# INVENTARIO DE FRECUENCIAS FONÉMICAS Y SILÁBICAS DEL CASTELLANO ESPONTÁNEO Y ESCRITO

*Antonio Moreno Sandoval[1], Doroteo Torre Toledano[1,2] Natalia Curto[1] y Raúl de la Torre[1]*

{antonio.msandoval,doroteo.torre}@uam.es {natalia,raul}@maria.lllf.uam.es

[1]LLI, [2]ATVS, Universidad Autónoma de Madrid, SPAIN.

**RESUMEN**

Este artículo presenta dos inventarios de frecuencias – fonémico y silábico – del castellano obtenidos a partir del corpus C-ORAL-ROM, que recoge español oral espontáneo en distintos contextos y registros. Estos inventarios se han desarrollado mediante un transcriptor fonológico y silábico cuyos resultados para el corpus C-ORAL-ROM han sido en su mayor parte revisados manualmente. Los inventarios incluyen la frecuencia absoluta de aparición de los diferentes fonemas y sílabas. Estos datos se han examinado junto a los extraídos de un corpus comparable de texto escrito y se ha hallado evidencia de que los inventarios de frecuencias obtenidos hasta ahora, basados fundamentalmente en textos, no describen adecuadamente el castellano oral espontáneo.

## 1. INTRODUCCIÓN

El primer inventario de frecuencias de fonemas para el castellano fue el elaborado por Zipf y Rogers [7] en 1939, sirviéndose de la descripción fonológica de Navarro Tomás. Posteriormente se han realizado varios estudios sobre esta cuestión, que aparecen recogidos en la Tabla 1. Dicha tabla indica el número total de fonemas/letras considerados para el cálculo de frecuencias así como el tipo de corpus empleado (oral o escrito).

| Autores | nº fon/let. | Tipo |
|---|---|---|
| Zipf y Rogers (1939) | 5.000 | escrito |
| Navarro Tomás (1946) [2] | 20.000 | Escrito |
| Guirao y Borzone (1972) [3] | 62.980 | Escrito |
| Quilis y Esgueva (1980) [4] | 160.000 | Oral |
| Rojo (1991) [5] | 3.641.915 | Escrito |
| UAM (2006) | 1.200.000 | Oral |

**Tabla 1.** *Estudios de frecuencias de los fonemas del castellano.*

Este trabajo ha sido finaciado parcialmente por el MEC-CICYT (TIN2004-07588-C03-02) y la Comunidad Autónoma de Madrid y la Universidad Autónoma de Madrid (05/TIC/001).

Los estudios relativos a las sílabas son aún más escasos: Álvarez, Carreiras y De Vega [7] y Alameda y Cuetos [1] son los más recientes. Los primeros autores consideraron 41.592 sílabas. Los últimos emplearon 3.930.954 en sus cálculos. En ambos casos, las fuentes utilizadas fueron textos escritos sin transcribir fonológicamente.

En el presente estudio se ha manejado un total de 1.244.411 fonemas y 558.982 sílabas.

La novedad de esta investigación viene dada por el empleo como fuente de un corpus de habla espontánea, diez veces mayor del utilizado por Quilis y Esgueva [4] y con un gran número de hablantes. El LLI-UAM ha compilado los dos corpus de español oral más importantes: CORLEC[9] y C-ORAL-ROM [10]. Este último ha constituido la base para nuestro estudio. Este corpus contiene más de 348.000 palabras (incluyendo algunas marcas prosódicas), de 192 textos transcritos. En total, 429 hablantes distintos y más de 42 horas de grabación. El corpus se halla dividido en tres grandes grupos: informal (165.210 palabras), formal (70.924) y medios de comunicación (97.170). Un pequeño subcorpus de 14,760 palabras recoge conversaciones telefónicas. La calidad de las transcripciones viene avalada por una validación interna (cada texto fue verificado por al menos 3 lingüistas) y externa (ELDA). La alineación de sonido y transcripción supone la garantía última de que la transcripción se ajusta a la grabación.

La sección 2 describe la metodología general seguida para la obtención del inventario y la sección 3 está dedicada a describir el transcriptor que se ha empleado. La sección 4 recoge los resultados más importantes e incluye una comparación de los corpus escrito y oral. La sección 5 presenta las conclusiones que hemos obtenido de los resultados.

