Frecuencia de fonemas en español actual

GUILLERMO ROJO

1. Los datos que expongo a continuación han sido obtenidos como derivación de una línea de trabajo muy alejada de los estudios fonéticos y fonológicos. En efecto, un grupo de investigadores y estudiantes vinculados al Departamento de Filología española, Teoría de la Literatura y Lingüística general de la Universidad de Santiago de Compostela está almacenando en ordenador un amplio conjunto de textos españoles e hispanoamericanos contemporáneos con el fin de constituir un corpus del español actual. Como parte del proceso de organización de los datos, se pasa a los textos introducidos un programa que hace recuentos de palabras gráficas (para elaborar luego índices, concordancias, etc.) (1) y que da como una de sus salidas posibles la relación de palabras diferentes que contiene el texto con el número de veces que aparece cada una de ellas. Situado ante este material, se me ocurrió que sería útil almacenar las palabras y sus frecuencias respectivas en una base de datos y escribir luego un programa que produjera la transcripción fonológica a la variedad estándar y, al tiempo, hiciera los recuentos de frecuencias (2).

Ya Lloyd y Schnitzer, de una parte, y Mosterín, de otra, habían utilizado el ordenador para sus recuentos. Los primeros perforaron tarjetas con la transcripción fonológica de 252.404 sílabas (70.755 palabras) y escribieron los programas necesarios para hallar las frecuencias de fonemas y tipos de sílabas (cf. Lloyd-Schnitzer, 1967, 60). Algunos años más tarde, Jesús Mosterín (1981, 203-205) grabó la transcripción fonológica de varios capítulos de una obra de Octavio Paz (58.620 fonemas en total) y le aplicó un programa para hallar la frecuencia de cada elemento. Para este trabajo se ha seguido un camino más largo, pero mucho más cómodo. Dadas las características del sistema ortográfico del español contemporáneo, no resulta excesivamente complicado escribir un programa que haga la transcripción fonológica. Una vez introducido el texto (en su forma ortográfica habitual), el proceso se hace de modo totalmente automático, con lo que el investigador se ahorra la parte más aburrida y menos gratificante del trabajo y, al tiempo, se hace posible pensar en trabajar con millones de fonemas. Para esta

(1) Empleamos para ello el programa Micro-OCP [= Oxford Concordance Program], de Oxford University Press.

(2) Para los ficheros de datos y los programas de transcripción utilicé dBASE IV, de Ashton-Tate. Los cuadros y gráficos que aparecen en este trabajo han sido creados en Lotus 1-2-3, de Lotus Development Corporation.

ocasión, he trabajado con diecisiete textos literarios y ensayísticos (3) que suponen un total de 817.085 palabras y 3.641.915 fonemas, lo que constituye el recuento fonológico más amplio que se ha hecho para el español y creo que para cualquier otra lengua.

2. Ya en el trabajo de Quilis y Esgueva (1980), que es, sin duda, el punto básico de referencia para cualquier estudio sobre frecuencias de fonemas en español, se destacaba la incómoda heterogeneidad que presentaban los recuentos realizados hasta aquel momento. En efecto, la esperable diversidad de los materiales empleados ha llegado al extremo de que Lloyd y Schnitzer (1967, 59-60) utilizan las principales entradas del DRAE (18ª ed., 1956) y algunos elementos de la letra *A* del *Diccionario* de Williams. No cabe duda de que el procedimiento empleado distorsiona los datos en dos sentidos distintos. De un lado, favorece los fonemas más frecuentes en las formas utilizadas tradicionalmente por los lexicógrafos como representantes del lexema (véase en la tabla 1, por ejemplo, las frecuencias que obtienen para /a/ y /e/, a las que no pueden ser ajenas la utilización sistemática de los infinitivos y la gran cantidad de verbos de la primera conjugación y el segundo lugar que ocupa /o/, único caso en todos los recuentos). De otro, en un diccionario las palabras aparecen una sola vez, sea cual sea su frecuencia. Con la técnica empleada, *a*, *de*, *en*, *el*, *la*, etc. tienen el mismo peso en el recuento que una palabra cuya frecuencia media en un texto sea del 0,01%.

El segundo factor de heterogeneidad radica en las unidades fonológicas empleadas. Zipf-Rogers (1939), Navarro Tomás (1946) y, parcialmente, Delattre (1965) consideran monofonemáticos los diptongos y los triptongos, con lo que se producen divergencias con otros recuentos (4). En el subsistema consonántico hay diferencias en el propio inventario (presencia o ausencia de los fonemas /w/ y / /, oposición de /s/ y /Θ/ y algunas otras), pero el factor que provoca las discrepancias fundamentales en los resultados obtenidos radica en la consideración o no de archifonemas y, en caso afirmativo, en qué posiciones se considera que hay neutralización.

