EL DISEÑO DE UNA FÓRMULA MATEMÁTICA PARA OBTENER UN ÍNDICE DE DISPONIBILIDAD LÉXICA CONFIABLE

Ha habido intentos diversos de ordenar de acuerdo con modelos matemáticos o estadísticos las respuestas obtenidas en distintas encuestas léxicas. A nosotros en un principio nos llamó la atención el planteamiento realizado en 1982 por Humberto López Morales y Roberto Lorán[1], planteamiento que se caracteriza por la búsqueda de un alto grado de formalización y —aun— de automatización. Entusiasmados por lo novedoso de tal propuesta, decidimos iniciar una serie de encuestas léxicas que nos permitieran conocer y ensayar esa nueva metodología, al mismo tiempo que podríamos obtener un reflejo del componente léxico del español de México.

Sin embargo, ya al trabajar las primeras muestras encontramos ciertas imprecisiones en la fórmula realizada por los profesores mencionados, razón por la cual empezamos a llevar a efecto toda una serie de pruebas, trabajos e investigaciones de laboratorio de la que publicamos ahora los resultados más relevantes. El objetivo principal de este artículo[2] es, pues, poner a la conside-

[1] Que hemos conocido por diversos manuscritos pero que no hemos podido consultar hasta la fecha —es decir hasta el momento de la edición del artículo que presentamos ahora— en ninguna publicación.

[2] Cuya versión preliminar se leyó en 1987 en el IV Simposio de la Asociación Mexicana de Lingüística Aplicada: "La lingüística compu-

ración de los lectores algunas modificaciones sustanciales que le hemos realizado, con base en nuestra experiencia en el tratamiento de los datos léxicos, a la fórmula del cálculo del índice de disponibilidad léxica propuesta en 1982 —lo dijimos ya—.

Como finalidades complementarias, aunque tal vez más funcionales, queremos mostrar las vicisitudes que hemos pasado en el tratamiento automatizado, desde la captura de datos hasta la presentación final, incluyendo la formulación de los diversos programas computacionales. Terminaremos con un extracto —para no extendernos excesivamente— de algunos resultados obtenidos —como acabamos de explicitarlo justo en el párrafo anterior—.

Es necesario, antes de entrar propiamente en materia, hacer algunas consideraciones.

Los diversos refinamientos a los que ha llegado la léxico-estadística han producido un número ya amplio de fórmulas, índices y factores que se aplican sobre los materiales recogidos en los textos con la finalidad de conseguir ciertas medidas y agrupaciones que han sido de alto rendimiento en la lexicografía, la lingüística e incluso la dialectología. Dos conceptos muy vigorosos surgidos de estas investigaciones son los de léxico *básico* y léxico *disponible*. Su uso y sus aplicaciones se han generalizado ampliamente por todo el ámbito de los estudios lexicológicos.

tacional en México". Hasta hoy había permanecido inédito, pero ahora nos hemos decidido a darlo a la luz porque tenemos la seguridad de que resultará de interés para quienes lean esta revista, dado que —por esgrimir una razón de peso— el pasado año se publicó el libro *El Centro de Lingüística Hispánica y la Lengua Española. Volumen conmemorativo del 30 aniversario de su fundación* (México, UNAM, 1999), en que uno de nosotros —Juan López Chávez— tuvo la fortuna de participar con el trabajo "Consideraciones acerca del índice de disponibilidad léxica" (pp. 419-430), investigación de que el presente estudio es antecedente y complemento.

Ocupémonos primero de la noción de léxico básico: con el afán de conocer el vocabulario usual de una comunidad hablante dada se diseñaron muestras de los textos que ella producía; de ahí se obtuvieron largas listas de frecuencias que eran ponderadas por un factor de dispersión con la finalidad de obtener los vocablos más usuales, en los diversos géneros de escritura, en una época precisa. Al resultado conseguido se le llamó léxico básico.

Sin embargo, pronto los estudiosos se dieron cuenta de que a pesar del enorme tamaño de las muestras y de lo apropiado de la aplicación de la medida de dispersión, no se lograban registrar algunas palabras que la inmensa mayoría de los individuos pertenecientes a la comunidad lingüística en cuestión conocía y empleaba, pero cuyo uso dependía de situaciones comunicativas concretas. Para dar cuenta de este campo léxico específico surgió el concepto de léxico disponible.

Hablemos brevemente, pues, de esto. Todo parisino —y es que fue en Francia donde originalmente se hicieron los trabajos lexicológicos a que hacemos referencia— con seguridad conocía y usaba palabras como *metro, autobus, lettre* o *timbre*; no obstante, en las listas que sirvieron de base para la elaboración del *Français Fondamental* no era posible encontrarlas. Por cierto que en un diccionario moderno y actual como el publicado por El Colegio de México —*Diccionario básico del español de México*[3]— sus equivalentes en español sí aparecen; sin embargo, palabras como *sandía, melón, fresa, salchicha, motocicleta, taxi, helicóptero, cabello, costilla, enfermera* o *anaranjado* —que nuestras investigaciones nos han señalado como familiares y disponibles para cualquier hablante de la ciudad de México— no están registradas en la obra mencionada. Este señalamiento debe considerarse —aclaremos— más como una ejemplificación que como una aco-

