k-1}\right)} \qquad [13]$$

A modo de ejemplo, si queremos que la curva de pesos tome el valor 0,1 en la posición 35 (w=0,1 y k=35) y que decaiga de tal forma que en la quinta posición el peso no haya bajado de 0,99 (w^*=0,99 y k^*=5, para que las cinco posiciones iniciales tengan pesos parecidos), el parámetro de modulación "*m*" debería ser:

$$m = \frac{\ln\left(\frac{\ln\left(0,99\right)}{\ln\left(0,1\right)}\right)}{\ln\left(\frac{5-1}{35-1}\right)} = 2,5392587 \approx 2,54$$

Para ilustrar numéricamente el comportamiento descrito de este nuevo índice, hemos calculado con él las disponibilidades de los vocablos (que llamaremos IDLP_st) para las tres muestras de prueba construidas en este trabajo. En concreto, se muestran los resultados obtenidos con dos parametrizaciones diferentes.

Con objeto de que los resultados puedan compararse directamente con los de la Tabla 1, derivados de la fórmula de López Chávez y Strassburger Frías (1987, 1991), la primera parametrización reproduce los pesos de la fórmula de los investigadores mexicanos para n=35 (w=0,10026, k=35 y m=1), por lo que cabe esperar que el índice aplique un peso de 0,10026 a las frecuencias relativas de la posición 35 y que las disponibilidades calculadas sean similares a las presentadas en dicha tabla (al ser m=1, salvo para la muestra 10x10i), ya que la nueva medida debe mostrarse insensible a la longitud de la lista más larga de palabras. La segunda parametrización, por su parte, se corresponde a w=0,1, k=33 y m=2,54, para cuya elección se ha tenido en cuenta por un lado que el vocablo "piscina" solo se menciona

por un único informante de la muestra original en la posición 31 (posición 33ª de nuestras tres muestras, y de aquí la elección de *k*), y por otro, que se pretende ilustrar el papel modulador del parámetro *m* (y de aquí la decisión de tomar el valor antes calculado que las cinco posiciones iniciales tengan pesos parecidos), por lo que se espera que las disponibilidades de los primeros casos sean mayores que en el caso anterior y que para el vocablo "piscina" sea de 0,1 dividido por el número de informantes.

Finalmente, los resultados son los presentados en la Tabla 3 y, como puede verse en ella, las disponibilidades calculadas para todos los vocablos que presentaron análogas frecuencias relativas de aparición en las tres muestras coinciden, no viéndose condicionados ni por el diferente número de informantes de los estudios, ni por las diferentes longitudes máximas de las listas de palabras dadas por ellos. Además, podemos ver cómo la disponibilidad calculada para el vocablo de mínima disponibilidad (VocMinDisp) disminuye con la frecuencia relativa de aparición (1/10 en la muestra 10i y 1/100 en la muestra 10x10i) así como con la tardanza de su mención (posición de aparición). Es por esto que, en nuestra opinión, la segunda parametrización calcula de manera más adecuada las disponibilidades de los vocablos de máxima disponibilidad (Voc1MaxDisp y Voc2MaxDisp), cuyos valores son más próximos que el reflejado con la fórmula original.

Tabla 3: Disponibilidad (IDLP_st) de los vocablos en las tres muestras

	IDLP_st (w=0.10026, k=35, m=1)			IDLP_st (w=0.1, k=33, m=2.54)		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
Voc1MaxDisp	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
Voc2MaxDisp	0,934590	0,934590	0,934590	0,999654	0,999654	0,999654
calle(s)	0,591980	0,591980	0,591980	0,779774	0,779774	0,779774
…	… …	… …	… …	… …	… …	… …
Piscina	0,011478	0,011478	0,011478	0,010000	0,010000	0,010000
VocMinDisp (pos35)	0,010026	0,0010026		0,006816	0,0006816	
VocMinDisp (pos50)			0,000363			0,000011

En todo caso, y como venimos diciendo, desde la perspectiva de la comparabilidad de resultados, deberá ser la comunidad científica la que decida sobre la combinación más adecuada (o combinaciones más adecuadas)

de parámetros que podrían considerarse "estándares" para referir a ellos los resultados de los diversos estudios que se pretendieran comparar.

4. COMPARABILIDAD DEL ÍNDICE DE DISPONIBILIDAD LÉXICA INDIVIDUAL (IDLI) DE LÓPEZ CHÁVEZ Y STRASSBURGER FRÍAS

Pretendiendo evaluar el grado de aportación de un individuo dentro de los listados generales de disponibilidad de vocablos, López Chávez y Strassburger Frías (1991) presentan su Índice de Disponibilidad Léxica Individual (IDLI). Así, para cada centro y empleando una formulación parecida a la empleada para calcular el IDLP, calculan un índice de disponibilidad léxica individual para dicho centro (en adelante, $IDLI_c$) que definen con la siguiente expresión:

$$D_c(S_j) = \frac{1}{k}\sum_{i=1}^{P_j} d\left(P_{ij}\right)\cdot e^{-2.3\left(\frac{i-1}{n-1}\right)} \qquad [14]$$

donde

$D_c(S_j)$ = disponibilidad del sujeto o informante j-ésimo" (en el centro *c*)

P_j = número de palabras mencionadas por el sujeto *j* en la posición *i*

$d(P_{ij})$ = disponibilidad de la palabra mencionada por el sujeto *j* en la posición *i*

n = máximo número de menciones en el centro *c* de interés

k = constante para ajustar las calificaciones

Y, a partir de estos, calcula la calificación total de disponibilidad léxica del individuo como la suma de las anteriores para todos los centros (en la fórmula, *NC* es el número de centros considerados en el estudio):

$$DT(S_j) = \sum_{c=1}^{NC} D_c(S_j) \qquad [15]$$

El índice de Disponibilidad Léxica Individual para un Centro ($IDLI_c$) asigna a cada informante un valor que trata de expresar su grado de participación en la producción léxica global del centro *c* de interés. El objetivo de la

medida, por tanto, sugiere que la constante *k* que se emplearía para ajustar las calificaciones debería ser la valoración, con la misma lógica, de la disponibilidad léxica del centro, alcanzada conjuntamente por todos los informadores.

Como la formulación del Índice de Disponibilidad Léxica Individual para un Centro ($IDLI_c$) depende de la disponibilidad léxica de los vocablos (IDLP) que menciona cada informante, y además las valora mediante el mismo sistema de ponderaciones que se empleaba para el cálculo de las mismas, hereda los comportamientos ya descritos en el apartado anterior cuando se considera desde la perspectiva de la comparabilidad de sus resultados entre estudios diferentes. De hecho, el $IDLI_c$ asigna valores que permiten medir el grado en que cada informante participa en la producción léxica global de un centro de interés, pero al igual como sucede con el Índice de Disponibilidad Léxica de una Palabra (IDLP), los valores obtenidos mediante el $IDLI_c$ nunca llegan a ser 0, ya que su valor mínimo también aquí viene determinado en función de la longitud de la lista más larga de vocablos mencionada por los informantes.

Para ilustrar este comportamiento, hemos utilizado nuevamente las tres muestras de informantes creadas artificialmente, donde observaremos el comportamiento del informante con mayor disponibilidad léxica (Inform_Max_Dispon), el de menor disponibilidad léxica (Inform_Min_Dispon) y también aquel informante que hemos llamado el "más locuaz"(Inform_más_locuaz), que ha escrito más palabras que los demás informantes, hasta llegar a la posición 35 en las muestras 10i y 10x10i, y a la posición 50 en la muestra 10x10i50p. En las tres muestras hemos calculado la disponibilidad léxica individual de los informantes, considerando como valor *k* de ajuste de las calificaciones en cada una de ellas, la disponibilidad léxica absoluta calculada para el conjunto de informantes en el centro de interés. En la expresión siguiente, *Nvoc* representa el número de vocablos mencionados en el centro, mientras que V_1, V_2, ..., V_{Nvoc} representan dichos vocablos ordenados de forma decreciente en función de su disponibilidad:

$$D_c = \sum_{i=1}^{Nvoc} d\left(V_i\right) \cdot e^{-2.3\left(\frac{i-1}{Nvoc-1}\right)} \qquad [16]$$

En este sentido, presentamos en la Tabla 4 las disponibilidades léxicas individuales relativas al centro para estos informantes más característicos de la muestra. Como puede comprobarse en ella, el índice produce una ligera variación entre las dos primeras muestras, explicable por el hecho de que la disponibilidad léxica del vocablo de mínima disponibilidad es menor en la

segunda muestra que en la primera, al ser mencionado en ambos casos por un único informante de los 10 y 100 sujetos que componen las respectivas muestras. Lo anterior produce a su vez una pequeña variación a la baja de la disponibilidad léxica del centro en la segunda muestra, así como un ligero aumento igualmente proporcional de la participación relativa de los informantes, con la excepción del "más locuaz" que es quien menciona dicho vocablo de mínima disponibilidad.

