

# Estudio estadístico de la ortografía castellana (2): Frecuencia de bigramas

CARLOS J. ALVAREZ, MANUEL CARREIRAS Y MANUEL DE VEGA  
*Universidad de La Laguna*



## Resumen

*A partir de la misma muestra de más de 25.000 palabras castellanas utilizada en el estudio silábico de Alvarez, Carreiras y De Vega (en artículo anterior), se realizaron tres tabulaciones de frecuencia posicional y de frecuencia total de bigramas (pares de letras): una para palabras de clase abierta, otra para palabras de clase cerrada y otra para la muestra total. El resultado son tres listados de bigramas con la frecuencia absoluta y ponderada correspondiente a cada bigrama en cada posición de palabra y a la frecuencia total de cada bigrama en la muestra. La frecuencia ponderada posicional es el resultado de dividir cada frecuencia posicional por el número de palabras que puede contener ese bigrama en esa posición.*

*Palabras clave:* Bigrama, frecuencia de bigramas, acceso léxico, frecuencia silábica.

## Abstract

*It were built on three dictionaries of total and by position in the words bigram frequency from a pool of 25.000 usual spanish words: one for closed class words, one for open class words and the other one for the whole sample. The result is three lists with each posible bigram and the times which it appears in each word position and further the total frequency of this bigram. There is a weighted frequency, and it is the result of dividing the positional frequency by the number of words which can contain that bigram in that position.*

*Agradecimientos:* Esta investigación fue subvencionada por el proyecto de la DGICYT Pb 88-0425, así como por la beca FPI (Subprograma General) del Ministerio de Educación y Ciencia concedida al primer autor. A su vez formó parte de los créditos de investigación del programa de Tercer Ciclo sobre "Procesamiento del Lenguaje" (Universidad de La Laguna). Los autores quieren agradecer especialmente la colaboración de Fernando Caballero de Rodas, quien realizó el programa informático utilizado en este trabajo.

*Dirección del autor:* Departamento de Psicología Cognitiva. Facultad de Psicología. Universidad de La Laguna. 38201 Tenerife.

## INTRODUCCION

Existen propiedades de la estructura ortográfica de un idioma que pueden influir en el reconocimiento de letras o de palabras, permitiendo un acceso más rápido al significado de una unidad informativa dada. Por ejemplo, en experimentos con reconocimiento de letras reconocemos mejor letras individuales cuando éstas pertenecen a una palabra que cuando no lo hacen (Reicher, 1969). Se ha distinguido entre dos grandes categorías de la estructura ortográfica: la regularidad gobernada por reglas y la redundancia estadística (Massaro, Taylor, Venezky y Jastrzembski, 1980). El término regularidad gobernada por reglas se refiere a aquellas propiedades de las letras o de las secuencias de letras fruto de las limitaciones o convenciones que un idioma impone a sus palabras o a su ortografía y que el lector generaliza (por ejemplo, el patrón de letras «nb» es imposible en castellado y es una norma que el lector acaba generalizando). Por su parte, la redundancia estadística se aplica a aquellas medidas que pueden derivarse de la frecuencia con las cuales las letras, las secuencias de letras o las palabras aparecen en los textos naturales. Dicha propiedad o característica del lenguaje es de tipo distributivo, en el sentido de que una unidad lingüística dada, pongamos por caso una letra determinada, puede aparecer con mayor o menor frecuencia que otras unidades del mismo tipo. La cuestión es si esta característica distributiva de las unidades ortográficas afecta de algún modo a su procesamiento.

Algunos autores han mantenido la postura de que los buenos lectores extraen información sobre las ocurrencias legales de letras o de secuencias de letras a partir de las exposiciones repetidas a palabras escritas, y que emplean esta información para facilitar el reconocimiento de palabras en tareas normales de lectura (Massaro et al., 1980; Seidenberg, 1989). Nuestra experiencia lectora nos va proveyendo de un conocimiento implícito acerca de las regularidades del idioma y de la distribución de frecuencias de ciertos patrones lingüísticos que posteriormente emplearemos para un mayor rendimiento lector. Esta es la razón por la cual desde muy pronto los psicólogos se han interesado en la medición de diversos tipos de redundancia estadística.