[3] México, El Colegio de México, 1991.

tación o crítica: una comparación de esta diversidad de datos necesita —por una parte— de una indagación más exhaustiva de la disponibilidad concreta de los hablantes de México, y —por otra parte— de un análisis más profundo del *Diccionario básico.* Sin embargo, algunas de las conclusiones consignadas por Marcela Castellanos en *Presencia de léxico escolar disponible en los diccionarios del español de México*[4] son extremadamente contundentes e incontrovertibles. Y es que, aunque la muestra del corpus de un diccionario sea muy ambiciosa y esté muy bien elaborada, ciertas parcelas del lexicón serán captadas sólo muy pobremente. Es por eso que las investigaciones acerca del léxico disponible de escolares mexicanos —publicadas incluso ya en partes[5]— podrían haberse tomado en cuenta para la elaboración del después impreso *Diccionario del español usual en México*[6] y habrían cubierto con eficiencia importantes omisiones[7].

Es así que precisamente para compensar las ausencias que se han detectado en los diccionarios se haya propuesto tener como auxiliares —específicamente— los valores que arroja la disponibilidad de una palabra. Se diseñaron, con tal fin, encuestas que con base en

[4] Tesis de licenciatura inédita, México, UNAM, 1998.

[5] JUAN LÓPEZ CHÁVEZ y ROSA MARÍA MEZA CANALES, *Léxico disponible de preescolares,* México, UNAM-Alhambra, 1993; JUAN LÓPEZ CHÁVEZ y MARÍA TRINIDAD MADRID GUILLÉN, *Léxico disponible de primer grado de primaria,* México, UNAM-Alhambra, 1993; JUAN LÓPEZ CHÁVEZ y MARTHA JULIÁN PEÑA, *Léxico disponible de segundo grado de primaria,* México, UNAM-Alhambra, 1993; JUAN LÓPEZ CHÁVEZ y ROSALÍA BOLFETA MONTES DE OCA, *Léxico disponible de tercer grado de primaria,* México, UNAM-Alhambra, 1993; JUAN LÓPEZ CHÁVEZ y MARCELA FLORES CERVANTES, *Léxico disponible de cuarto grado de primaria,* México, UNAM-Alhambra, 1993; JUAN LÓPEZ CHÁVEZ y LILIA CASTELLANOS MEDINA, *Léxico disponible de quinto grado de primaria,* México, UNAM-Alhambra, 1993; JUAN LÓPEZ CHÁVEZ y EVA NÚÑEZ ALONSO, *Léxico disponible de sexto grado de primaria,* México, UNAM-Alhambra, 1993.

[6] México, El Colegio de México, 1996.

[7] Ejemplifiquemos con *carroza, jerga, joyero, lapicero, maratón, palillo, tobimedias.*

conceptos fijos y bien estudiados marcaban unos "centros de interés" que obligaban a reproducir por asociación el léxico usual en determinadas situaciones comunicativas que, por un lado, eran situaciones frecuentes en la vida cotidiana y que, por otro, eran tan cotidianas que no había textos de fácil captura que las reprodujeran.

De esta manera, para conseguir la obtención de esta importante parcela del léxico se eligieron 16 centros de interés[8] que fueron los más generalizadamente aceptados —porque hay quienes han trabajado con más, con menos, o con diferentes centros—.

Vale decir que en las primeras investigaciones de disponibilidad el único parámetro considerado era la frecuencia de aparición. Muy pronto, empero, se modificó esta posición basada en un único criterio y surgieron diferentes propuestas, asunto del que hablaremos más adelante y que precisamente es el punto central de esta exposición.

Por cierto que la técnica misma de encuesta también ha cambiado de manera fundamental: de —en un principio— obtener un número fijo y determinado de palabras por cada centro de interés se pasó a la búsqueda de listas de palabras en número desigual regidas solamente por la predeterminación de la duración de la encuesta, un número específico de minutos por cada centro.

Veamos con un poco de detenimiento ambas posibilidades.

Originalmente, por cada centro de interés se pedía a cada informante que dijera un número fijo de palabras que asociara con este centro —podían ser 20 o 30, u otra cantidad, dependiendo del investigador—. Lo que se esperaba era que señalara las mismas palabras que habría

[8] GUSTAVE GOUGENHEIM *et al.*, *Français Fondamental (1er. degré)*, París, Didier, 1958.

de usar en una situación comunicativa donde se tratara el tema relacionado con el centro en cuestión; por ejemplo, se preguntaba sobre el léxico de los medios de comunicación y de transporte, o de las partes del cuerpo, o de los enseres de la casa, o de las comidas y alimentos, o de cualquier otro centro que las necesidades de la investigación requirieran.