Sin embargo, también se observa que las distintas longitudes de las listas emitidas por los informantes de los diferentes estudios (el caso de la tercera muestra) afectan de forma no proporcional a las valoraciones de los grados de disponibilidad léxica individuales en el centro realizadas por estos índices. Estas variaciones no proporcionales observadas para los informantes no podrán ser corregidas por ninguna otra determinación de la constante de escalado "*k*", que siempre produciría alteraciones proporcionales para todos los informantes.

Tabla 4: Disponibilidad ($IDLI_c$) de los Informantes para el centro, en las tres muestras

	$IDLI_c$		
	10i	**10x10i**	**10x10i50p**
Informante 1	0,525792	0,525854	0,516755
Informante 2	0,531523	0,531586	0,537339
Inform_Mín_Dispon	0,312226	0,312263	0,286071
Informante 4	0,523303	0,523364	0,527910
Informante 5	0,421363	0,421412	0,406338
Inform_Máx_Dispon	0,567753	0,567821	0,579247
Inform_más_locuaz	0,479477	0,479416	0,498962
Informante 8	0,523941	0,524003	0,516204
Informante 9	0,510473	0,510533	0,514900
Informante 10	0,468253	0,468308	0,473388

4.1. Índice Estandarizado de Disponibilidad Léxica Individual

El índice que proponemos emplea para su obtención una lógica similar a la expuesta anteriormente, pero utilizando las medidas de disponibilidad estandarizadas propuestas en este trabajo para los vocablos, así como las funciones de peso que se utilizaban en su definición. Así, fijados unos parámetros *w, k* y *m* para la medición de los índices estandarizados de disponibilidad léxica de los vocablos del centro, proponemos medir la disponibilidad léxica (absoluta) de un informante en el centro, como:

$$Dabs_c^{st}(S_j) = \sum_{i=1}^{P_j} D^{st}{}_{w,k,m}(V_i).w^{\left(\frac{i-1}{k-1}\right)^m} \qquad [17]$$

A partir de esta y para construir un índice relativo de disponibilidad léxica de los informantes que sea comparable para distintos estudios y fácilmente interpretable, proponemos tomar una referencia clara y referir a ella el nivel obtenido con la expresión anterior. En este sentido, proponemos tomar como referencia la de un informante ideal que mencionara *k* vocablos con la máxima disponibilidad teóricamente posible, y en consecuencia, su disponibilidad absoluta valdría:

$$K = Dabs_c^{st}(S_{ref}) = \sum_{i=1}^{k} w^{2.\left(\frac{i-1}{k-1}\right)^m} \qquad [18]$$

Una vez adoptada esta referencia común, definimos el Índice Estandarizado de Disponibilidad Léxica Individual ($IDLI_c$_st, en adelante) como:

$$D_c^{st}\left(S_j\right) = \frac{1}{K}\sum_{i=1}^{P_j} D^{st}{}_{w,k,m}(V_i).w^{\left(\frac{i-1}{k-1}\right)^m} \qquad [19]$$

Y, análogamente, el Índice Estandarizado de Disponibilidad Léxica del conjunto de Informantes para el Centro (IDL_c_st, en adelante), como:

$$D_c^{st} = \frac{1}{K}\sum_{i=1}^{Nvoc} D^{st}{}_{w,k,m}(V_i).w^{\left(\frac{i-1}{k-1}\right)^m} \qquad [20]$$

donde *Nvoc* representa el número de vocablos mencionados en el centro y V_1, V_2, ..., V_{Nvoc} representan a dichos vocablos ordenados de forma decreciente

en función de su disponibilidad calculada en el mismo a partir del conjunto de informantes. Es de notar que estos índices de disponibilidad léxica así calculados se pueden interpretar como la proporción (porcentaje si se multiplican por 100) que representan los correspondientes niveles léxicos de los informantes, o del conjunto de informantes respectivamente, comparado con el nivel de disponibilidad léxica que le correspondería al informante ideal de referencia.

Finalmente, como índice estandarizado del nivel de disponibilidad léxica del informante alcanzado dentro de su grupo en el centro ($IDLI_c$_st%, en adelante), proponemos:

$$D_c^{st\%}\left(S_j\right) = \frac{D_c^{st}\left(S_j\right)}{D_c^{st}} \qquad [21]$$

En este caso, este índice se puede interpretar, análogamente a los anteriores, como la proporción (porcentaje si se multiplican por 100) que representa el correspondiente nivel léxico del informante considerado, comparativamente con el nivel global alcanzado por el conjunto de informantes, tomado como un todo, en el centro *c* en cuestión.

Como ilustración del comportamiento de estos nuevos índices propuestos, los hemos calculado para las tres muestras de informantes creadas artificialmente, y presentamos los resultados más representativos en las tablas siguientes (tablas 5 y 6). En la Tabla 5, los hemos calculado para el conjunto de parámetros que aproximaría el índice al de López Chávez y Strassburger Frías. Como puede verse en ella, el comportamiento del índice para los informantes con mayor disponibilidad léxica (Inform_Max_Dispon) y menor disponibilidad léxica (Inform_Min_Dispon) se mantiene para las tres muestras, variando ligeramente la del informante que hemos llamado el "más locuaz". Esta ligera variación es explicable de forma lógica por el hecho de que el vocablo de mínima disponibilidad, que siempre ha sido señalado por un único informante, ha sido mencionado en la posición 35 por el 10% de la muestra 10i, en la posición 35 por el 1% de la muestra 10x10i, y en la posición 50 por el 1% de la muestra 10x10i50p, lo que hace variar ligeramente la disponibilidad de este vocablo entre muestras, así como la valoración global de la disponibilidad léxica del centro (IDL_c). En consecuencia, estas ligeras variaciones se transmiten a los índices calculados, especialmente a los del informante más locuaz. En cualquier caso, se observa una gran estabilidad en los índices de las tres muestras, a pesar de sus diferencias en el número de informantes y en las longitudes máximas de sus listas, como se pretendía con la propuesta.

Tabla 5: Índices de Disponibilidad Léxica de los Informantes del Centro. Cálculos usando la medida de disponibilidad IDLP_st y, en ambas fórmulas, w=0,10026; k=35; m=1

	IDLI$_c$_st			IDLI$_c$_st%		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
IDL$_c$_st	**0,650322**	**0,650321**	**0,650322**	**(1,000000)**	**(1,000000)**	**(1,000000)**
Informante 1	0,512878	0,512878	0,512878	78,86	78,87	78,87
Informante 2	0,518468	0,518468	0,518468	79,73	79,73	79,73
Inform_ Mín_Dispon	0,304557	0,304557	0,304557	46,83	46,83	46,83
Informante 4	0,510450	0,510450	0,510450	78,49	78,49	78,49
Informante 5	0,411014	0,411014	0,411014	63,20	63,20	63,20
Inform_ Máx_Dispon	0,553809	0,553809	0,553809	85,16	85,16	85,16
Inform_ más_locuaz	0,467701	0,467586	0,467663	71,91	71,90	71,91
Informante 8	0,511073	0,511073	0,511073	78,59	78,59	78,59
Informante 9	0,497936	0,497936	0,497936	76,57	76,57	76,57
Informante 10	0,456752	0,456752	0,456752	70,23	70,24	70,23

Tabla 6: Índices de Disponibilidad Léxica de los Informantes del Centro. Cálculos realizados usando la medida de disponibilidad IDLP_st y, en ambas fórmulas, w=0,1; k=33; m=2,54

	IDLI$_c$_st			IDLI$_c$_st%		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
IDL$_c$_st	**0,832181**	**0,832181**	**0,832181**	**(1,000000)**	**(1,000000)**	**(1,000000)**
Informante 1	0,597534	0,597534	0,597534	71,80	71,80	71,80
Informante 2	0,640216	0,640216	0,640216	76,93	76,93	76,93
Inform_ Mín_Dispon	0,292496	0,292496	0,292496	35,15	35,15	35,15
Informante 4	0,628572	0,628572	0,628572	75,53	75,53	75,53
Informante 5	0,449976	0,449976	0,449976	54,07	54,07	54,07
Inform_ Máx_Dispon	0,686909	0,686909	0,686909	82,54	82,54	82,54