Es necesario hacer una diferenciación entre dos tipos de recuentos de frecuencia utilizados en estos estudios: el recuento tipo «token» y el recuento tipo «type» (Massaro et al., 1980). El primero consiste en contabilizar las letras o secuencias de letras en todas las palabras de una muestra aun cuando algunas de estas palabras se repitan varias veces en la muestra. En el tipo «type» se cuenta también la ocurrencia de una letra o secuencia de letras en las palabras que la contienen, pero sin tener en cuenta la frecuencia de la palabra en la muestra. Por ejemplo, si la palabra «libro» aparece diez veces, sólo se contaría la ocurrencia de «li» en esa palabra como una ocurrencia.

A nivel subléxico uno de los primeros análisis de este tipo fue realizado por Solso y King (1976), quienes tabularon la frecuencia de letras individuales (cuántas veces aparece una letra en un millón de palabras) y su versatilidad (el número de veces que una letra aparece en una posición específica y el número de palabras diferentes que contienen esa letra). Anteriormente ya se habían tabulado los bigramas y trigramas de muestras pequeñas de palabras (2.090) utilizando recuentos tipo «token», con el fin de obtener un índice de familiaridad, sumando las frecuencias de los bigramas y trigramas de una palabra (Underwood y Schultz, 1960). Sin embargo, en estos estudios no se tuvo en cuenta

una variable de gran importancia: la longitud de las palabras, variable que explicaba gran parte de la varianza en estudios experimentales que utilizaban la frecuencia de bigramas como variable independiente (Spreeen y Schultz, 1960). Con el fin de solventar este problema se realizaron otro tipo de recuentos de frecuencia de bigramas, trigramas e incluso tetragramas, que tuvieron en cuenta también la posición dentro de la palabra (lo que llamaremos Frecuencia Posicional) y la longitud de la palabra (Mayzner y Tresselt, 1965; Massaro et al., 1980). En todos los trabajos mencionados se utilizó recuento tipo «token». No obstante, sobre las listas de palabras de Kucera y Francis (1967) también se han construido recuentos de frecuencia de bigramas tipo «type», teniendo en cuenta tanto la longitud de la palabra como la posición dentro de ésta (Solso, Barbutto y Juel, 1979; Solso y Juel, 1980).

Trabajos de esta índole también se han dado en otros idiomas aunque en menor medida que en el inglés. Por ejemplo, Content y Radeau (1988) realizaron una tabulación de frecuencia posicional de letras, bigramas y trigramas en 30.000 palabras francesas.

En castellano no abundan precisamente este tipo de estudios y por ello el presente trabajo, junto al estudio estadístico de frecuencia silábica (Alvarez, Carreiras y De Vega, en prensa), y en su misma línea, pretende suplir de alguna manera esta carencia, permitiendo la manipulación de este tipo de variables en la investigación psicolingüística.

## METODO

El primer paso fue obtener una muestra representativa de palabras castellanas de uso relativamente común. Para ello se muestrearon diversos tipos de fuentes impresas. Concretamente, periódicos nacionales y locales, revistas de divulgación científica general y ensayos de autores nacionales contemporáneos así como novelas actuales. Dicha muestra fue prácticamente idéntica, con algunas variaciones, a la utilizada en el estudio de frecuencia silábica (Alvarez, Carreiras y De Vega, en prensa), donde figuran con más detalle sus características técnicas. Se extrajo aproximadamente el mismo número de párrafos de cada uno de los tipos de publicaciones, párrafos de entre 25 y 70 palabras extraídos al azar. No se tuvieron en cuenta las palabras técnicas, cultas ni pertenecientes a otras lenguas. El número total de palabras fue de aproximadamente 25.000.