Una vez obtenido el material, se procedía a realizar los cálculos estadísticos. La fórmula que para el caso particular de las muestras con número de respuestas preestablecido era usada por los autores de la propuesta de 1982 se reproduce a continuación.

$$D(P_j) = X_{1j} + CX_{2j} + C^2X_{3j} + C^3X_{4j} + \ldots + C^{p-1}X_{pj}$$

siendo

$$C = 0.9 \text{ y } X = f_{ji}/I_1$$

Como bien puede verse, los parámetros considerados son tres: la frecuencia con que fue dicha la palabra —$\mathbf{f_{ji}}$—, el número de informantes —$\mathbf{I_1}$ (que, considerado como divisor de la frecuencia, da la frecuencia relativa)— y, por último, un ponderador de la posición para valorar el lugar en que fue dicha cada palabra —**C**—.

Para entender este último factor hay que aclarar que se parte de la base de que una palabra tendrá mayor peso si es sugerida en las primeras posiciones que si se obtiene en las últimas.

Por otra parte, cuando se utiliza un tiempo límite para otener la muestra, necesariamente se tienen como resultado listas de tamaño diferente, ya que cada uno de los informantes cuenta con la oportunidad de emitir **n** palabras en **n** minutos, y la probabilidad de que haya diferencias existe y es inmediata. Y aunque en este caso el problema estadístico es semejante, al parecer intervienen en el asunto factores que merecen un análisis detallado.

Pero volvamos por ahora al punto y transcribamos la fórmula que para las muestras con tiempo específico se sugirió en 1982.

$$D(P) = \sum_{i=1}^{n} \lambda^{(i-1)} \ (N_i/N_{i-1})X_i$$

donde

$$N_0=N_1$$

donde *lambda* es el factor de ponderación de la posición, con un valor determinado empíricamente por expertos de **0.9** elevado a una potencia igual a la posición menos uno. N_i entre N_{i-1} es el factor que pondera el tamaño desigual de la lista; y X_i es la frecuencia relativa.

Al aplicar esta fórmula encontramos —según nuestra interpretación, claro— que en realidad se trataba de una mera simplificación de la otra, pues comprende un quebrado igual a la unidad que no aporta función alguna; por lo tanto, podría quedar del modo siguiente:

$$D(P) = \sum_{i=1}^{n} \lambda^{(i-1)} \ (F_{ji}/N_{i-1})$$

porque

$$X_i=f_{ji}/N_{i-1}$$

Pero inmediatamente salta a la vista una adecuación caprichosa. En efecto, ¿por qué obtener la frecuencia relativa dividiendo la frecuencia absoluta entre el número de informantes que llegaron a la posición inmediatamente anterior? No se explica.

Aunado a esto, nos topamos con incongruencias en la clasificación, ya que —por ejemplo— en una matriz donde hubo 49 informantes y la lista más larga obtenida era de 13 posiciones, se encontró lo que sigue:

Frecuencia	Posición	Valor disponible	
1 en	9ª	0.061495	←
1 en	8ª	0.053144	
1 en	7ª	0.037960	
1 en	6ª	0.028119	
1 en	5ª	0.025235	
1 en	1ª	0.022222	←
1 en	2ª	0.020000	
1 en	3ª	0.019286	
1 en	4ª	0.019184	

Se evidencia así que no existe ninguna congruencia en los resultados. La explicación es muy simple, en realidad: todo depende del número de informantes y de la proporción que guarden en el descenso. Sin embargo, como esto no nos resultó satisfactorio —por razones obvias—, iniciamos un análisis más profundo.

Como veremos en seguida, el camino que preferimos seguir fue el de proponer una única fórmula para cualquier tipo de muestra, es decir, una que fuera igualmente eficaz en el caso de las muestras obtenidas por número fijado previamente y en el de las muestras que resultaran de una cantidad preestablecida de minutos.

Detengámonos, sin embargo, en algunas cuestiones de importancia. Y es que hemos de recordar que nos interesa particularmente hacer partícipe al lector de los vericuetos del camino que hemos seguido para llegar a donde hemos llegado. Sigamos, pues.

Si analizamos la matriz de vectores de frecuencia nos encontramos con que los factores que debemos considerar en el tratamiento estadístico son los siguientes: la

frecuencia con que fue dicha cada palabra en cada posición, la suma de esas frecuencias —que da la frecuencia absoluta de la palabra—, el número de informantes que participaron en la encuesta, el número de informantes que llegó a cada posición y —finalmente— el número de posiciones. Ello sin olvidar, claro, el número de palabras diferentes —es decir, de *vocablos*[9]— obtenidos.

De esta manera, y para comprender todos los valores en cuestión, hemos elaborado —tras cuidadoso diseño— la siguiente fórmula:

$$D(P) = \sum_{i=1}^{n} e^{-2.3 \cdot (i-1/n-1)} \; (f_{ji}/I_i)$$

En ella se obtiene la frecuencia relativa como resultado de dividir la frecuencia absoluta en cada posición entre el número total de informantes; la consideración directa de los informantes que llegan a cada posición para este aspecto no puede mantenerse, como lo veremos más adelante, pues —además— empíricamente no lo pudimos conservar. Sin embargo, sí hay una consideración indirecta, pues el total de frecuencias absolutas entre el total de informantes, resultado de la suma de frecuencias en cada posición, encierra las veces que fue dicha cada palabra, el número de informantes de la muestra y el número de informantes que llega a cada posición, al ponderarlo por un factor que dispersa de 1 a 0.1, que es el sustituto de *lambda* que hemos utilizado. Este último valor está fijado en **–2.3** para lograr una dispersión constante (entre 1 y 0.1) y toma en cuenta cada posición (**i**) y el total de posiciones alcanzadas (**n**).