	IDLI$_c$_st			IDLI$_c$_st%		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
Inform_ más_locuaz	0,570659	0,570620	0,570628	68,57	68,57	68,57
Informante 8	0,597550	0,597550	0,597550	71,81	71,81	71,81
Informante 9	0,606713	0,606713	0,606713	72,91	72,91	72,91
Informante 10	0,555356	0,555356	0,555356	66,74	66,74	66,74

Por otra parte, en la Tabla 6 se observan efectos análogos cuando se emplea la parametrización *w*=0,1; *k*=33; *m*=2,54 para calcular los correspondientes índices de disponibilidad. Ahora, el valor *m*=2,54 hace que los pesos bajen muy lentamente en las primeras posiciones, pero de manera muy rápida después de la posición 33 (al ser *k*=33), por lo que los últimos vocablos tienen una influencia poco apreciable tanto en la valoración de las disponibilidades léxicas de los vocablos (Tabla 3) como en la de los informantes que los mencionan en las posiciones más tardías. Por ello, la valoración global del centro (IDL$_c$) realizada a partir de esta parametrización permanece prácticamente constante. Lógicamente, las únicas variaciones que se aprecian se observan en el individuo "más locuaz", siendo menos importantes aún que las observadas en la tabla anterior, como cabía esperar de acuerdo con lo expuesto.

En consecuencia, la nueva fórmula propuesta generaliza la fórmula original de López Chávez y Strassburger Frías (que se reproduciría con la parametrización *w*=0,10026; *k*=*n*; *m*=1), y permite comparar los resultados de distintas investigaciones aunque presenten diferente número de informantes y distintas longitudes máximas en las listas de palabras recogidas.

5. COMPARABILIDAD DEL ÍNDICE DE COMPATIBILIDAD LÉXICA DE UN VOCABLO (COMP) DE ÁVILA MUÑOZ Y SÁNCHEZ SÁEZ

Continuando con la perspectiva de la comparabilidad entre diferentes estudios, hemos procedido de forma análoga para evaluar el comportamiento de los índices presentados por Ávila Muñoz y Sánchez Sáez (2010) con respecto de la medición de la Compatibilidad léxica de los vocablos y la Descentralización léxica de los informantes. En este apartado comenzamos por el análisis del Índice de Compatibilidad Léxica de un Vocablo (COMP).

En este caso, el índice propuesto por Ávila Muñoz y Sánchez Sáez (2010) para calcular el grado de compatibilidad de los vocablos respecto al núcleo prototípico de un centro de interés, presenta una situación contraria a la descrita en los apartados anteriores: si bien las fórmulas originales de López Chávez y Strassburger Frías (1987, 1991) para calcular el Índice de Disponibilidad Léxica de una Palabra (IDLP) y el Índice de Disponibilidad Léxica Individual (IDLI) no llegaban a tomar nunca el valor cero, por menos disponible que fuera un vocablo o menos disponibilidad léxica que presentara un informante, la fórmula propuesta por Ávila Muñoz y Sánchez Sáez nunca llega a tomar el valor 1, por más compatible que sea el vocablo considerado.

A modo de ilustración, hemos calculado el índice de compatibilidad léxica de Ávila Muñoz y Sánchez Sáez para las tres muestras de contraste ya introducidas y empleadas en los apartados anteriores. En la Tabla 7 se presentan los principales valores obtenidos para cada una de dichas muestras, adoptando dos criterios diferentes sobre el valor del parámetro *k* que aparece en su fórmula. Así, en la parte izquierda de la tabla se presentan los resultados cuando se emplea un único valor de *k* común para las tres muestras (se ha tomado *k*= 0.1, a modo de ejemplo), mientras que en la parte derecha se presentan los resultados cuando se ha calculado en cada caso un valor distinto de *k* para que la compatibilidad calculada para el vocablo de máxima disponibilidad se aproxime a 1 tanto como queramos (se ha tomado un error de aproximación ε de una millonésima, que conduce a valores calculados para el vocablo de máxima disponibilidad de 0,999999).

Tabla 7: Compatibilidades en las tres muestras al aplicar la fórmula de Ávila Muñoz y Sánchez Sáez

	COMP (k=0.1)			COMP (ε=0.000001)		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
Voc1MaxDisp	0,651322	0,999973	0,999973	0,999999	0,999999	0,999999
Voc2MaxDisp	0,401263	0,994079	0,994079	0,990818	0,998731	0,998731
calle(s)	0,152336	0,808470	0,808470	0,738971	0,882328	0,882328
…						
piscina	0,003030	0,029893	0,029893	0,022691	0,038421	0,038421
VocMinDisp (pos35)	0,002857	0,002857		0,021395	0,003687	
VocMinDisp (pos50)			0,002000			0,002581

Para comprender mejor el comportamiento del índice que muestra dicha tabla, consideremos la fórmula explícita derivada del proceso de cálculo propuesto por Ávila Muñoz y Sánchez Sáez (2010) que nos ha servido para el cálculo:

$$COMP_k\left(V_j\right) = 1 - \prod_{i=1}^{n}\left(1 - \frac{k}{i}\right)^{F_{ji}} \qquad [22]$$

siendo V_j el vocablo considerado, n es el número de la máxima posición que puede alcanzar un vocablo en las menciones que hacen de él los informantes; i representa a cada una de las posibles posiciones de mención del vocablo; F_{ji} la frecuencia absoluta de aparición del vocablo en la posición i-ésima, y k es la constante introducida por los autores en la valoración inicial de un vocablo que aparece en cada posición i ($t=k/i$).

Si pensamos en la hipotética situación en la que todos los informantes de una muestra escriben un determinado vocablo (VocMaxDisp) en la primera posición de los listados correspondientes a un centro de interés, su índice de compatibilidad debería ser máximo (1), sin embargo, al aplicarle esta fórmula, el valor calculado para su compatibilidad nunca llega a ser 1 para ningún valor válido de k:

$$COMP_k\left(VocMaxDisp\right) = 1 - \left(1 - k\right)^{I_1} \qquad [23]$$

Como se observa en la parte izquierda de la Tabla 7, el valor de máxima compatibilidad resultante depende sensiblemente del número de informantes de la muestra, cuando se prefija un valor concreto para el parámetro k, siendo sensiblemente mayor su alejamiento con respecto del valor máximo de referencia (1) en la muestra 10i (a pesar de que el Voc1MaxDisp ha sido mencionado por todos los informantes en la primera posición). Esto provoca que, en circunstancias como ésta, el límite superior efectivo de la escala de medida que subyace bajo este índice pueda ser sustancialmente inferior a 1, especialmente cuando se consideran muestras relativamente pequeñas, lo que resulta un inconveniente desde la óptica de la comparabilidad del índice para distintos estudios. Sin embargo, no parece que afecte a la medida de compatibilidad el efecto que el informante más locuaz produce sobre las longitudes máximas de las listas de palabras, al menos de forma apreciable; comportamiento éste plausible para el índice, ya que su interés está en enfatizar la compatibilidad de los vocablos con el núcleo prototípico del centro, no debiendo verse alterado por comportamientos atípicos o muy específicos.