La muestra total de palabras fue introducida en un ordenador compatible IBM. Posteriormente se crearon dos ficheros diferentes y se dividió la muestra en dos tipos de palabras: palabras de clase abierta y de clase cerrada. Esta distinción, aunque discutida, se realizó debido al hecho de que son muchas las investigaciones que destacan la importancia de la diferenciación lingüística entre estos dos tipos de vocabulario, que encuentra una correspondencia a nivel psicológico, y que parecen desempeñar funciones distintas tanto en la comprensión como en la producción y adquisición del lenguaje (Sánchez-Casas y García-Albea, 1986).

Se realizó un programa informático para la tabulación mecánica de los bigramas. Este algoritmo que utilizamos para aplicar a cada uno de los tres ficheros (clase abierta, clase cerrada y muestra total) produjo tres listados, uno para cada fichero. En estos listados aparecen los distintos bigramas (pares de letras)

existentes en la muestra por orden alfabético y a continuación las frecuencias de cada uno según su posición en la palabra. La primera columna (PP) se refiere a los bigramas en posición de principio de palabra. La segunda (1L) son aquellos bigramas con una letra antes («le» en «aleja», por ejemplo). La tercera (1B) son los bigramas con un bigrama antes («le» en «paleta»). La cuarta, bigramas con dos bigramas antes («co» en «discos») y así sucesivamente hasta ocho bigramas antes o más (8Bo+). En la columna 8Bo+ se incluyen aquellos bigramas a los que les preceden ocho o más, a excepción de los bigramas que se encuentran en posición final de palabra. Estos son listados en una columna distinta (FP), independientemente de si están en una palabra corta o larga (el bigrama «es» será situado en Posición Final tanto en «pies» como en «rotuladores»). La última columna es la frecuencia total del bigrama en la muestra (TOTAL). Al final de cada una de las tres tablas se encuentra un listado similar pero más corto correspondiente a las frecuencias de las palabras monosílabas de cada tipo de palabra (clase abierta, clase cerrada y total). Esta diferenciación entre monosílabos y polisílabos se ha realizado con el fin de hacer comparable el presente trabajo con el Diccionario de Frecuencia Silábica en palabras castellanas (Alvarez, Carreiras y De Vega, en prensa). Al final de cada listado se ofrece un cómputo del número de bigramas contados, así como del número de palabras según el número de letras.

Para cada posición de bigramas se tabularon dos índices distintos: la Frecuencia Absoluta (FA) y la Frecuencia Ponderada (FP). La Frecuencia Absoluta es el número de veces que aparece ese bigrama en esa posición en toda la muestra (sea la total, la de palabras de clase abierta o la de palabras de clase cerrada), mientras que la Frecuencia Ponderada es un índice corrector que utilizamos con el fin de tener en cuenta la longitud de las palabras. Es el resultado de dividir la Frecuencia Absoluta por el número de palabras que realmente puede contener ese bigrama en esa posición y multiplicarlo por cien. Por ejemplo, un bigrama precedido por tres bigramas jamás podrá encontrarse en palabras de menos de seis letras, por lo cual la frecuencia de ese bigrama en esa posición será dividida por el número total de palabras de siete o más letras y multiplicado por cien. El motivo de dividirlo por siete y no por seis es el siguiente: si un bigrama está en posición 3B (con tres bigramas antes), quiere decir que tiene cuatro letras antes, más las dos del bigrama y al menos está seguido por una letra, ya que si ese bigrama se encontrara en posición final, sería tabulado como FP (final de palabra). Por tanto, las palabras que lo pueden contener son aquellas de siete o más letras. La misma lógica se sigue para cualquier bigrama en cualquier posición. Los bigramas en principio de palabra, en final de palabra y la frecuencia total están divididos por el total de palabras de su muestra.

El sistema de tabulación utilizado fue de tipo «token», contándose la ocurrencia de un bigrama en una posición dada en todas las palabras que la contengan, esté o no esté esa palabra repetida en la muestra. Sin embargo, se trató de eliminar el sesgo local de repetición, peligro inherente a este tipo de tabulación y que consiste en la repetición excesiva de una palabra en un texto debido a ser un tópico central a dicho texto, aunque esa palabra sea infrecuente en el idioma. Para ello, y como se mencionó anteriormente en lo relativo a la muestra, se extrajeron párrafos pequeños.