[9] Se entiende por *palabra* cada una de las variaciones del *vocablo*, al tiempo que éste es la forma que engloba a las susodichas variaciones. Así, *perrita* será una palabra, pero el vocablo correspondiente será *perro*.

De lo dicho se desprende que resulta inútil conservar dos fórmulas para los dos diversos tipos de muestras, ya que la que hemos diseñado puede dar cuenta de ambas sin perder información ninguna.

Estamos convencidos —además— de que con la fórmula que proponemos logramos describir un sistema léxico en el que hay varios elementos conjuntados: la frecuencia de aparición de cada palabra, una norma marcada por todos los informantes, y varias subnormas contenidas en las diferentes agrupaciones descendentes que podemos hacer de acuerdo con los informantes que llegan a cada posición. Así pues, la realidad lingüística quedará reflejada con fidelidad en este tratamiento estadístico.

Nos parece que de más está decir que es justamente esto lo que perseguimos: una precisión de la mayor confiabilidad posible en los instrumentos —que no son más que eso, cosa que jamás olvidamos— que apliquemos a la descripción de la lengua.

Queremos, sin embargo, explicar cómo es la solución computacional que le dimos al problema —que en la práctica era francamente un escollo de mucha consideración— de tener que emplear dos fórmulas —no demasiado eficientes, aparte, como luego podrá verse—.

Interesa especificar que independientemente de la fórmula que se elija para calcular el índice de disponibilidad léxica, el proceso de computación para preparar las bases que permitan dicho cálculo es un problema determinístico, gracias a lo cual fue y es posible efectuar múltiples pruebas simplemente cambiando en el programa de cómputo la expresión que representa a la fórmula que se desee. Conviene decir que dedicamos cientos de horas a la investigación y la experimentación requeridas, pero nuestro esfuerzo dio frutos claros, así que podemos afirmar que hasta donde se puede humanamente estar seguro de algo nosotros lo estamos de que es posible reflejar fielmente el índice de disponibilidad léxica con la ahora llamada fórmula de "López

Chávez-Strassburger", que acabamos de presentar arriba y cuyo proceso de gestación —como dijimos ya— tenemos la intención de explicar en las líneas que siguen[10].

Lo primero de que hay que hablar es de la captura de los materiales. A las encuestas o pruebas para hacer análisis de disponibilidad o riqueza léxica se les diseñó un programa en lenguaje BASIC para capturar la información en un microprocesador con el objeto de minimizar los errores. Los datos que se capturaron son el número de informantes, las claves para identificar a cada sujeto (sexo, edad, escolaridad, turno, ubicación, etcétera) —que sirven para hacer agrupaciones (cf. *infra*)— y la serie de palabras que se habrían de analizar colocándolas en campos con un máximo de 24 casillas, quedando abierto el número de palabras que se podrían capturar por informante. Los datos quedan archivados en un disco flexible.

Pertinente es detallar que tanto los errores ortográficos, las duplicaciones y las incorrecciones de cualquier tipo, así como cuestiones convenidas previamente por los especialistas —la reducción de los plurales a singulares y de las variantes de género a uno solo, por ejemplo— se corrigieron utilizando editores de las computadoras hasta obtener el corpus que a juicio del experto podía ser analizado ya.

Finalmente, lo que se ha conseguido con la captura nítida de los materiales es tener una matriz con los datos proporcionados por los informantes en la forma siguiente:

[10] Por cierto que creemos oportuno señalar que es esta fórmula y no otra la que desde hace ya tiempo —desde su diseño mismo, podríamos decir— se emplea en el mundo hispánico para hacer los cálculos del índice de disponibilidad léxica. En efecto, con ella —y en México, incluso— se hicieron los trabajos estadísticos de los léxicos disponibles de Madrid, de Las Palmas de Gran Canaria, de Puerto Rico y de la República Dominicana.

MATRIZ 1: DATOS DE LOS INFORMANTES

S_1 - $clave_1$ - p_{11} p_{12} p_{13} ··· p_{1a}
S_2 - $clave_2$ - p_{21} p_{22} p_{23} ··· p_{2b}

S_3 - $clave_3$ - p_{31} p_{33} p_{33} ··· p_{3b}
.
.
S_w - $clave_w$ - p_{w1} p_{w2} p_{w3} ··· p_{wb}
.
.
S_x - $clave_x$ - p_{x1} p_{x2} p_{x3} ··· p_{xb}

siendo

Sx = el sujeto o informante x
n = el número de posiciones máximas alcanzadas
pij = la palabra dada por el sujeto i en la posición j

Procedamos ahora a la descripción de los procesos de datos. Digamos primero que con el fin de acelerar el trabajo, se elaboraron los programas con lenguaje ALGOL para un computador Burroughs B-7831 —posteriormente se inició la adaptación en BASIC y PASCAL para microprocesadores, con la finalidad de facilitar el uso de esta herramienta tan útil para el cálculo de disponibilidad léxica; fue hasta hace muy poco, sin embargo, que uno de nosotros (Juan López Chávez) consiguió por fin encontrar la manera de calcular índices de disponibilidad en los nuevos microprocesadores en una simple hoja de cálculo.