## 2. METODOLOGÍA

El corpus C-ORAL-ROM incluye la transcripción ortográfica de todas las grabaciones pero no la fonémica ni la silábica. Éstas han sido obtenidas de manera semi-automática.

Para la compilación del inventario del corpus oral se han dado los siguientes pasos:

1. El punto de partida fue un transcriptor desarrollado previamente (sección 3).
2. Basándose en el corpus CORLEC, los investigadores del LLI identificaron posibles problemas del transcriptor.
3. A partir de los resultados de esta investigación se realizaron mejoras en las reglas del transcriptor.
4. Se obtuvo una transcripción fonémica y silábica preliminar de C-ORAL-ROM.
5. Las transcripciones se revisaron manualmente: todas ellas fueron revisadas por al menos un lingüista y el 60% de ellas por dos personas distintas. Con estas revisiones, se volvieron a mejorar las reglas y las excepciones del transcriptor.
6. Finalmente, se llevó a cabo una transcripción definitiva que ha sido la base para los inventarios que aquí presentamos.

## 3. DESARROLLO DEL TRANSCRIPTOR

El transcriptor empleado en este trabajo utiliza un mecanismo de reglas de reescritura dependientes del contexto y excepciones, con la palabra como unidad. En primer lugar, la palabra se busca en una lista de excepciones que contiene transcripciones fonológicas y silábicas para casos especiales (normalmente extranjerismos). Si la palabra se halla en la lista se toma su correspondiente transcripción. Si la palabra no figura en la lista se aplica una serie de reglas de reescritura dependientes del contexto para obtener la transcripción fonológica. Después se derivan las sílabas a partir de las vocales mediante otro conjunto de reglas. Por último, otro conjunto de reglas interviene para determinar si las sílabas son acentuadas o no dependiendo de las reglas acentuales y ortográficas que rigen el empleo de la tilde. La transformación de la representación ortográfica de una palabra a su transcripción fonológica se basa en reglas de reescritura dependientes del contexto con el siguiente formato:

signo → [context-i] nuevo(s)-signo(s) [context-d]

en el que [context-i] y [context-d] son opcionales y pueden incluir cualquier número de signos (que pueden representar letras o fonemas dependiendo de la regla). Signo es la letra o fonema que debe reescribirse y nuevo(s) -signo(s) puede ser cero, uno o más signos que representan letras o fonemas (según la regla). Estas reglas se aplican una por una en un orden predeterminado. Dada la regularidad de la correspondencia entre letras y sonidos en español, bastó con 50 reglas para obtener buenos resultados. A esto también contribuyó nuestra decisión de utilizar un conjunto mínimo de 23 fonemas y de tomar en cuenta únicamente las transcripciones fonológicas canónicas, obviando variantes regionales y reducciones. Una limitación importante de esta transformación es el hecho de que la unidad de transcripción empleada es la palabra y no oraciones enteras con lo que fenómenos fonológicos intraoracionales no pueden tratarse.

Una vez que la palabra se ha transformado en secuencia de fonemas, cada vocal se marca inicialmente como una sílaba. A continuación, unas reglas determinan qué pares de vocales pertenecen a la misma sílaba y cuáles a sílabas diferentes. A continuación, ocho reglas añaden a la sílaba que corresponda las consonantes que aparecen antes y después de la(s) vocal(es). Si al final de este proceso hay consonantes no asignadas a ninguna sílaba, se informa de un error de silabificación, lo que sucedió fundamentalmente con palabras extranjeras y con acrónimos, que se incluyeron en la lista de excepciones.

Cuando la palabra ha sido transcrita fonológicamente y silabificada, otro conjunto de reglas asigna acento a una de las sílabas de la palabra según las convenciones ortográficas del español. En este punto la limitación de tomar la palabra como unidad es de nuevo importante, dado que es frecuente en español oral el agrupar varias palabras (p.e. el nombre y palabras funcionales como el artículo) con una única sílaba acentuada, lo que nuestro transcriptor no puede reflejar.