Dado que algunas de estas cuestiones remiten a zonas de la teoría fonológica muy alejadas de la finalidad de este estudio, decidí escribir el programa de transcripción fonológica de tal forma que fuera fácil la comparación con los recuentos anteriores más útiles y mejor organizados. En consecuencia, adopté exactamente el criterio utilizado por Quilis y Esgueva (1980), casi idéntico al empleado por Alarcos (1971). Así pues, manejo aquí las unidades fonológicas siguientes:

---

(3) Vid. apéndice.

(4) Cf., por ejemplo, la tabla 1, correspondiente a las vocales. He recalculado los porcentajes de Navarro Tomás y Delattre. No es suficiente, como hacen Quilis y Esgueva, sumar los porcentajes de las vocales y los diptongos. Los diptongos tienen dos vocales y, por tanto, hay que contarlos dos veces, con lo que luego es necesario reconvertir en porcentajes las cantidades que han dejado de serlo. No puedo hacer lo mismo con los datos de Zipf y Rogers porque no he podido consultar su trabajo. Según indican Quilis y Esgueva, los diptongos y triptongos suponen el 2,02% del total en el recuento de estos autores. Tras las reconversiones oportunas, su porcentaje de vocales debe de situarse entre el 45% y el 45,5%.

a) cinco fonemas vocálicos;
b) diecinueve fonemas consonánticos;
c) cinco archifonemas (/B/, /D/, /G/, /N/ y /R/).

Estos últimos son considerados únicamente en posición postnuclear, lo mismo que Quilis y Esgueva (1980, 2-3). La diferencia con Alarcos está fundamentalmente en el caso de /R/, elemento en el que este autor incluye "todos los casos en que la oposición *r* / $\bar{r}$ no es pertinente" (1971, 199) (5).

Esas veintinueve unidades, con las peculiaridades que acabo de indicar para los archifonemas, han sido las empleadas para todos los textos. Queda claro, pues, que la transcripción fonológica se ha hecho en todos los casos de acuerdo con las características del que se considera habitualmente español estándar peninsular. Es evidente que un estudio de estas características exige trabajar como si todos los textos perteneciesen a la misma variedad lingüística, por lo que la elección de uno de los varios sistemas fonológicos existentes era forzosa. La distorsión que ello pudiera originar queda virtualmente anulada por el hecho de que casi siempre es posible hallar la frecuencia de un fonema determinado de un sistema distinto mediante la suma de los que le corresponden en la variedad seleccionada aquí (/s/ y /Θ/, / / y /j/, por ejemplo).

3. Las tablas 1 y 2 resumen (retocados en algunos casos) los resultados que arrojan los recuentos de que tengo noticia. La primera impresión que producen es, sin duda, la de una excesiva divergencia en los datos. A partir de lo apuntado en el parágrafo anterior es esperable que haya diferencias importantes en tal o cual fonema, pero no lo es tanto la aparición de una distancia excesiva en factores más generales. Sin embargo, la tabla 1 muestra que el porcentaje total que alcanzan las vocales oscila entre el 42,83% (recalculado) que se encuentra en Delattre (6) y el 47,55% obtenido por Quilis y Esgueva. Son casi cinco puntos, lo cual constituye una horquilla excesivamente ancha para un factor tan general.

Todos los recuentos realizados, salvo el de Mosterín, proceden de la integración de calas realizadas sobre diferentes textos, lo cual significa que las frecuencias y porcentajes que manejamos han sufrido ya un proceso de integración que ha ocultado sus diferencias internas. Tratando de medir el alcance que pueden tener las divergencias entre textos distintos, he aplicado la prueba del chi

(5) Cf. tabla 2. El alto porcentaje de /R/ que obtiene Alarcos se debe, claro, a la atribución a este archifonema de los casos que se dan en posición inicial.