Tratando de evitar el efecto inconveniente que produce el número de informantes sobre el índice, podría pensarse (y así lo proponen sus autores) en calcular un valor de *k* particular para cada muestra que asegurase una compatibilidad tan próxima a uno como deseemos para el vocablo de máxima disponibilidad. Para ello, basta con considerar que:

$$COMP_k\left(VocMaxDisp\right) = 1-\varepsilon \quad \Leftrightarrow \quad k = 1-\varepsilon^{1/I_1} \qquad [24]$$

por lo que, por ejemplo, si queremos que en un estudio que dispone de 100 informantes el valor de la compatibilidad del vocablo de máxima disponibilidad alcance el valor 0.999999 (ε=1-0.999999=0.000001), entonces el valor de *k* debe tomarse como:

$$k = 1-\varepsilon^{1/I_1} = 1-0.000001^{1/100} = 0.12903641 \approx 0.129$$

Y, análogamente, si el estudio tiene 10 informantes, el valor de *k* debería ser:

$$k = 1-\varepsilon^{1/I_1} = 1-0.000001^{1/10} = 0.748811356 \approx 0.749$$

De esta forma hemos construido la parte derecha de la Tabla 7, en la que observamos que al aplicar la fórmula COMP, el Voc1MaxDisp ahora se aproxima al valor 1 tanto como hemos establecido, aunque sigue sin alcanzarlo exactamente en ninguna de las tres muestras. Pero se observa también que la diferente elección de *k* en la primera muestra produce valoraciones diferentes de las posiciones de mención de las palabras a la hora de medir sus compatibilidades, con respecto de las empleadas en las dos muestras de 100 informantes. En consecuencia, la compatibilidad medida para cada vocablo también varía, siendo las variaciones de los distintos vocablos no proporcionales (por ejemplo, si comparamos las muestras de 100 informantes con las de 10 informantes, el mero hecho de considerar más informantes en la primera que en la segunda produce que la compatibilidad medida para el vocablo *calle(s)* se vea incrementada en un 19,4%, mientras que la del vocablo *piscina* se incremente un 69,3%, siendo iguales sus respectivas frecuencias relativas de aparición en cada posición para las tres muestras).

Por una parte, este hecho nos hace pensar que una cierta variación en el sentido observado para el grado de compatibilidad medido para los vocablos parece adecuada, pues aunque las frecuencias relativas de mención en las tres muestras han sido las mismas, parece lógico que el grado de compatibilidad

se reafirme con un mayor número de informantes que empleen el vocablo evaluado y, en consecuencia, la medida de su compatibilidad aumente con la frecuencia absoluta de los informantes que los mencionan. Pero, por otra parte, el vocablo Voc2MaxDisp ha sido mencionado por todos los informantes de la muestra, de forma que, en consecuencia, ya no ha podido ser mencionado por nadie más en ninguna de las tres muestras. Desde la perspectiva de la comparabilidad de los resultados de diferentes estudios, ¿no debería presentar este vocablo la misma medida de compatibilidad en todos los casos? Aunque consideramos que la respuesta a esta pregunta debe ser debatida y acordada por la comunidad científica, en cualquier caso, la formulación anterior no admite esta posibilidad.

5.1. Índice Estandarizado de Compatibilidad Léxica de un Vocablo

Un primer factor que puede explicar parte de los comportamientos diferentes observados para las muestras de contraste es el hecho de haber empleado distintos valores de *k* en las muestras, al tratar de paliar el efecto que produce el número de informantes sobre el máximo efectivo que podría tomar el índice original. Para comprenderlo mejor, analicemos más en profundidad el índice original de Ávila Muñoz y Sánchez Sáez (2010). Si suponemos que un vocablo perteneciente al núcleo prototípico tiende a aparecer en las primeras posiciones de mención con una ley de probabilidades decrecientes del tipo:

$$\Pr\left(X = i\right) = \begin{cases} \dfrac{k}{i} & , \quad si\, 1 \le i \le n^* \quad ; \quad k > 0\,,\, n^* \ge 1 \end{cases} \qquad [25]$$

siendo X la variable aleatoria que representa la posición de aparición de dicho vocablo, entonces la medida de compatibilidad de un vocablo que se define en Ávila Muñoz y Sánchez Sáez (2010) puede interpretarse como la probabilidad que tendría el vocablo de haber sido mencionado en alguna de las ocasiones en que lo fue, en el supuesto de que dicho vocablo perteneciera al núcleo prototípico, y en consecuencia, su complemento a uno (1-COMP) sería la probabilidad de que el vocablo, en el supuesto de que perteneciera al núcleo prototípico, no hubiera sido mencionado en ninguna de las ocasiones en que lo fue.

De esta forma, si COMP toma un valor bajo, 1-COMP toma un valor alto indicando que, si el vocablo fuera del núcleo prototípico, también sería alta la probabilidad de que no hubiera sido mencionado en ninguna de las ocasiones en que lo fue. Pero sin embargo fue mencionado en todas ellas, por lo que cuanto mayor sea el valor 1-COMP, mayor será la inconsistencia

de lo observado con la suposición de que el vocablo pertenece al núcleo prototípico, informándonos dicho valor de un cierto grado de alejamiento de núcleo, o de un cierto grado de "especificidad" del vocablo.

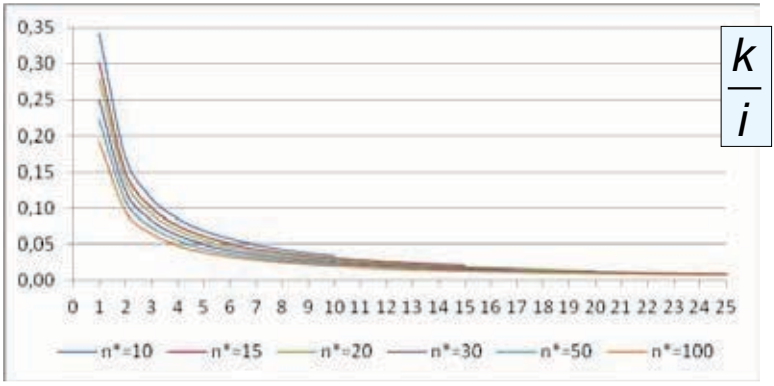
Pero para que dicha ley de probabilidad sea realmente una ley bien definida, el parámetro k no puede tomar cualquier valor arbitrario, sino que debe elegirse para que la suma de las probabilidades asociadas a todas las posiciones en que pueda aparecer mencionado un vocablo del núcleo prototípico valga 1. Así, si admitimos que un vocablo del núcleo prototípico puede aparecer según esta ley, con menor probabilidad cuanto más tarde aparezca, pero solo en un número finito de posiciones, n^*, los valores de k y n^* se encontrarían relacionados por las expresiones:

$$k = \frac{1}{\sum_{i=1}^{n^*} \frac{1}{i}} \approx \frac{1}{\ln\left(n^*\right) + \gamma} \quad \Leftrightarrow \quad n^* \approx e^{\frac{1}{k} - \gamma} \qquad [26]$$

donde γ es la constante de Euler-Mascheroni (0,577215664901532860606...), y consecuentemente, tanto el valor de la constante k, como las probabilidades que se emplearían para valorar en el índice de compatibilidad, especialmente las primeras posiciones, estarían entonces condicionados por el valor de n^*. A modo de curiosidad, y desde esta visión probabilística de la fórmula [22], los valores de k implícitamente empleados para construir los índices de la Tabla 7 (k=0.1, 0.129 y 0.749) conllevarían que los vocablos del núcleo prototípico deberían aparecer respectivamente antes de la posiciones 12367, 1306 y 2, respectivamente.

En el Gráfico 4 se muestra cómo sería el decaimiento de dichas probabilidades asociadas a cada posición, para distintos valores de n*. Obsérvese que, efectivamente, las probabilidades asociadas a cada posición que se considerarían para evaluar los correspondientes índices de compatibilidad en los diferentes estudios que se pretendieran comparar, podrían ser distintas si lo son las respectivas longitudes n^*. Concretamente, puede apreciarse cómo para los diferentes valores de n^*, las diferencias sustanciales se aprecian casi exclusivamente en las primeras posiciones de mención, haciéndose mínimas a medida que avanzamos sobre las posiciones (diferencias prácticamente despreciables a partir de la posición 25). Este hecho nos aconsejaría tomar un mismo valor de referencia n^* común para todos los estudios comparados, independientemente de sus longitudes máximas de sus listas, n. Así, el índice referiría siempre a una misma ley de probabilidad común, de forma que los resultados obtenidos serían comparables.

Gráfico 4: Caída de las probabilidades con la posición (eje horizontal) para varios valores de n* (curvas)



Otro factor importante para explicar los diferentes comportamientos observados en la Tabla 7 es el efecto que imprime sobre la valoración de cada posición el diferente número de informantes de las muestras. De hecho, el número de informantes (I_1) influye indirectamente en la fórmula [1] a través del exponente F_{ji} que aparece en el numerador, ya que estas frecuencias absolutas de aparición del vocablo en cada posición aumentan también de forma natural con el número de informantes de la encuesta.

Proponemos corregir este efecto redefiniendo la medida de especificidad del vocablo, 1-COMP, como 1-COMP elevado a la potencia $100/I_1$; lo que sería equivalente a calcular el índice referido a una muestra estandarizada de 100 informantes que hubiera presentado las mismas frecuencias relativas de mención del vocablo en cada posición que la muestra original.