El Anexo 1 es el listado de las frecuencias de bigramas en las palabras de clase abierta, el Anexo 2 es el de las palabras de clase cerrada y el Anexo 3 es el de las frecuencias de la muestra total.

## ESTUDIO CORRELACIONAL

## Método

La hipótesis de la Redundancia Ortográfica (Seidenberg, 1989) propone que las sílabas están compuestas por bigramas de mayor frecuencia que los límites entre sílabas. Esta diferencia podría explicar los efectos atribuidos a las sílabas encontrados en la literatura experimental sobre reconocimiento visual de palabras. Con el fin de explorar esta hipótesis en castellano elegimos 50 palabras al azar, 25 bisílabas y 25 trisílabas. Tabulamos la media de los bigramas intrasílaba, la media de los bigramas intersílaba (límites entre sílabas, por ejemplo, «as» en «casa») y la media de la frecuencia silábica, esta última según el trabajo de Alvarez, Carreiras y De Vega, en prensa) para cada palabra por separado. A continuación se llevaron a cabo correlaciones entre las tres variables. Los resultados pueden verse en la tabla 1.

TABLA 1

	FRE. SIL.	F. INTRA.
F. INTRA.	.0698	
F. INTER.	-.1580	.1493

*Correlaciones entre la frecuencia silábica, la frecuencia de bigramas intersílaba y la frecuencia de bigramas intrasílaba para una muestra de 50 palabras.*

De acuerdo con la hipótesis de la redundancia ortográfica, cabría esperar obtener una correlación positiva y alta entre la frecuencia intra y la frecuencia silábica. Como puede apreciarse la correlación obtenida es despreciable y no significativa. Lo mismo ocurre con la frecuencia silábica y la frecuencia de bigramas inter, aunque este resultado sí era esperable.

Las palabras utilizadas en el estudio fueron elegidas al azar y podría haber ocurrido que todas o gran parte de ellas pertenecieran a un mismo rango de frecuencia (por ejemplo, que todas fuesen de alta frecuencia silábica). Por ello, en una segunda fase, ampliamos la muestra a 288 palabras, siendo la mitad de ellas bisílabas y la otra mitad trisílabas. Por otro lado, la mitad fueron palabras de alta frecuencia silábica (con puntuaciones mayores de 100) y la otra mitad lo fueron de baja (menores de 50). Se controló también la frecuencia léxica (en cada celdilla la mitad eran de alta y la mitad de baja). Como puede verse en la tabla 2, los resultados predichos por Seidenberg sólo se obtienen en las palabras de alta frecuencia silábica, mientras que en aquellas palabras que tenían una frecuencia silábica baja no se obtuvieron correlaciones dignas de mención. Este resultado es similar al obtenido por Gernsbacher (1984) con familiaridad y frecuencia léxica, donde las palabras de baja frecuencia y alta familiaridad subjetiva eran reconocidas más rápidamente que las palabras de idéntica frecuencia pero baja familiaridad, siendo dos variables muy relacionadas.

TABLA 2

ALTA. FRE. SIL.			BAJA. FRE. SIL.		
	FRE. SIL.	F. INTRA.		FRE. SIL.	F. INTRA.
F. INTRA.	.3681**	—	F. INTRA.	.0384	—
F. INTER.	— .1546	.1129	F. INTER.	— .0415	.0067

\*\* *Correlaciones significativas* ( $-0.0001$ )

*Correlaciones entre la frecuencia silábica, la frecuencia de bigramas intersílaba y la frecuencia de bigramas intrasílaba para palabras de alta frecuencia silábica.*

### Discusión

Las correlaciones obtenidas en el presente estudio no parecen sostener la hipótesis de Seidenberg, según la cual las sílabas estarían compuestas de bigramas de alta frecuencia mientras que las uniones entre sílabas lo estarían por bigramas de baja frecuencia. Sólo se obtuvo dicho patrón en las palabras compuestas por sílabas muy frecuentes. Parece que, a pesar de todo, son dos variables diferentes y que no podemos reducir la frecuencia silábica a la frecuencia de bigramas.