Refirámonos antes que nada a un punto específico: la lista desordenada y su clasificación.

Este proceso toma los datos de los informantes (matriz 1) y procede a formar una lista desordenada que se archiva como vector 2, que contiene lo que sigue:

VECTOR 2: LISTA DESORDENADA

P_{11}	P_{23}	•	P_{13}	•
P_{12}	•	•	•	•
P_{13}	•	P_{3c}	•	P_{x1}
•	•	•	•	P_{x2}
•	P_{2b}	•	P_{wn}	P_{x3}
•	P_{31}	•	•	•
P_{1a}	P_{32}	•	•	•
P_{21}	P_{33}	P_{w2}	•	•
P_{22}	•	P_{w2}	•	P_{xd}

De aquí se obtiene el total de palabras (T) suministradas por los informantes, siendo "T" el contador de ocurrencias desde p_{11} hasta p_{xd}, que servirá para obtener la frecuencia relativa.

Este proceso termina al formarse el vector clasificado en orden alfabético que denominamos vector 3, que obtiene T ocurrencias, y queda como sigue:

VECTOR 3: VECTOR CLASIFICADO DE PALABRAS

P_1
•
•
P_I
P_2
•
•
P_2
•
•
•
P_j
•
•

donde y = número de palabras diferentes

P_1 = j-ésima palabra en orden alfabético

•
P_j
•
•
P_y
•
•
•
P_y

Precisemos ahora lo referente a las listas de frecuencias. Del vector 3 clasificado se van contando las ocurrencias de palabras idénticas y se forma la matriz de frecuencias totales absolutas y relativas de cada palabra (matriz 4), que tiene la formación siguiente:

MATRIZ 4: TABLA DE FRECUENCIAS

Vocablo	Frecuencia absoluta	Frecuencia relativa
P_1	FA_1	$FR_1 = FA_1/T$
P_2	FA_2	$FR_2 = FA_2/T$
P_3	FA_3	$FR_3 = FA_3/T$
•	•	•
•	•	•
P_J	FA_J	$FR_J = FA_J/T$
•	•	•
•	•	•
P_Y	FA_Y	$FR_Y = FA_Y/T$

donde:

FA_J = frecuencia absoluta de la j-ésima palabra
FR_J = frecuencia relativa de la j-ésima palabra
T = total de palabras ocurridas
P_J = j-ésima palabra dentro del orden alfabético.

Lista $P_1 \ldots P_y$ = "diccionario" de la encuesta

Partiendo de aquí, el programa de computación emite dos listados, uno en el mismo orden alfabético, con el fin de localizar las frecuencias que corresponden a una palabra, y otro en orden descendente de frecuencia-orden alfabético de palabra para apreciar la importancia de las palabras según su frecuencia, factor que —como lo dijimos ya— en un principio se consideró una forma adecuada —y hasta la única— de valorar la disponibilidad léxica.

Hablemos ahora de la construcción de la matriz de frecuencias parciales.

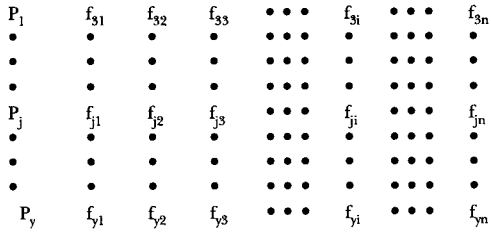
El proceso previo al cálculo de los índices de disponibilidad léxica consiste en formar una matriz 5 con los vectores de frecuencia parcial por posición de enunciamiento de cada palabra (f_{ji}), esto es, la matriz de frecuencias parciales tienen **"y"** renglones según el "diccionario" de la encuesta y **"n"** columnas según el máximo número de posición alcanzada.

Esta matriz 5 se llena con el proceso simultáneo de la matriz 4 ("diccionario" de la encuesta) y de la matriz 1 (datos de los informantes), siguiendo el algoritmo de rastrear por cada informante cada palabra dada desde la primera posición hasta el número de posición alcanzada por éste, entonces la palabra dada en la columna o posición **"i"** se localiza según el "diccionario" de la prueba, se obtiene el renglón **"j"** y se suma una ocurrencia a la f_{ji}, que es iniciada con valor cero.

La matriz 5, entonces, se ilustra como:

MATRIZ 5: MATRIZ DE FRECUENCIAS PARCIALES O MATRIZ DE VECTORES DE FRECUENCIAS

posición / *palabra*	*1*	*2*	*3*		*i*		*n*
P_1	f_{11}	f_{12}	f_{13}	• • •	f_{1i}	• • •	f_{1n}
P_1	f_{21}	f_{22}	f_{23}	• • •	f_{2i}	• • •	f_{2n}



donde f_{ji} = frecuencia de la j-ésima palabra en la i-ésima posición.