(6) Considero que el porcentaje de Delattre es el más bajo porque los resultados de Zipf y Rogers son más altos aunque nos limitemos simplemente a añadir el porcentaje de diptongos sin recalcular el resultado final: 43,56%. Debo añadir, de todas formas, que el recuento realizado por este autor, aparte de ciertas peculiaridades en el inventario de fonemas, quizá explicables por su deseo de comparar varias lenguas, presenta algunos problemas matemáticos. La suma de los porcentajes sobre el total atribuidos a los fonemas vocálicos españoles suma 42,69% (Delattre, 1965, 97). Sin embargo, poco antes ha indicado que el español, como el francés, muestra "a high proportion of vowels (against consonants): 43,6%" (ib., 63). Una divergencia semejante se observa en las otras tres lenguas comparadas.

cuadrado (chi^2) (7) a la distribución de vocales y consonantes de los cinco textos más extensos de mi muestra (8). El resultado obtenido (63,47) resulta muy superior a 9,49, que es el máximo que podría ser atribuido al azar en una tabla de este tipo (cuatro grados de libertad) con un margen de seguridad del 95%.

Aunque es bien sabido que el chi^2 de una tabla asciende de forma espectacular con el aumento del tamaño de la muestra (9), la única conclusión estadística posible ante estos resultados sigue siendo que hemos de rechazar la hipótesis nula y considerar que las diferencias observadas no pueden ser debidas al azar. Dado que estos cinco textos han sido transcritos exactamente con el mismo criterio, debemos concluir que hay factores internos a los textos que producen discrepancias de entidad en la frecuencia de los fonemas. Volveremos sobre esta cuestión en el apartado 5.

4. Veamos ahora los resultados más destacados de la muestra estudiada (cf. tablas 3, 4, 5 y 6). En primer lugar, vocales y consonantes presentan una distribución global próxima al 50%: el 47,12% y el 52,88%, respectivamente. Constituye el de las vocales un porcentaje más bajo que el obtenido en los recuentos de Alarcos, Guirao-Borzone y Quilis-Esgueva (cf. tabla 1). Según era de esperar, los primeros lugares de la escala de frecuencias (cf. tabla 6) están ocupados por las vocales /a/, /e/ y /o/. El fonema /i/ ocupa el quinto lugar.

Como ya señalaron Quilis y Esgueva (1980, 15), los recuentos difieren en cuál es el fonema más frecuente en español. Estos dos autores consideran que /e/ ocupa el primer lugar en los recuentos realizados sobre textos orales, mientras que pasa al segundo cuando se trabaja con textos escritos. Los datos de la tabla 3 indican con toda claridad que esta hipótesis ha de ser rechazada: /e/ ocupa el primer lugar en siete de los textos examinados (todos ellos escritos, por supuesto) y lo mismo ocurre en la muestra manejada por Mosterín.

De los diecisiete textos examinados aquí se deduce más bien que estamos ante dos fonemas de frecuencias tan próximas entre sí que los factores específicos de cada texto (los mismos que hacen que un fonema presente porcentajes ligeramente distintos en cada caso) pueden provocar que ambos elementos intercambien sus posiciones. En diez textos, el fonema que ocupa el primer lugar es /a/, lo cual se refleja en la pequeña ventaja (0,13) que obtiene con respecto a /e/ en los valores medios (cf. tabla 4). En el conjunto de la muestra, en cambio, el peso de

(7) Es una prueba que indica si las diferencias entre las frecuencias halladas y las esperadas pueden ser debidas o no al azar. Para detalles, vid. cualquiera de los manuales de Estadística lingüística citados en las referencias bibliográficas.

(8) Son *Ratón*, *Sonrisa*, *Usos*, *Tiempo* y *Laberinto*, todos ellos con más de trescientos mil fonemas cada uno.

(9) Veámoslo de forma práctica. El texto más largo de los examinados contiene 180.961 vocales y 206.316 consonantes. Supongamos que existiera un texto con únicamente una diferencia equivalente al 0,25% de la muestra (968 vocales más y 968 consonantes menos, por ejemplo), que es una diferencia que intuitivamente consideramos perfectamente atribuible al puro azar. El chi^2 correspondiente daría como resultado 4,86, que ya es superior al máximo que podemos atribuir al azar en una tabla con un grado de libertad y un margen de seguridad del 95% (3,84). Una oscilación porcentualmente idéntica con una muestra diez veces menor arroja un chi^2 de 0,49, que está comprendida dentro de los márgenes que pueden ser debidos al azar.

los textos más largos ha eliminado esa diferencia y una inesperada casualidad ha hecho coincidir los porcentajes: 13,46% en ambos casos (10).