Para desarrollar el inventario hemos afinado las reglas y las excepciones del transcriptor, basándonos en las correcciones manuales de las transcripciones automáticas del corpus C-ORAL-ROM. Debe señalarse que sólo un 2% de las palabras transcritas automáticamente contenían un error de transcripción fonémico o silábico. Para el presente trabajo hemos omitido la información acentual y hemos tenido en cuenta únicamente la transcripción fonológica y la silabificación. Por lo tanto dos sílabas o fonemas que difieren únicamente en el acento se consideran equivalentes. Las excepciones corresponden sobre todo a palabras extranjeras y acrónimos con los cuales el transcriptor produjo errores de silabificación.

## 4. RESULTADOS

Con el fin de facilitar una comparación entre las frecuencias obtenidas a partir del corpus escrito y las del oral, seleccionamos aleatoriamente 500.000 palabras del corpus de una agencia de noticias (EFE) de 150 millones de palabras. El procedimiento de selección tomaba una palabra de cada 300. De esta forma contamos con un corpus escrito representativo con el que comparar nuestro corpus oral.

| Fonemas | Español oral | | | | Español escrito | | | |
|---|---|---|---|---|---|---|---|---|
| | Lexicón | | Corpus | | Lexicón | | Corpus | |
| | Fr. Absoluta | Fr. Relat. | Fr. Absoluta | Fr. Relat. | Fr. Absoluta | Fr. Relat. | Fr. Absoluta | Fr. Relat. |
| a | 23294 | 13.87 | 152664 | 12.27 | 46488 | 13.48 | 323783 | 12.89 |
| b | 5036 | 2.99 | 31126 | 2,50 | 12513 | 3.63 | 64170 | 2.55 |
| θ | 3623 | 2.15 | 18940 | 1.52 | 7469 | 2.16 | 50301 | 2.00 |
| ʧ | 592 | 0.35 | 3744 | 0.30 | 1043 | 0.30 | 4463 | 0.18 |
| d | 7521 | 4.48 | 54284 | 4.36 | 14479 | 4.20 | 136187 | 5.42 |
| e | 18337 | 10.92 | 188196 | 15.12 | 34510 | 10.00 | 320140 | 12.74 |
| f | 1630 | 0.97 | 6217 | 0.50 | 4381 | 1.27 | 23042 | 0.92 |
| g | 1995 | 1.19 | 11359 | 0.91 | 4108 | 1.19 | 26138 | 1.04 |
| i | 14623 | 8.71 | 89799 | 7.22 | 31626 | 9.17 | 190756 | 7.59 |
| x | 1548 | 0.92 | 7681 | 0.62 | 3031 | 0.88 | 19362 | 0.77 |
| k | 6981 | 4.16 | 55863 | 4.49 | 13466 | 3.90 | 95427 | 3.80 |
| l | 5627 | 3.35 | 56107 | 4.51 | 14633 | 4.24 | 137148 | 5.46 |
| m | 5438 | 3.24 | 39278 | 3.15 | 9847 | 2.85 | 69445 | 2.76 |
| n | 11394 | 6.78 | 87775 | 7.05 | 23366 | 6.77 | 178012 | 7.09 |
| ɲ | 320 | 0.19 | 2427 | 0.19 | 1451 | 0.42 | 7729 | 0.31 |
| o | 15399 | 9.17 | 129208 | 10.38 | 31187 | 9.04 | 234238 | 9.32 |
| p | 4582 | 2.73 | 34135 | 2.74 | 7899 | 2.29 | 68687 | 2.73 |
| r | 1783 | 1.06 | 5236 | 0.42 | 6561 | 1.90 | 25016 | 0.99 |
| ɾ | 10992 | 6.55 | 63702 | 5.12 | 23603 | 6.84 | 155632 | 6.19 |
| s | 12453 | 7.42 | 100881 | 8.11 | 23998 | 6.96 | 184085 | 7.33 |
| t | 9168 | 5.46 | 56287 | 4.52 | 17201 | 4.99 | 108398 | 4.31 |
| u | 4601 | 2.74 | 39146 | 3.14 | 9091 | 2.64 | 76390 | 3.04 |
| ʎ | 959 | 0.57 | 10356 | 0.83 | 2905 | 0.84 | 13307 | 0.53 |
| TOTAL | 167896 | 100 | 1244411 | 100 | 344856 | 100 | 2511856 | 100 |

**Tabla 1.** *Frecuencia de los fonemas españoles.*

El procedimiento de extracción del inventario ha sido el mismo para los dos corpus.