Aquí observamos que

$$\sum_{i=1}^{n} f_{ji} = FA_j$$

y ahora denotaremos

$$\sum_{j=1}^{y} f_{ji} = I_i$$

siendo **Ii** el número de informantes que pasaron por la posición "i"; y desde la captura se tiene cuidado de no saltar posiciones, por lo que $I_i \geq I_i + 1$, y además **Ii = X** donde **X** es el total de informantes porque son todos los que pasaron o llenaron la primera posición cuando menos.

Por lo que toca a la construcción de la matriz total de frecuencias parciales hemos de decir que hasta ahora hemos visto cómo se trabajan los datos sin tomar en cuenta la clave del informante, o sea que se trabaja con el universo, es decir la totalidad, de la prueba.

Sin embargo, hemos de ocuparnos también de la construcción de la matriz agrupada de frecuencias parciales.

Para formar grupos existe un programa selector de informantes que por medio de la fijación de una expresión lógica permite agruparlos según se desee: se puede, por ejemplo, obtener todos los informantes del sexo femenino, y que además estén en el turno vespertino, o que tengan una edad entre los 10 y los 15 años.

Con base en las instrucciones del programa —la mencionada expresión lógica—, entonces, se obtiene un subconjunto de la prueba capturada, al que se le considera cada vez como la matriz 1, con lo que el programa de computadora automáticamente sigue los pasos antes descritos.

Pero, retomando lo que ya hemos dicho, importa destacar que empleamos la fórmula propuesta en 1982 para las muestras limitadas en el número de respuestas, así como que lo utilizamos en varias pruebas y que tomamos la información obtenida sin cuestionamiento alguno; sin embargo, en cuanto hicimos uso del índice para ponderar las muestras limitadas en el tiempo de llenado detectamos notorias incongruencias —como lo señalamos más arriba—. Esto nos llevó —reiteramos— a analizar los resultados y a buscar una solución mejor —una solución congruente— para los trabajos que se estaban realizando y para los que se realizarían después; la facilidad de automatizar las bases para obtener el índice de disponibilidad léxica nos permitió hacer múltiples pruebas simplemente cambiando las fórmulas que se deseaban probar.

Abundamos en ciertas cuestiones para que se pueda entender con mayor claridad cómo diseñamos la ya aludida fórmula de López Chávez-Strassburger.

Así, repetimos que al programar la fórmula de 1982 apreciamos que era innecesariamente compleja, y que en realidad se podía reducir a

$$D(P_j) = \sum_{i=1}^{n} 0.9^{(i-1)} \bullet (f_{ji} / (N_{i-1})$$

considerando $N_0 = N_1$

Claramente, sin embargo, se notaba en los resultados que surgían problemas varios, por ejemplo en casos como el siguiente:

posición	1	2	3	4
•	•	•	•	•
•	•	•	•	•
palabra a	0	0	1	0
palabra b	0	0	0	1
•	•	•	•	•
•	•	•	•	•
Informantes	20	20	15	15

D(a) = 0.81 * 1/20 = 0.0405
D(b) = 0.739* 1/15 = 0.0486
o sea D(b) > D(a) !!! (ilógico e incongruente)

La dificultad se debe aquí a que se divide la frecuencia absoluta de la posición (i) entre el número de informantes de la posición anterior (N_{i-1}); y como el N_{i-1} es decreciente a medida que aumenta la posición, la "frecuencia relativa" aumenta y el coeficiente *lambda* no corrige tal situación.

Nos abocamos, pues, a buscar una mejor solución.

Dado que la fórmula ha de constar —necesariamente— de dos partes —la tasa de sustitución y la frecuencia relativa—, primero probamos con una tasa de sustitución exponencial cuadrática, además de que la "frecuencia relativa" la obtuvimos dividiendo la frecuencia absoluta de la posición entre la diferencia del doble de informantes en la primera posición y el número de informantes que pasaron por esa posición, todo ello con el fin de corregir los —así llamados en la jerga estadística— "caballos".

La primera fórmula que elaboramos fue, entonces, la que sigue:

$$D(P_j) = \sum_{i=1}^{n} e^{-(i/n)^2} \bullet (f_{ji}/(2N_1 - N_i))$$

Sin embargo, nos percatamos de que la tasa de sustitución dispersaba entre 0.99 y 0.40, o sea que daba, por un lado, valores poco dispersos pero, por otro lado, descubrimos que intervenía el rango de la prueba (**n**) o número máximo de posiciones, de tal forma que los valores primeros y último coincidían para cualquier **n**; esto nos llevó a examinar con más detenimiento la tasa de sustitución, de manera que observamos lo que viene a continuación:

a) Para la *lambda* se tiene que 10 ocurrencias en la posición **i** + 1 equivale a 9 ocurrencias en la posición **i** —según la tasa obtenida por los expertos mencionados en la propuesta de 1982—, lo que hace que se necesiten frecuencias cada vez mayores en el último lugar para igualar a 1 en el primero, en la medida que aumentan las posiciones máximas para diversas pruebas —cosa que podría ser cierta—. Veamos el cuadro siguiente:

Posición i	0.9**(i-1)	Recíproco de lambda
1	1.000000	1.00
5	0.656100	1.52
10	0.387420	2.58
20	0.135085	7.40
23	0.098477	10.15
30	0.047101	21.23
40	0.016423	60.89
50	0.005726	174.63
60	0.001997	500.85

El recíproco de *lambda* —o de la tasa de sustitución— indica el número de ocurrencias en la posición valorada para igualar una ocurrencia en la primera posición, y apreciamos que los valores de *lambda* para

i > 23 decrecen fuertemente, lo que provoca que se hagan despreciables los vocablos más allá de la posición 40 —siendo que en pruebas reales hay informantes que sobrepasan las 60 posiciones en tres minutos—.

b) Para la tasa de sustitución exponencial (**t**) que de hecho es

$$t = e^{-c.(i/n)^2}$$

se tiene que el número de ocurrencias en la última posición para un mismo coeficiente **C** es el mismo independientemente del tamaño de **n**.

Así, como no nos resultaba convincente la variación de *lambda*, ya que —por ejemplo— para **n** = 10 se necesitan 2.58 ocurrencias en el último lugar para equipararse a una en primera posición, y para **n** = 30 se necesitan 21.23 ocurrencias para lo mismo, entonces calculamos el coeficiente **"c"** de tal forma que siempre 10 ocurrencias en la última posición se equipararan a 1 en la primera; de este modo, **"c"** resultó ser **2.3**, con lo que se obtienen tasas de sustitución entre 1 y 0.1 a lo largo de todas las posiciones —cosa que pudiera no ser cierta—. Sin embargo, por la facilidad de la exponencial se puede cambiar **"c"** si esta dispersión no se considera adecuada. Ahora presentaremos el cuadro siguiente para **i/n** = **1** (en la última posición):

c	EXP(−c)	Recíproco
1.61	0.2	5
2.30	0.1	10
3.00	0.05	20
3.40	0.0333	30
3.70	0.025	40

En la misma forma que **"c"** se cambie se puede adaptar la base 0.9 de *lambda* a otros valores, por ejemplo 0.85 o 0.93, pero esto lo abandonamos por ser materia de

expertos y decidimos continuar la experimentación para obtener resultados lógicos y congruentes.

La segunda prueba que hicimos fue con la fórmula

$$D(P_j) = \sum_{i=1}^{n} e^{-2.3(i/n)^2} \bullet (f_{ji}/(2N_1 - N_i))$$

Como este cálculo arrojó resultados que se consideraron buenos en un principio, se realizaron con él las investigaciones de varias tesis de licenciatura. Pero salieron a la luz algunas incongruencias, como por ejemplo

posición	1	2	3
palabra a	0	2	0
palabra b	1	0	1

Es decir que esta fórmula plantea que dos posiciones en segundo lugar tienen mayor disponibilidad que una en primero y una en tercero, lo cual es falso, porque la tasa de sustitución debe ser creciente y lleva implícito que

$$t_i - t_{i+1} > t_{i+1} - t_{i+2} \text{ para } i = 1, 2, 3, \ldots, n$$

siendo t_i el valor de la tasa de sustitución en la posición **i**; entonces sumando t_{i+1} a cada miembro queda

$$t_i > 2t_{i+1} - t_{i+2}$$

$$\therefore \quad t_i + t_{i+2} > 2t_{i+1}$$

Y, en efecto, al revisar nuestra fórmula provisional vimos que la exponencial cuadrática —Campana de Gauss— tiene un punto de inflexión, y arriba de este punto originaba que $2t_{i+1}$ fuera mayor que $t_i - t_{i+2}$.

Esto lo corregimos cambiando la tasa de sustitución a una exponencial simple, quitando el exponente del cociente **i/n** para que la fórmula quedara de la siguiente forma:

$$D(Pj) = \sum_{i=1}^{n} e^{-2.3(i/n)} \bullet (f_{ji}/(2N_1 - N_i))$$

Finalmente ajustamos la dispersión entre **1** y **0.1**, poniendo el cociente **i/n** como **(i-1)/(n-1)**, y dejamos que la frecuencia relativa fuera en toda la amplitud de su concepto f_{ji}/I_1 —ya que no hay explicación para dividir entre N_{i-1} o entre $2N_1 - N_1$—, con el objeto de que coincidan los vectores de frecuencias absolutas con los de frecuencias relativas. Veamos la siguiente figura:

Posición/ Informante	1	2	3	4
1	p_2	p_4	p_7	p_1
2	p_1	p_3	p_6	•
3	p_3	p_5	p_4	•
4	p_4	p_3	•	•
5	p_3	•	•	•

Aquí se ilustran las siete palabras dadas por cinco informantes que llegaron hasta la cuarta posición, de modo que el vector de frecuencias absolutas es como sigue:

P_1	(1,0,0,1)	=	2
P_2	(1,0,0,0)	=	1
P_3	(2,2,0,0)	=	4
P_4	(1,1,1,0)	=	3
P_5	(0,1,0,0)	=	1
P_6	(0,0,1,0)	=	1
P_7	(0,0,1,0)	=	1
Suma ...	5,4,3,1	=	13

Nótese que P_2 y P_7 sólo fueron dadas una vez y que sus frecuencias absolutas se consideran ceros donde no aparecen.