### DISCUSION

Los resultados en inglés acerca de la influencia de la variable frecuencia de bigramas en el procesamiento son variados. Algunos autores han obtenido correlaciones significativas entre la frecuencia posicional de letras individuales y la exactitud en el informe de ítems de cuatro letras en mayor medida que cuando ésta es correlacionada con la frecuencia de bigramas (McClelland y Johnston, 1977). En tareas de decisión léxica y de reconocimiento también se han obtenido efectos significativos de la frecuencia posicional de letras individuales (Bouwhuis, 1979; Massaro, Venezky y Taylor, 1979). Sin embargo, las correlaciones disminuían considerablemente al incluir la frecuencia posicional de bigramas (Massaro et al., 1980). Esta variable fue utilizada en experimentos perceptivos, encontrándose que las palabras de baja frecuencia son más fácilmente percibidas cuando están formadas por bigramas de baja frecuencia (Broadbent y Gregory, 1968). Para Solso y Juel (1980) la frecuencia posicional de bigramas es más adecuada para estimar regularidades ortográficas que la frecuencia de letras solas, ya que refleja múltiples regularidades de conexiones entre letras que no puede reflejar la frecuencia de letras solas.

En los actuales modelos conexionistas se da una gran importancia a la coocurrencia de letras, concretamente a la frecuencia de los bigramas (Seidenberg, 1989). Para estos modelos conexionistas no es necesario postular ninguna unidad subléxica de procesamiento visual de palabras, a no ser la letra y se basan en la redundancia ortográfica (frecuencia de bigramas) para explicar algunos efectos atribuidos previamente a ciertas unidades subléxicas como la sílaba, el morfema, etc. Nuestro estudio correlacional, y a pesar de su carácter exploratorio, no parece apuntar en esta dirección, aunque evidentemente es una cuestión que

requiere una mayor profundidad y una constatación experimental. Otros autores, sin embargo, defienden la existencia de una unidad subléxica de procesamiento visual (Taft, 1989; Pritzmetal, Treiman y Rho, 1986; De Vega et al., 1990), destacando que las diferencias idiomáticas pueden determinar diferentes tipos de segmentación subléxica, tal y como afirman Cutler, Mehler, Norris y Segui (1986) en estudios con material auditivo. Por ello resulta sumamente importante disponer de este tipo de estudios en castellano, que posibilitan el diseño de investigaciones tendentes a contrastar las similitudes y diferencias del procesamiento subléxico en castellano y en otros idiomas.

La importancia de este tipo de estudios reside en su valor instrumental para las investigaciones psicolingüísticas. El poder disponer en castellano de una medida de frecuencia de bigramas, por un lado, y de una medida de frecuencia silábica (Alvarez, Carreiras y De Vega, en prensa), por otro, permite poner a prueba hipótesis acerca del rol final de ambos tipos de unidades subléxicas en el procesamiento de palabras. Entre otras cuestiones ayudará a conocer en qué medida los resultados obtenidos con la lengua inglesa son universales o difieren según la regularidad ortográfica particular de cada idioma.

## ANEXOS

### *Claves de las abreviaturas*

- BIG: Bigrama.  
 PP: Posición de bigrama de principio de palabra.  
 1L: Posición de bigrama precedido por una letra.  
 1B, 2B, 3B...: Posición de bigrama precedido por un, dos, tres, etc., bigramas.  
 8BO + : Posición de bigrama precedido por ocho o más bigramas.  
 FP: Posición de bigrama en final de palabra.  
 TOTAL: Frecuencia total del bigrama en la muestra.  
 FA: Frecuencia Absoluta.  
 FP: Frecuencia Ponderada.