En consecuencia, estos razonamientos nos llevan a un primer Índice Estandarizado de Compatibilidad Léxica de un Vocablo (en adelante COMP_st) como la transformación lineal al intervalo [0,1] del índice así modificado del propuesto por Ávila Muñoz y Sánchez Sáez, teniendo en consideración el valor real de su máximo efectivo:

$$C_k^{st}\left(V_j\right)=\frac{1-\prod_{i=1}^{n}\left(1-\frac{k}{i}\right)^{100\frac{F_{ji}}{I_1}}}{1-\left(1-k\right)^{100}} \qquad [27]$$

Como podemos observar en la Tabla 8, en la que se presentan sus resultados al calcularlos sobre las tres muestras de referencia, se alcanzaría el valor

máximo de referencia para el vocablo de máxima disponibilidad y se mantienen razonablemente las medidas de compatibilidad en todos los casos.

Tabla 8: Compatibilidades en las tres muestras al aplicar el índice COMP_st

	COMP_st (k=0.1)			COMP_st (k=0.16)		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
Voc1MaxDisp	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
Voc2MaxDisp	0,994106	0,994106	0,994106	0,999761	0,999761	0,999761
calle(s)	0,808491	0,808491	0,808491	0,930263	0,930263	0,930263
…						
Piscina	0,029894	0,029894	0,029894	0,047441	0,047441	0,047441
VocMinDisp (pos35)	0,028208	0,002857		0,044785	0,004571	
VocMinDisp (pos50)			0,002000			0,003200

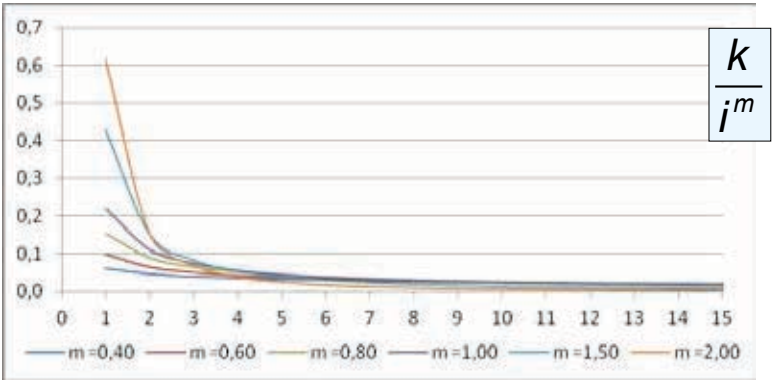
De todas formas, en el Gráfico 4 se observa también que para un determinado valor de n^*, y en consecuencia, para un valor fijo de k, la curva de decaimiento de las probabilidades estaría totalmente determinada, sin permitir la más mínima flexibilización de la misma para poder modular una mayor o menor intensidad en el decaimiento de la función de probabilidad en las primeras posiciones, a conveniencia de la investigación. En este sentido, parece conveniente dotar al índice de un cierto grado de flexibilización para que la curva de probabilidades pudiera adaptarse mejor a comportamientos hipotéticamente más adecuados para modelar la aparición de los vocablos del núcleo prototípico en las sucesivas posiciones de mención. En consecuencia, proponemos a continuación un par de formulaciones más flexibles, respetando siempre que un vocablo perteneciente al núcleo prototípico debe tender a aparecer prioritariamente en las primeras posiciones de mención con una ley de probabilidades decrecientes en la posición i.

En primer lugar, alterando de manera muy simple la ley básica empleada hasta ahora, proponemos como ley de probabilidad de aparición de los vocablos del núcleo prototípico asociada al índice, la que hace depender las probabilidades de una cierta potencia m del indicador de la posición (en lugar de emplear el indicador directamente, como hasta ahora). Así, la ley sería del tipo:

$$\Pr\left(X=i\right)=\begin{cases}\dfrac{k}{i^{m}} & , \quad si\ 1\le i\le n^{*} \quad ; \quad k>0\,,\,m>0\,,\,n^{*}\ge 1\end{cases} \qquad [28]$$

Con esta definición, el parámetro *m*, que llamaremos parámetro de modulación, permite flexibilizar algo el grado de decaimiento de las correspondientes probabilidades, diferenciando (o igualando) más o menos las probabilidades de las primeras posiciones, principalmente, según se observa en el Gráfico 5.

Gráfico 5: Caída de las probabilidades con la posición (eje horizontal) para diferentes grados de modulación, 'm' (curvas)



Y el valor de *k*, para que esta ley sea una verdadera ley de probabilidad, ahora puede determinarse mediante la expresión:

$$k=\frac{1}{\sum_{i=1}^{n^{*}}\frac{1}{i^{m}}} \qquad [29]$$

Así, prefijado un valor del parámetro "*k*", la siguiente fórmula muestra cómo puede determinarse de forma general el valor de "*m*", que consigue que la probabilidad asociada a una posición determinada "n_p" baje hasta una proporción "*p*" de la probabilidad asociada a la primera posición (y a su lado un ejemplo para el caso en que se desea que en la posición 3ª, la probabilidad sea el 90% de la de la 1ª):

$$m=-\frac{\ln\left(p\right)}{\ln\left(n_{p}\right)} \quad ; \quad m=-\frac{\ln\left(0.9\right)}{\ln\left(3\right)}=0,0959\approx 0,1 \qquad [30]$$

Y escalando el índice resultante para que tome el valor 1 para el vocablo de máxima disponibilidad, llegamos a proponer el siguiente Índice Estandarizado de Compatibilidad Léxica (en adelante, COMP1_st). A modo de ejemplo, en la Tabla 9 se presentan los resultados obtenidos sobre las tres muestras de prueba para un par de combinaciones paramétricas:

$$C^{st}_{1;n^*,m}\left(V_j\right)=\frac{1-\prod_{i=1}^{n^*}\left(1-\frac{k}{i^m}\right)^{100\frac{F_{ji}}{I_1}}}{1-\left(1-k\right)^{100}} \quad ,con \quad k=\frac{1}{\sum_{i=1}^{n^*}\frac{1}{i^m}} \qquad [31]$$

Tabla 9: Compatibilidades en las tres muestras al aplicar el índice COMP1_st

	COMP1_st (n*=50, m=1)			COMP1_st (n*=50, m=0.1)		
	10i	**10x10i**	**10x10i50p**	**10i**	**10x10i**	**10x10i50p**
Voc1MaxDisp	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
Voc2MaxDisp	0,999992	0,999992	0,999992	0,985687	0,985687	0,985687
calle(s)	0,975930	0,975930	0,975930	0,900685	0,900685	0,900685
…						
Piscina	0,065347	0,065347	0,065347	0,185968	0,185968	0,185968
VocMinDisp (pos35)	0,061719	0,006350		0,184968	0,020115	
VocMinDisp (pos50)			0,004445			0,019410

En segundo lugar, proponemos alterar algo más la ley básica empleada hasta ahora, adoptando como ley de probabilidad de aparición de los vocablos del núcleo prototípico asociada al índice, una forma similar a la ya introducida para el Índice de Disponibilidad Léxica Estandarizado, que hace que las probabilidades decaigan a un ritmo modulado por un parámetro *m*, para llegar en una determinada posición *k*, a alcanzar un nivel igual a la proporción *w* de la probabilidad asignada a la primera posición; luego seguir decreciendo hasta la última posición permisible para los vocablos del núcleo prototípico, n*. Así, la ley sería del tipo:

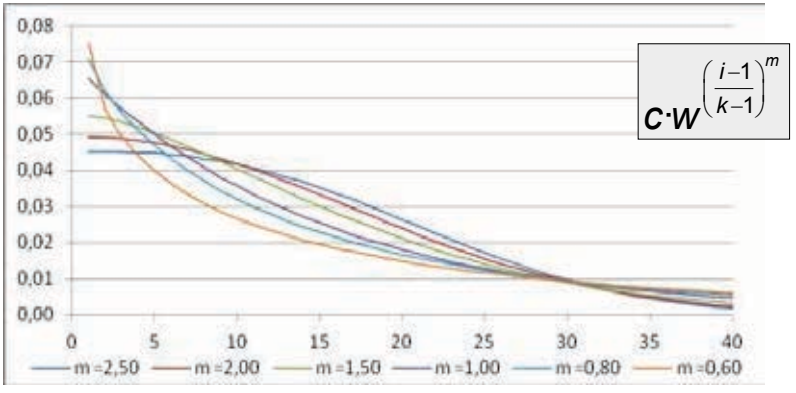
$$\Pr\left(X=i\right)=\begin{cases} C\cdot w^{\left(\frac{i-1}{k-1}\right)^m} & , \quad 1\le i\le n^* \quad ; \quad C>0\,,\,0<w<1\,,\,k<n^*\,,\,m>0 \end{cases} \qquad [32]$$

donde, para que esta ley sea una verdadera ley de probabilidad, el parámetro *C* no puede ser cualquiera, sino que debe tomar el valor:

$$C = \frac{1}{\sum_{i=1}^{n^*} w^{\left(\frac{i-1}{k-1}\right)^m}} \qquad [33]$$

Con esta definición, el parámetro *m* (parámetro de modulación) permite flexibilizar aún más el grado de decaimiento de las correspondientes probabilidades, imprimiendo un mayor o menor ritmo relativo de decaimiento en las primeras posiciones con respecto de las más posteriores, según se observa en el Gráfico 6 (para w=0.10, k=35 y n^*=300).