Con independencia de cuál sea el más frecuente, es importante observar la distancia que existe entre ellos y los demás fonemas. Suponen en conjunto el 26,92% del total y el 57,12% de las vocales (esto es, una de cada cuatro realizaciones de fonemas en un texto corresponde a /a/ o /e/). El tercero de la escala, /o/, aparece ya a casi cuatro puntos de distancia. Sólo el recuento de Lloyd y Schnitzer, muy probablemente por las características de su muestra, presenta una ordenación diferente: /a/, /o/, /e/ y, además, /i/ está a poco más de un punto de /e/ (cf. tabla 1).

Como deja ver la tabla 6, las frecuencias de los fonemas están distribuidas de un modo muy jerarquizado: sólo hay dos fonemas que superen el 10% y únicamente cinco están situados entre el 5% y el 10%; ahora bien, estos siete fonemas (solo la cuarta parte de los componentes del sistema) suponen en conjunto el 61,75% de las frecuencias totales. Al otro extremo, los once fonemas que no superan el 1% suman en conjunto solamente el 4,61% (esto es, más o menos el peso de un fonema de frecuencia media como /d/ o la tercera parte de las apariciones de /a/).

Entre las consonantes destaca, con respecto a otros recuentos, la aparición de /s/ en cuarto lugar (superando, aunque por muy poco, a /i/) y de /l/, que ocupa el segundo lugar entre las consonantes (el quinto en Quilis y Esgueva).

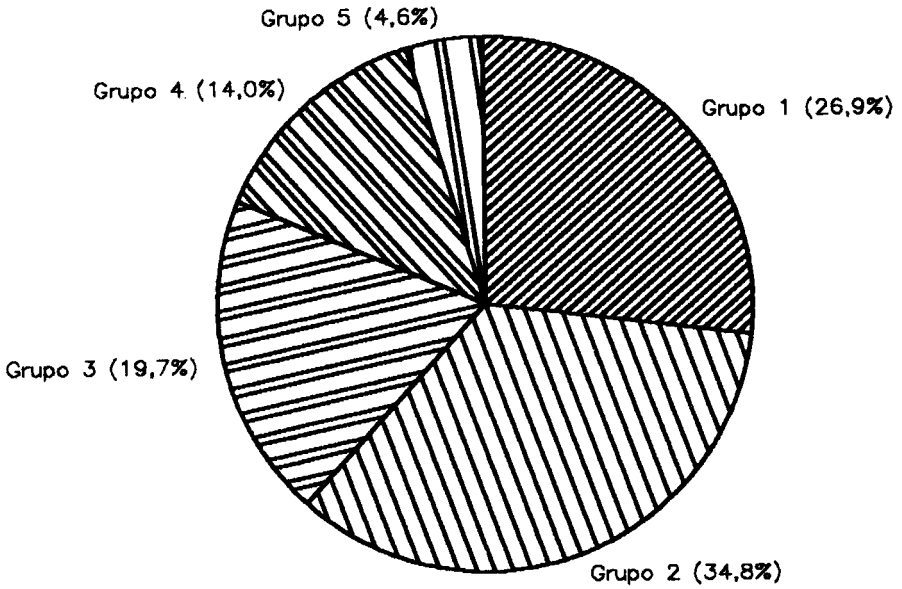
Sin pretender negar esas discrepancias, hay que dar más importancia al hecho de que los siete fonemas consonánticos más frecuentes sean los mismos en ambos recuentos. Las diferencias comienzan a partir del octavo lugar, ya por debajo del 3% sobre el total. Por el extremo opuesto de la escala, coinciden los trece fonemas de menor frecuencia. Aquí hay mayor coincidencia todavía que en la zona alta. El factor más divergente es la posición de /ǰ/, que aparece como antepenúltimo en mi recuento y cuatro puestos más arriba en el de Quilis y Esgueva.

Así pues, en términos generales podemos establecer cinco grupos de fonemas según su situación en la escala de frecuencias. En primer lugar, /a/ y /e/, los únicos que rebasan el 10%. Viene luego el grupo de los fonemas comprendidos entre el 5% y el 9,9%, en el que se encuentran los cinco fonemas ya mencionados. Otros cinco elementos (cuatro consonantes y la vocal /u/) presentan frecuencias comprendidas entre el 3% y el 4,9%. Hay seis entre 1% y 2,9% y, por fin, los once ya aludidos que no llegan al 1%. Los porcentajes acumulados de la tabla 6 muestran la asimetría a que he hecho referencia antes: los cinco fonemas más frecuentes (cuatro vocálicos y /s/ superan el 50% del total; los doce más frecuentes (las cinco vocales y siete consonantes, todos ellos con más del 3%) suponen más del 80%; el 18,6% restante se reparte (desigualmente, por supuesto) entre los otros diecisiete fonemas y archifonemas. El gráfico 1 da una visión clara de todo ello.