Primero se extraen dos listas de palabras de cada corpus: el conjunto de las formas que aparecen y este mismo conjunto más el número de apariciones de cada forma. Es decir, en el primero estarían sólo los "types" y en el segundo tanto "types" como "tokens". El transcriptor se aplica a ambas listas y se obtiene respectivamente un léxico y un corpus fonológicos con las apariciones de cada palabra.

A continuación se silabifican las formas y se obtienen dos conjuntos de sílabas fonológicamente transcritas: las correspondientes al léxico y las correspondientes al corpus. Las sílabas de los dos conjuntos se cuentan y se ordenan por frecuencia, de mayor a menor. Con esto averiguamos la distribución de las sílabas en el léxico y la distribución real en el corpus.

En este punto obtenemos una tabla de sílabas para el léxico y otra para el corpus, datos de las apariciones de cada sílaba, su frecuencia relativa al total y su distribución (Tabla 2). Ahora se pueden examinar las distintas estructuras de sílaba y hallar su frecuencia en el léxico y en el corpus. Por limitaciones de espacio presentamos únicamente las 10 sílabas más frecuentes. El orden de presentación es la frecuencia del corpus oral.

El último paso supone el recuento de fonemas y el cálculo de la frecuencia total de cada unidad. A continuación se repite el proceso tomando en cuenta el contexto silábico. Dada la frecuencia de aparición de cada fonema se pueden hallar las probabilidades de su presencia en cualquier combinación. La Tabla 1 muestra los resultados de los 23 fonemas.

La Figura 1 muestra la distribución de las sílabas en el corpus escrito y en el oral. Un resultado interesante y significativo es que las primeras 100 sílabas representan más del 80% del corpus oral. Las primeras 650 sílabas cubren más del 99 % del corpus.

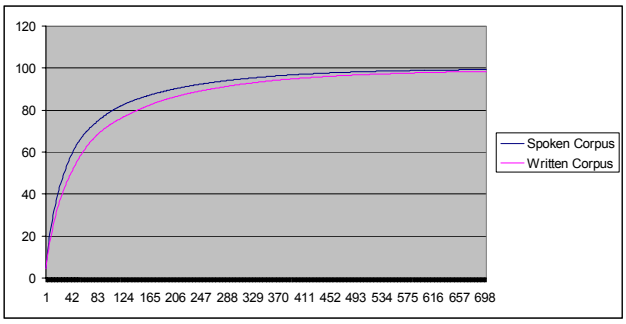


**Figura 1.** *Distribución de las sílabas en los corpus oral y escrito. La figura muestra la frecuencia relativa acumulada de las sílabas en español, ordenadas en orden decreciente.*

| Corpus oral | | | | | | Corpus escrito | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Lexicón | | | Corpus | | | Lexicón | | | Corpus | | |
| .a. | 2506 | 3,53 | .a. | 27606 | 4,94 | .a. | 4220 | 3,15 | .de. | 46748 | 4,49 |
| .do. | 1854 | 2,61 | .ke. | 21070 | 3,77 | .ta. | 2918 | 2,18 | .a. | 37021 | 3,55 |
| .ta. | 1621 | 2,28 | .de. | 19638 | 3,51 | .do. | 2613 | 1,95 | .la. | 27138 | 2,61 |
| .te. | 1406 | 1,98 | .es. | 13703 | 2,45 | .ka. | 2392 | 1,78 | .ta. | 17885 | 1,72 |
| .ka. | 1314 | 1,85 | .i. | 13102 | 2,34 | .ti. | 2264 | 1,69 | .ke. | 17704 | 1,70 |
| .de. | 1235 | 1,74 | .no. | 12781 | 2,28 | .te. | 2263 | 1,69 | .en. | 17203 | 1,65 |
| .ti. | 1182 | 1,66 | .te. | 10620 | 1,89 | .ra. | 2138 | 1,60 | .do. | 16840 | 1,62 |
| .to. | 1026 | 1,44 | .el. | 10282 | 1,84 | .na. | 1936 | 1,45 | .te. | 16610 | 1,59 |
| .ko. | 1019 | 1,43 | .la. | 10281 | 1,84 | .de. | 1828 | 1,37 | .na. | 15872 | 1,52 |
| .ra. | 992 | 1,39 | .do. | 10172 | 1,82 | .ko. | 1828 | 1,37 | .ma. | 15463 | 1,48 |