Gráfico 6: Caída de las probabilidades con la posición (eje horizontal) para diferentes grados de modulación, 'm' (curvas)



De manera análoga a lo expuesto en el caso de los pesos del Índice Estandarizado de Disponibilidad Léxica de una Palabra (IDLP_st), prefijado uno de los valores de los parámetros w y k, puede determinarse de forma general el valor de m que consigue mantener las probabilidades que actúan sobre cada posición, desde la primera hasta una cierta posición k^*, por encima de una cierta proporción w^* de la probabilidad asignada a la primera posición, empleando la expresión dada en la fórmula [13].

Y escalando el índice resultante para que tome el valor 1 para el vocablo de máxima disponibilidad, llegamos a proponer el siguiente Índice de compatibilidad léxica estandarizado (en adelante, COMP2_st). A modo de ejemplo, en la Tabla 10 se presentan los resultados obtenidos sobre las tres muestras de prueba para un par de combinaciones paramétricas:

$$C^{st}_{2;n^*,w,k,m}(V_j) = \frac{1-\prod_{i=1}^{n^*}\left(1-C\cdot w^{\left(\frac{i-1}{k-1}\right)^m}\right)^{100\frac{F_{ji}}{I_1}}}{1-(1-C)^{100}} \quad , con \quad C = \frac{1}{\sum_{i=1}^{n^*} w^{\left(\frac{i-1}{k-1}\right)^m}} \quad [34]$$

Tabla 10: Compatibilidades en las tres muestras al aplicar el índice COMP2_st

	COMP2_st (n^*=50 , w=0.1, k=35, m=1)			COMP2_st (n^*=50 , w=0.1, k=33, m=2.54)		
	10i	**10x10i**	**10x10i50p**	**10i**	**10x10i**	**10x10i50p**
Voc1MaxDisp	1,000000	1,000000	1,000000	1,000000	1,000000	1,000000
Voc2MaxDisp	0,999455	0,999455	0,999455	0,999987	0,999987	0,999987
calle(s)	0,984590	0,984590	0,984590	0,985287	0,985287	0,985287
…						
Piscina	0,075017	0,075017	0,075017	0,047071	0,047071	0,047071
VocMinDisp (pos35)	0,065803	0,006783		0,032303	0,003278	
VocMinDisp (pos50)			0,002456			0,000054

6. COMPARABILIDAD DEL ÍNDICE DE DESCENTRALIZACIÓN LÉXICA DE UN INFORMANTE (IDD) DE ÁVILA MUÑOZ Y SÁNCHEZ SÁEZ

Con objeto de medir hasta qué punto la disponibilidad léxica de un informante se limita a los vocablos del núcleo prototípico, o si por el contrario, utiliza vocablos poco comunes entre el resto de informantes, Ávila Muñoz y Sánchez Sáez proponen su Índice de Descentralización Léxica de un Informante en el centro (IDD) calculado para cada informante mediante la siguiente expresión, inspirada en la formulación empleada por estos mismos autores para medir la compatibilidad léxica de un vocablo:

$$IDD_c\left(Inf_j\right) = 1 - \prod_{i=1}^{P_j}\left(1 - k\left(1 - COMP_K\left(Pal_{j,i}\right)\right)\right) \qquad [35]$$

siendo:

Inf_j = el informante j-ésimo de la muestra

P_j = número de palabras que mencionó el informante j-ésimo, en el centro

i = indicador de la posición en que el informante j-ésimo mencionó alguna palabra, en el centro

k = constante de ajuste para controlar la magnitud de los resultados

$Pal_{j,i}$ = palabra mencionada en la posición i-ésima por el informante j-ésimo, en el centro

$COMP_K(\cdot)$ = Índice de compatibilidad léxica de un vocablo (Ávila Muñoz y Sánchez Sáez)

Los problemas ya señalados para el Índice de Compatibilidad Léxica de un Vocablo de Ávila Muñoz y Sánchez Sáez, desde la perspectiva de la comparabilidad de resultados, se transmiten ahora al Índice de Descentralización Léxica de un informante. Como puede apreciarse en la Tabla 11, se vuelve a producir una cierta dependencia de los resultados con el tamaño de la muestra empleada, lo que supone un primer inconveniente.

Por otra parte, y como los propios autores advierten[9], para aplicar esta fórmula es preciso utilizar, en cada ocasión, un valor de *k* que module convenientemente los resultados obtenidos mediante esta fórmula, ya que si como es deseable, los indicadores de compatibilidad de los vocablos se miden en una escala con una adecuada graduación de resultados entre 0 y 1, sus productos tenderían rápidamente a 0 a medida que aumente el tamaño de la lista de palabras que mencionara el informante, incluso cuando tuvieran un grado importante de compatibilidad (piénsese que si no se impusiera un valor modulador de *k* en esta fórmula, un informante cuya lista estuviera constituida exclusivamente por 10 palabras de compatibilidad 0.8, compatibilidad ésta

[9] En palabra de los autores "[…] este proceso no está exento de problemas: uno de ellos es que el índice siempre tiende a crecer, nunca a disminuir […]", es por esto que "[…] se ponderan las descentralizaciones iniciales por un valor *k* para controlar el proceso de incremento; en principio hemos tomado este valor k de forma experimental como 0.2." (Ávila Muñoz y Sánchez Sáez, 2010: 79).

relativamente alta, obtendría un indicador de descentralización bastante alto, aproximadamente 0.9). La necesidad de elegir este *k* en cada ocasión y su dependencia de las longitudes de las listas de palabras en cada estudio, en nuestra opinión, supone un segundo inconveniente para la aplicación de esta fórmula, desde la perspectiva de su comparabilidad con otros estudios.

Tabla 11: Descentralización Léxica (IDD) de los Informantes para el centro, en las tres muestras

Cálculos usando la medida de compatibilidad COMP (para K=0,16)

	IDD (k=0,2)		
	10i	10x10i	10x10i50p
Informante 1	0,898506	0,594491	0,594491
Informante 2	0,985778	0,920546	0,920546
Inform_Mín_Dispon	0,958228	0,916106	0,916106
Informante 4	0,977795	0,883404	0,883404
Informante 5	0,970669	0,918615	0,918615
Inform_Máx_Dispon	0,992560	0,951966	0,951966
Inform_más_locuaz	0,999212	0,993945	0,999784
Informante 8	0,920299	0,696690	0,696690
Informante 9	0,991214	0,956524	0,956524
Informante 10	0,983476	0,919134	0,919134

6.1. Índice estandarizado de Descentralización Léxica de un Informante

Con la idea de mitigar estos inconvenientes, realizamos aquí una propuesta de índice estandarizado para medir la descentralización léxica de un informante, a partir de ciertas modificaciones del propuesto por los autores originales. Para solucionar el primero de los citados inconvenientes proponemos emplear, en lugar de los índices de compatibilidad léxica de los vocablos de Ávila Muñoz y Sánchez Sáez, alguno de los índices estandarizados de compatibilidad léxica propuestos en el apartado anterior que han sido convenientemente parametrizados.