(10) Los porcentajes están redondeados al segundo decimal. La tabla 5 muestra que las frecuencias son ligeramente distintas (43 casos), por lo que aparecen ya diferencias si se llega al tercer decimal:

/a/	13,457
/e/	13,456

Gráfico 1



Una asimetría semejante se observa al comparar las frecuencias conjuntas de tipos de sonidos (cf. tabla 7). Las oclusivas orales (que incluyen todas las realizaciones de /b/, /d/ y /g/) suponen casi el 40% del total de las consonantes. Si sumamos las nasales, los dos tipos de oclusivas alcanzan el 56,25% (casi tres de cada cinco consonantes son oclusivas). En cuanto a la clasificación por punto de articulación, la tabla deja bien claro el predominio de las alveolares (que incluyen /N/): suponen la mitad de todas las consonantes (el 49,24% en el recuento de Quilis y Esgueva). Trabajando con grandes zonas de articulación, la escasa importancia de la que podemos considerar posterior (palatales y velares) es evidente: sólo el 12,67%.

5. Ya hice referencia en el § 3 a la impresión de excesiva divergencia que produce la comparación de los resultados obtenidos en los recuentos realizados hasta el momento. La importancia de este rasgo deriva fundamentalmente del hecho de que la frecuencia de los elementos del sistema fonológico propio de un estado de lengua determinado constituye un caso evidente de fenómeno cuya distribución real en la población (el conjunto de los textos) no puede ser conocido nunca. De ahí que sea forzoso trabajar mediante la proyección al conjunto de la

población de los valores obtenidos en el estudio de una muestra. El problema radica, naturalmente, en que las muestras analizadas proporcionan resultados excesivamente distanciados entre sí, por lo que surgen dudas acerca de la validez de las proyecciones posibles.

Hay varios factores que pueden explicar estas divergencias. En primer lugar, las muestras han sido seleccionadas de modos distintos y con criterios también divergentes. Como ya he indicado, Lloyd y Schnitzer (1967) trabajan con una muestra muy amplia (debe de estar alrededor de los 500.000 fonemas), pero no utilizan textos, sino entradas de diccionario, con lo que sus resultados forzosamente han de ser muy diferentes de los obtenidos en el análisis de textos. Otras muestras resultan excesivamente reducidas (la de Zipf y Rogers (1939), por ejemplo, con solo 5.000 fonemas), con lo que el riesgo de que estén sesgadas es bastante alto.

En segundo lugar, los criterios empleados en la transcripción son muy distintos. Nótese, por ejemplo, lo que muestra la tabla 2 para /n/, que presenta una desviación estándar equivalente al 49% de la media (11), o para /R/, que llega al 41% teniendo en cuenta únicamente aquellos recuentos en que se considera su existencia (12). Los diecisiete textos que constituyen la muestra estudiada aquí son bastante diferentes entre sí, pero las desviaciones estándar que muestran en estos casos son mucho más bajas: suponen el 11,56% y el 9,48% de la media, respectivamente.

Todas estas discrepancias en el modo de seleccionar las muestras o los criterios empleados para la transcripción hacen que la comparación de los resultados que arrojan los distintos recuentos ofrezca escaso interés. Las medias que aparecen en las tablas 1 y 2 están realizadas, como es lógico, dando el mismo peso a cada recuento, con lo que, al ser pocos, cualquier desviación tiene un fuerte efecto sobre las medias y las distorsiona.

De todos modos, es también evidente que la frecuencia de los fonemas varía según los textos. En los diecisiete que componen la muestra estudiada aquí, transcritos todos con el mismo criterio y ninguno de ellos inferior a 100.000 fonemas, existen las fuertes diferencias que aparecen en la tabla 4. Es cierto que las mayores oscilaciones tienen lugar en elementos de escasa frecuencia, en los que cualquier alteración tiene grandes repercusiones, pero también aparecen variaciones de cierta importancia en fonemas como /b/, /l/, /m/, /n/, /R/, /s/, /Θ/ o /i/, en todos los casos iguales o superiores al 8% del valor medio correspondiente.

En realidad, sería ingenuo esperar mayor uniformidad. No se puede perder de vista que la frecuencia de los fonemas en un texto depende básicamente de la frecuencia de las palabras que lo forman. En los textos más amplios de los estudiados aquí, algunas de las palabras más frecuentes en español muestran también

(11) Este mismo coeficiente es mucho menor en el caso de /N/ porque sólo cuentan los casos en que se ha trabajado con el archifonema.