## Referencias

- ALVAREZ, C.; CARREIRAS, M., y DE VEGA, M.: Estudio estadístico de la ortografía castellana (I): la frecuencia silábica (1992): *Cognitiva 4 (1)*, 75-105. Madrid. Aprendizaje.  
 BOWHUIS, D. G. (1979): *Visual Recognition of Words*. Tesis Doctoral, Katholieke Universiteit, Nijmegen, Holanda.  
 BROADBENT, D. E., y GREGORY, M. (1968): Visual perception of words differing in letter digram frequency. *Journal of Verbal Learning and Verbal Behavior*, 7, 569-571.  
 CONTENT, A., y RADEAU, M. (1988): Données statistiques sur la structure orthographique du Français. *European Bulletin of Cognitive Psychology*, 8, (4), 339-404.  
 CUTLER, A.; MEHLER, J.; NORRIS, D., y SEGUI, J. (1986): The Syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385-400.  
 DE VEGA, M.; CARREIRAS, M.; GUTIERREZ, M., y ALONSO, M. L. (1990): *Lectura y comprensión: una perspectiva cognitiva*. Madrid. Alianza Editorial.

- GERNSBACHER, M. A. (1984): Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, 2, 256-363.
- KUCERA, H., y FRANCIS, W. (1967): Computational analysis of present day american English. Providence, R.I.: Brown University Press.
- MCCLELLAND, J. L., JHONSTON, J. C. (1977): The role of familiar units in perception of words and non words. *Perception and Psychophysics*, 22, 249-261.
- MASSARO, D.; TAYLOR, G.; VENEZKY, R., y JASTRZEMBSKI, J. (1980): *Letter and world perception*. Amsterdam, North Holland Publishing Company.
- MAYZNER, M. S., y TRESSELT, M. E. (1965): Tables of single letters and digram frequency counts for various world length and letter position combination. *Psychonomic Monograph Supplements*, 1, 13-32.
- SPREEN, O., y SCHULZ, R. W. (1966): Parameters of abstraction, meaningfulness and pronounceability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5, 459-468.
- PRITZMETAL, W., T;REIMAN, R., y RHO, S. H. (1986): How to see a reading unit. *Journal of Memory and Language*, 25, 461-475.
- REICHER, G. M. (1969): Perceptual recognition as a function of the meaningfulness of the stimulus material. *Journal of Experimental Psychology*, 81, 275-280.
- SÁNCHEZ-CASAS, R. M., y GARCÍA-ALBEA, J. E. (1986): Dos vocabularios: diferencias computacionales en el estudio del lenguaje. En M. Siguán (ed.), *Estudios de Psicolingüística*, 87-103, Ed. Pirámide, Madrid.
- SEIDENBERG, M. (1989): Reading complex word. En G. Carlson y M. Tanenhaus (eds.), *Linguistic Structure in Language Processing*, 53-105. Kluwer Academic Publishers.
- SOLSO, R., y KING, J.: Frequency and versatility of letters in the English language. *Behavior Research Methods and Instrumentation*, 8, 283-286.
- SOLSO, R.; BARBUTO, P., y JUEL, C. (1979): Bigram and trigram frequencies and versatilities in the English language. *Behavior Research Methods and Instrumentation*, 11, 474-484.
- SOLSO, R., y JUEL, C. (1980): Positional frequency and versatility of bigrams for two through nine letter English words. *Behavior Research Methods and Computer*, 12, (3), 297-343.
- TAFT, M. (1989): Morphographic processing: the BOSS re-emerges. En M. Coltheart (ed.), *Attention and Performance, XII: reading*. Hillsdale, NJ: Erlbaum.
- UNDERWOOD, B. J., y SCHULTZ, R. W. (1960): Meaningfulness and verbal learning. Nueva York: Academic Press.





Table with columns: DCR, FA, PP, IL, VA, IB, VA, ZB, SB, AB, SB, GB, VA, 7B, GP, VA, PP, TOTAL, PP. Rows include letters 'a' through 'z' and numbers 1-19. Data is organized in columns of pairs of values.











Frecuencia posicional absoluta y ponderada y frecuencia total absoluta y ponderada de bigramas de palabras polisilabos en la muestra total

Table with columns: BGR, PP, 1L, 1B, 2B, 3B, 4B, 5B, 6B, 7B, 8B+, 9B, TOTAL. Rows list bigrams (e.g., ab, ac, ad) and their respective frequencies in various positions.