Para solucionar el segundo de ellos, hemos retomado las consideraciones originales de Ávila Muñoz y Sánchez Sáez (2010: 78). Como estos autores señalan, el Índice de Compatibilidad Léxica de un informante mide la

acumulación de la descentralidad del léxico mencionado por él mismo, y proponen dos procedimientos de entre los que finalmente eligen el segundo que deriva en la fórmula presentada en el apartado anterior. El primer procedimiento, que descartan, consistía en sumar las descentralidades de cada término mencionado en un centro de interés. A pesar de que opinan que este procedimiento permite calcular fácilmente el indicador e interpretar directamente los resultados, lo descartan fundamentalmente porque, a diferencia del segundo procedimiento, el índice no establece unos límites bien definidos.

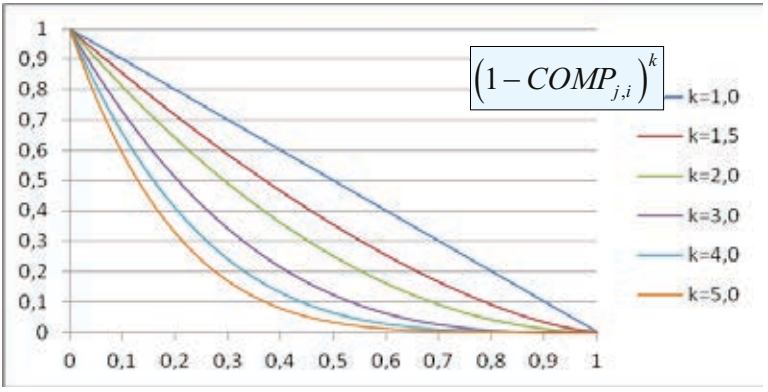
En nuestra opinión, tanto la indefinición de estos límites para la vía desechada, como la necesidad de elegir el parámetro que gradúe adecuadamente los resultados que proporciona el índice finalmente adoptado por aquellos autores, son parte de un mismo problema: ambos procedimientos calculan la descentralidad acumulada a partir de la acumulación de ciertas evaluaciones de las compatibilidades de las palabras mencionadas por el informante (en un caso sumando y en otro multiplicando), para evaluar un efecto total, sin relativizar este acúmulo con respecto del número de palabras que el informante menciona.

En este sentido, el término que en la fórmula propuesta por dichos autores para la evaluación de las compatibilidades, $1\text{-}k(1\text{-}COMP_K)$, es de tipo lineal, lo que significa que le da la misma importancia a una cierta variación absoluta en la compatibilidad de un vocablo, $COMP_K$, independientemente de su nivel. Sin embargo, sería posible pensar que palabras con niveles altos de compatibilidad con el núcleo prototípico (pongamos, por ejemplo, con valores de $COMP_K$ entre 0.8 y 1.0) deberían influir poco en la medición de la descentralidad del individuo que las menciona, y de manera más o menos parecido, mientras que palabras más específicas (digamos con valores entre 0.1 y 0.3) no solo deberían influir mucho más que las anteriores, sino que también progresivamente más cuanto más específica sea la palabra. Es por ello que proponemos incorporar en la formulación un término más flexible para la evaluación de las especificidades de los vocablos, de la forma:

$$\left(1 - COMP_{j,i}\right)^k \qquad [36]$$

Así, como podemos ver en el Gráfico 7, la contribución que realizaría un vocablo de compatibilidad determinada (medida en el eje horizontal de abscisas) a la evaluación de la descentralización del individuo (medida en el eje vertical de ordenadas) podría modularse convenientemente mediante el parámetro *k*, de forma que los vocablos más compatibles contribuyan poco y con efectos parecidos, mientras que las vocablos menos compatibles contribuyan mucho más, y tanto más, cuanto más específicos sean.

Gráfico 7: Contribución de la compatibilidad de un vocablo para diferentes potencias 'k' (curvas)



Siempre puede calcularse el valor de k que garantice que dichas contribuciones a la especificidad, para que los vocablos con compatibilidad mayor o igual que un valor de referencia c_0 dado, sea menor o igual que otro valor de referencia ε dado.

Así, la siguiente fórmula muestra cómo puede determinarse de forma general dicho valor de "k" para tales circunstancias (y a su lado un ejemplo ilustrativo para cuando se quiere que la contribución de un vocablo sea de 0.01, a lo sumo, si su compatibilidad es mayor o igual que 0.8):

$$k \geq \frac{\ln(\varepsilon)}{\ln(1-c_0)} \quad ; \quad k \geq \frac{\ln(0.01)}{\ln(1-0.8)} = 2,861353 \approx 3 \qquad [37]$$

En consecuencia, y no queriendo abandonar ninguna de las dos vías de medición propuestas originariamente por Ávila Muñoz y Sánchez Sáez, proponemos las dos que se presentan a continuación. En primer lugar, el Índice Estandarizado Multiplicativo de Descentralización Léxica de un Informante, para un centro c (en adelante, $IDD_{m\_}st$):

$$IDD_{m}\_st_{c,k}\left(Inf_j\right) = 1 - \left\{\prod_{i=1}^{P_j}\left(1-\left(1-COMP_{j,i}\right)^k\right)\right\}^{1/P_j} \qquad [38]$$

Y en segundo lugar, el Índice Estandarizado Aditivo de Descentralización Léxica de un Informante, para un centro c (en adelante, $IDD_{a\_}st$):

$$IDD_{a}\_st_{c,k}\left(Inf_{j}\right)=1-\sum_{i=1}^{P_j}\frac{1-\left(1-COMP_{j,i}\right)^{k}}{P_j} \qquad [39]$$

Como ilustración del comportamiento de estos nuevos índices propuestos, los hemos calculado para las parametrizaciones *k=1* y *k=3*, para las tres muestras de informantes creadas artificialmente, y presentamos sus resultados más representativos en las Tablas 12 y 13.

Tabla 12: Descentralización Léxica de los Informantes para el centro, en las tres muestras

Cálculos usando la medida de compatibilidad COMP_st (para K=0,16)

	IDD$_m$_st (k=1)			IDD$_m$_st (k=3)		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
Informante 1	0,398305	0,398305	0,398305	0,134152	0,134152	0,134152
Informante 2	0,661483	0,661483	0,661483	0,388614	0,388614	0,388614
Inform_Mín_ Dispon	0,786904	0,786904	0,786904	0,531052	0,531052	0,531052
Informante 4	0,632305	0,632305	0,632305	0,361416	0,361416	0,361416
Informante 5	0,751958	0,751958	0,751958	0,495261	0,495261	0,495261
Inform_Máx_ Dispon	0,692297	0,692297	0,692297	0,426532	0,426532	0,426532
Inform_ más_locuaz	0,785234	0,798791	0,939129	0,549041	0,577017	0,857789
Informante 8	0,483332	0,483332	0,483332	0,194422	0,194422	0,194422
Informante 9	0,733485	0,733485	0,733485	0,480826	0,480826	0,480826
Informante 10	0,634394	0,634394	0,634394	0,319818	0,319818	0,319818

En la Tabla 12 se muestran los resultados obtenidos para la versión multiplicativa del índice. Como podemos ver en ella, el comportamiento del índice para los informantes con mayor disponibilidad léxica (Inform Max Dispon) y menor disponibilidad léxica (Inform Min Dispon) se mantiene para las tres muestras, variando ligeramente la del informante que hemos llamado el "más locuaz", por motivos análogos a los argumentados cuando comentábamos la Tabla 5. En cualquier caso, se observa una gran estabilidad en los índices de las tres muestras a pesar de sus diferencias en el número de informantes y en las longitudes máximas de sus listas, como se pretendía con la propuesta.

Por otra parte, en la Tabla 13 se presentan de manera análoga los resultados obtenidos para la versión aditiva del índice, observándose los comportamientos estables pretendidos para estos índices, a pesar de las diferencias en las muestras empleadas, deseables desde la perspectiva de la comparabilidad de los estudios.