Entonces, el vector de frecuencias relativas es el siguiente:

P1	(0.2,	0,	0,	0.2)	=	0.4
P2	(0.2,	0,	0,	0)	=	0.2
P3	(0.4,	0.4,	0,	0)	=	0.8
P4	(0.2,	0.2,	0.2,	0)	=	0.6
P5	(0,	0.2,	0,	0)	=	0.2
P6	(0,	0,	0.2,	0)	=	0.2
P7	(0,	0,	0.2,	0)	=	0.2
Suma ...	1.0,	0.8,	0.6,	0.2	=	2.6

Se puede apreciar aquí que 13/5 = 2.6 es el promedio matemático que resulta de dividir todas las ocurrencias entre los informantes.

Así, la versión final de la fórmula López Chávez-Strassburger es la siguiente:

$$D(P_j) = \sum_{i=1}^{n} e^{-c((i-1)/(n-1))} ((f_{ji}/I_1)$$

donde

- n = máxima posición alcanzada
- i = número de posición
- j = índice de la palabra
- f_{ji} = frecuencia absoluta de la palabra **j** en la posición **i**
- I_1 = número de informantes que participaron en la encuesta
- c = coeficiente de dispersión (se recomienda **2.3**)
- $D(P_j)$ = disponibilidad de la palabra **j**

Esta manera de calcular el índice de disponibilidad léxica se probó con datos reales de tesistas y no encontramos incongruencia alguna.

Además, hicimos pruebas en que deliberadamente tratábamos de verificar con puntos de control que (...,1,0,1...) fuera mayor que (...,0,2,0...), y que (...,5,3,4,...) fuera mayor que (...,4,5,3,...), porque

$$(5,3,4) = (4,3,3)+(1,0,1)$$
$$(4,5,3) = (4,3,3)+(0,2,0)$$

por lo tanto (5,3,4) > (4,5,3). Utilizando, pues, esta matemática, se pusieron varios ejemplos que funcionaron como debe ser: en las pruebas para diferentes coeficientes de dispersión el comportamiento de la disponibilidad es casi el mismo, y se puede decir con certeza que no se afectan los resultados finales.

El caso es que tuvimos la suerte de dar con una fórmula de disponibilidad léxica que reflejaba con fidelidad prístina la realidad lingüística —que es justamente lo que nos interesaba conseguir—, una fórmula que supo salir airosa de todo tipo de pruebas.

Importa, sin embargo, aclarar que si bien este cálculo es —afortunadamente— muy confiable, se refiere al índice de disponibilidad del vocablo.

Es así que nos pusimos a trabajar en la búsqueda de una igualmente eficaz fórmula para el cálculo del índice de disponibilidad individual —es decir, el que se refiere a cada uno de los informantes—.

En una primera instancia, diseñamos la fórmula siguiente:

$$D(S_w) = 1000 \sum_{c_i=1}^{m} \sum_{i=1}^{n_{ci}} e^{-2.3((i-1)/(nci-1))} \bullet D(p_{ci},w,i)$$

donde:

D(Sw)	=	disponibilidad léxica del sujeto **w**
$\mathbf{c_i}$	=	centro de interés
m	=	número de centros de interés de la prueba
$\mathbf{n_{ci}}$	=	máximo número de posiciones para el centro de interés en cuestión

$e^{-2.3((i-1)/(nci-1))}$	= tasa de dispersión para ponderar según la posición en que fue dicha cada palabra
P_{ci}	= palabra del centro de interés en cuestión
$D(P_{ci},w,i)$	= índice de disponibilidad léxica de la palabra **P** del centro de interés en cuestión, mencionada por el sujeto **w** en la posición **i**
1000	= constante para obtener cifras significativas.

No abundaremos aquí sobre este punto, puesto que en 1991 publicamos un artículo en que se trata ampliamente el tema[11].

Hemos de decir únicamente que este otro cálculo también es eficaz y congruente, además de que no contradice en absoluto las formulaciones de base —lo que tiene una importancia incontrovertible—.

Volviendo, entonces, al índice de disponibilidad del vocablo, queremos finalizar destacando el hecho de que llegar a una sola fórmula estadística para tratar dos tipos muy diferentes de muestras constituye una simplificación muy considerable —sobre todo en comparación con la planteada en 1982— en la aplicación y el tratamiento de los datos, lo que, aunado a que la mencionada fórmula tiene un —reiteradamente probado— alto grado de confiabilidad, nos concede el derecho de sentirnos muy satisfechos de la aportación que el trabajo realizado significa para las investigaciones del léxico disponible de la lengua española.

JUAN LÓPEZ CHÁVEZ
Facultad de Filosofía y Letras.

CARLOS STRASSBURGER FRÍAS
Facultad de Ingeniería.

[11] "Un modelo para el cálculo del índice de disponibilidad léxica individual", *La enseñanza del español como lengua materna. Actas del II Seminario internacional sobre aportes de la lingüística a la enseñanza del español como lengua materna*, Río Piedras, Universidad de Puerto Rico, 1991, pp. 91-112.