**Tabla 2.** *Las 10 sílabas más frecuentes en castellano.*

Por último mostramos el orden de distribución silábica por tipos (distintas combinaciones de vocal y consonante que aparecen en las sílabas en castellano)

| Tipo de sílaba | Frecuencia Relativa | Tipo de sílaba | Frecuencia relativa |
|---|---|---|---|
| .CV. | 51,35 | .CVV. | 3,37 |
| .CVC. | 18,03 | .CVVC. | 3,31 |
| .V. | 10,75 | .CCV. | 2,96 |
| .VC. | 8,60 | .CCVC. | 0,88 |

**Tabla 3.** *Distribución de frecuencias por tipos de sílaba en el corpus oral.*

## 5. CONCLUSIONES Y TRABAJO FUTURO

Éste es el primer inventario de frecuencias de fonemas y sílabas del castellano que emplea un corpus oral y otro escrito de tamaños comparables, y que emplea los mismos criterios y herramientas para segmentar las unidades.

Dos conclusiones importantes pueden derivarse de los datos:

1. El corpus escrito y el oral arrojan diferentes frecuencias. El orden de algunas unidades y el porcentaje de uso son diferentes. Esto es especialmente notable en el caso de las vocales /a, e, o/. Por lo tanto, el entrenamiento con modelos de lengua basados en textos escritos producirá peores resultados que el entrenamiento con textos orales.
2. Unas pocas sílabas en castellano permiten cubrir una parte significativa de un texto. De esto se sigue que el empleo de sílabas en lugar de fonemas como unidades en el desarrollo de tecnologías del lenguaje para el español parece prometedor. Nuestro equipo continuará investigando la segmentación en sílabas para el entrenamiento de sistemas de reconocimiento de español espontáneo.

Una vez que hayamos estimado la relevancia estadística de estos resultados con respecto a otro corpus oral similar, como nuestro corpus CORLEC, podremos ofrecer las garantías de un trabajo empírico sobre un número importante de datos orales espontáneos. Téngase en cuenta, que la obtención de un corpus oral de gran tamaño es una tarea muy costosa y que los corpus empleados están entre los más grandes disponibles para cualquier lengua.

## 10. BIBLIOGRAFÍA


[1] Alameda, J.R. & F. Cuetos (1995) *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Servicio de publicaciones de la Universidad de Oviedo.

[2] Álvarez, C.J., M. Carreiras & M. De Vega (1992) "Estudio estadístico de la ortografía castellana: (1) la frecuencia silábica" *Cognitiva* 4, pp.75-105.

[3] Guirao, M. & A. Borzone de Manrique (1972) "Fonemas, sílabas y palabras del español de Buenos Aires" *Filologia*, XVI, pp.135-165.

[4] Navarro Tomás, T. (1946) "Escala de frecuencia de los fonemas españoles" *Estudios de fonología española*. Syracuse, pp.15-30

[5] Quilis, A. & M. A. Esgueva Martínez (1980)"Frecuencia de fonemas en el español hablado"*Lingüística Española Actual*, 2.

[6] Rojo, G. (1991) "Frecuencia de fonemas en el español actual" en Brea, M. & F. Fernández Rei (Coords.) *Homenaxe ó profesor Constantino García.* Universidade de Santiago. pp.451-457.

[7] Zipf, G. K. & J. M. Rogers (1939) "Phonemes and Variphones in four present-day romance Languages and Classical Latin from the viewpoint of dynamic Philology" Archives Néerlandaises de Phonétique Expérimentale" 15, pp. 111-147.

[8] Alameda, J.R. & F. Cuetos (1995) *Diccionario de frecuencias de las unidades lingüísticas del castellano*. Servicio de publicaciones de la Universidad de Oviedo.

[9] Marcos Marín, F. (1992) “El Corpus Oral de Referencia de la Lengua Española contemporánea” Project Report. Madrid. Publisher in ftp://ftp.lllf.uam.es/pub/corpus/oral.