Tabla 13: Descentralización Léxica de los Informantes para el centro, en las tres muestras

Cálculos usando la medida de compatibilidad COMP_st (para K=0,16)

	IDD_a_st (k=1)			IDD_a_st (k=3)		
	10i	10x10i	10x10i50p	10i	10x10i	10x10i50p
Informante 1	0,328467	0,328467	0,328467	0,105489	0,105489	0,105489
Informante 2	0,532492	0,532492	0,532492	0,303200	0,303200	0,303200
Inform_Mín_ Dispon	0,708865	0,708865	0,708865	0,483083	0,483083	0,483083
Informante 4	0,497741	0,497741	0,497741	0,277106	0,277106	0,277106
Informante 5	0,639337	0,639337	0,639337	0,422264	0,422264	0,422264
Inform_Máx_ Dispon	0,560776	0,560776	0,560776	0,330379	0,330379	0,330379
Inform_ más_locuaz	0,668341	0,669490	0,767515	0,442734	0,446013	0,608839
Informante 8	0,399402	0,399402	0,399402	0,163224	0,163224	0,163224
Informante 9	0,600863	0,600863	0,600863	0,383575	0,383575	0,383575
Informante 10	0,556515	0,556515	0,556515	0,275606	0,275606	0,275606

En consecuencia, las nuevas fórmulas propuestas generalizan en esencia la correspondiente fórmula original de Ávila Muñoz y Sánchez Sáez (2010) y permiten comparar los resultados de distintas investigaciones, aunque presenten diferente número de informantes y distintas longitudes máximas de las listas de palabras recogidas.

7. CONCLUSIONES

Desde sus inicios, los estudios de disponibilidad léxica han apostado por la utilización de herramientas matemáticas que permitan una medida fiable de la disponibilidad, entendida como el grado de inmediatez con que actualizamos las palabras relacionadas con un área temática o centro de interés. Y es precisamente la necesidad de reflejar "ese grado de inmediatez" lo que ha llevado a autores como Butrón, López Morales, Lorán, López Chávez y Strassburger Frías, Ávila Muñoz y Sánchez Sáez, entre otros, a proponer fórmulas matemáticas que sirven para determinar la relación que existe entre las palabras, los informantes y los centros de interés cuando son evocados en las pruebas de disponibilidad léxica.

Sin embargo, al aplicar dichas fórmulas en la investigación sobre el *Léxico disponible de estudiantes de español como Lengua Extranjera en la Comunidad de Madrid* desarrollada por Gallego Gallego, se apreciaron algunos inconvenientes en sus resultados que han sido analizados en detalle en los apartados 3, 4, 5 y 6 del presente trabajo. En concreto, se observó que estas medidas se ven influidas por factores como el número de informantes o la longitud de la lista de palabras en el centro (a veces provocada por un simple comportamiento atípico de uno de los informantes), lo que puede producir indeterminación de los límites en sus rangos de medida y, en consecuencia, dificulta su utilización para comparar resultados de diversos estudios realizados sobre diferentes muestras de informantes.

Considerando que cualquier estudio o propuesta que se haga en torno al mecanismo matemático subyacente a los estudios de disponibilidad debe partir por reconocer y utilizar los planteamientos anteriores de aquellos autores, hemos propuesto algunas formulaciones alternativas con el ánimo de corregir aquéllos inconvenientes detectados en las formulaciones originales, especialmente tratando de facilitar la comparabilidad de los resultados obtenidos en investigaciones de disponibilidad léxica realizadas con diferentes muestras de informantes. En este sentido, en los apartados 3.1, 4.1, 5.1 y 6.1 hemos propuesto algunos índices de disponibilidad y compatibilidad léxica estandarizados que generalizan las fórmulas originales, independientes del número de informantes y las longitudes de los listados de vocablos.

Como no puede ser de otra forma, la valoración de la utilidad de estas nuevas formulaciones propuestas debe ser realizada, mediante su aplicación en otros estudios, por la comunidad científica en el campo de la disponibilidad léxica. De esta forma, se podría avanzar en la búsqueda de unas parametrizaciones concretas y aceptadas de forma general por la

comunidad científica –que quedan abiertas en este trabajo–, que condujeran a la utilización de expresiones matemáticas comúnmente aceptadas y que permitan la comparación fiable entre diferentes estudios.

REFERENCIAS BIBLIOGRÁFICAS

Ávila Muñoz, Antonio Manuel y José María Sánchez Sáez. 2010. La disponibilidad léxica. Antecedentes y fundamentos. En Ávila Muñoz y Villena Ponsoda (eds.). *Variación social del léxico disponible en la ciudad de Málaga: diccionarios y análisis*, pp. 37 - 81. Málaga: Editorial Sarriá, S. L.

Azurmendi Ayerbe, María José. 1983. *Elaboración de un modelo para la descripción sociolingüística del bilingüismo y su aplicación parcial a la comarca de San Sebastián, Guipúzcoa*. San Sebastián: Caja de Ahorros Provincial de Guipúzcoa.

Butrón, Gloria. 1987. *El léxico disponible: índices de disponibilidad*. Tesis doctoral inédita. Río Piedras: Universidad de Puerto Rico.

_____ 1991. Nuevos índices de disponibilidad léxica. En López Morales (ed.). *La enseñanza del español como lengua materna. Actas del II Seminario sobre "Aportes de la lingüística a la enseñanza de la lengua materna"*, pp. 79 - 89. Río Piedras: Universidad de Puerto Rico.

Cañizal Arévalo, Alva. 1987. *Disponibilidad léxica en escolares de primaria terminada. Análisis de seis centros de interés*. Tesina inédita. México: Universidad Nacional Autónoma de México.

Dimitrijévic, Naum. 1969. *Lexical Availability*. Heidelberg: Julius Gross Verlag.

Echeverría, Max, Mª Olivia Herrera, Patricio Moreno y Francisco Pradenas. 1987. Disponibilidad léxica en Educación Media. *Revista de Lingüística Teórica y Aplicado*, 25: 55-115.

Gallego Gallego, Diego Javier. 2014. *Léxico disponible de estudiantes de español como lengua extranjera en la Comunidad de Madrid*. Tesis doctoral inédita. Alcalá de Henares: Universidad de Alcalá.

Gougenheim, Georges, René Michéa, Paul Rivenc y Aurélien Sauvageot. 1956. *L´elaboration du français élémentaire. Étude sur l`établissement d´un vocabulaire et d´une grammaire de base*. Paris: Didier.

_____ 1964. *L`elaboration du français fondamental (1er degrée). Étude sur l`établissement d`un vocabulaire et d`une grammaire de base*. Paris: Didier.

Justo Hernández, Hortensia. 1986. *Disponibilidad léxica en colores*. Tesina inédita. México: Universidad Nacional Autónoma de México.

Lakoff, G. 1987. *Women, fire and dangerous things: What categories reveal about the mind*. Chicago-Londres: The Universtity of Chicago Press.

López Chávez, Juan y Carlos Strassburguer Frías. 1987. Otro cálculo del índice de disponibilidad léxica. En *Actas del IV Simposio de la Asociación Mexicana de Lingüística Aplicada, Presente y perspectiva de la lingüística computacional en México.* México: Universidad Nacional Autónoma de México.

_____ 1991. Un modelo para el cálculo del índice de disponibilidad léxica individual. En López Morales (ed.). *La enseñanza del español como lengua materna. Actas del II*

Seminario sobre "Aportes de la lingüística a la enseñanza de la lengua materna", pp. 91-112. Río Piedras: Universidad de Puerto Rico.

López Morales, Humberto. 1973. *Disponibilidad léxica en escolares de San Juan*. MS.

_____ 1978. Frecuencia léxica, disponibilidad y programación curricular. En Humberto López Morales (ed.). *Aportes de la Lingüística a la Enseñanza del Español como Lengua Materna*, pp. 73-86. Edición especial de BAPLE, 6.

_____ 1979. Disponibilidad léxica y estratificación socioeconómica. *Dialectología y Sociolingüística. Temas puertorriqueños*. Madrid: Hispanova de Ediciones.

_____ 1999. *Léxico disponible de Puerto Rico*. Madrid, Arco Libros.

Lorán, Roberto y Humberto López Morales. 1983. *Nouveau calcul de l´indice de disponibilité*. MS.

Mackey, William C. 1971. *Le vocabulaire disponible du français* (2 Vols.). Paris – Bruxelles – Montreal: Didier.

Mena Osorio, Mónica. 1986. *Disponibilidad léxica infantil en tres niveles de enseñanza básica*. Tesis de maestría inédita. Concepción: Universidad de Concepción.

Román, Belén. 1985. *Disponibilidad léxica en escolares de Dorado*. Puerto Rico. Memoria de licenciatura inédita. Río Piedras: Universidad de Puerto Rico.

Rosch, E. 1978. Principles of categorization. En E. Rosch y B. Lloyd (eds.). *Cognition and categorization,* pp. 27-48. Hillsdale: Lawrence Erlbaum.

Wittgenstein, L. 1953. *Philosophical invetigations*. Nueva York: McMillan.