(12) En este caso está clara la causa: Alarcos considera más casos de neutralización. Cf. supra, nota 5.

importantes oscilaciones (13). Es lógico, por tanto, que encontremos discrepancias importantes en la frecuencia con que aparecen los fonemas en textos distintos y, como consecuencia de lo anterior, los resultados que arrojan los recuentos (que suponen la integración previa de los fragmentos o los textos estudiados) también presentan divergencias.

La única forma de vencer esas dificultades o, cuando menos, de paliar sus efectos consiste en trabajar sobre muestras amplias y suficientemente variadas. En ese sentido, el corpus estudiado en este trabajo, constituido por algo más de tres millones y medio de realizaciones de fonemas, constituye un conjunto de entidad suficiente para ser considerado como representativo de la población (14). Tal como deja ver la tabla 3, contiene textos de características bastante heterogéneas y, como consecuencia de ello, es también un buen reflejo de la diversidad existente.

Su mayor defecto es, sin duda, el no contener transcripciones de textos orales, lo cual resultaría un inconveniente de cierta importancia si llegara a demostrarse que las frecuencias fonemáticas en textos orales difieren sistemáticamente de las que se encuentran en textos escritos. Para evitar esa dificultad y aprovechar al máximo las ventajas de haber adoptado un sistema de transcripción idéntico al empleado por Quilis y Esgueva, me he permitido fundir los dos recuentos en un conjunto único (cf. tabla 8). Se alcanza así un total de algo más de 3.800.000 realizaciones de fonemas. La diferencia de tamaños de ambos recuentos hace que el peso de la muestra de Quilis y Esgueva (que supone el 4,21 % del total) sea reducido, pero introduce todavía mayor variedad en el conjunto y, por tanto, permite llegar a resultados más fiables (15).

En la tabla 8 encontramos, entre otros datos, los porcentajes de aparición correspondientes a cada fonema y, por tanto, la probabilidad general de aparición de cada uno de ellos: la probabilidad de encontrar una realización de /a/ es 0,134, la de dar con una realización de /k/ es 0,0382, la de una consonante cualquiera es igual a 0,5286, etc. La probabilidad de aparición de cada uno de los elementos en la muestra que hemos construido es, sin duda, la mejor estimación que podemos manejar para intentar conocer la que cada uno de ellos presenta en la población (el conjunto de los textos). El cálculo no es complicado. Además de la proporción de los distintos elementos, necesitamos conocer el error estándar de cada uno de ellos (16). El producto del error estándar por la puntuación *z* correspondiente al

(13) Así, *a* oscila entre el 1,90% y el 2,86%, *de* lo hace entre el 4,58% y el 6,71%, etc. Inciden también sobre la frecuencia de los fonemas las diferencias en el léxico empleado, aunque este factor es mucho más difícil de cuantificar.

(14) El tamaño de muestra que se necesita en un fenómeno con un porcentaje de aparición del 12,5% para que podamos estar seguros de no cometer una desviación superior al 0,1% en el 99% de los casos (cf. Butler, 1985, 62) es 728.044. La muestra estudiada aquí es cinco veces mayor que la exigida con esos requisitos, de modo que podemos estar razonablemente seguros de los valores obtenidos.

(15) Nótese que la adición de esta segunda muestra deshace el empate entre /a/ y /e/. /e/ pasa a primera posición, aunque sea con solo 0,11 de diferencia, a pesar del carácter mayoritariamente escrito del corpus manejado.

(16) El error estándar de una proporción está en función de la proporción y del tamaño de la muestra. La fórmula es

$\sqrt{p(1-p)/N}$, donde p es la probabilidad del elemento y N es el tamaño de la muestra.

margen de confianza con que deseamos trabajar (en este caso he utilizado el 99%, con $z = 2{,}58$) da las cantidades que hay que sumar o restar a la proporción obtenida para obtener los límites entre los que debe encontrarse el valor del total de la población.

Esos datos, reconvertidos ya en porcentajes, figuran también en la tabla 8. Significan que podemos tener una seguridad del 99% en que el porcentaje de /a/ en el total de la población oscila entre 13,36 y 13,45 (como se ve, a una distancia de 0,04 ó 0,05 del resultado obtenido en el análisis de la muestra) (17). De modo semejante, el porcentaje de /m/ estará situado entre 2,56 y 2,60, el del total de las vocales entre 47,07 y 47,20, etc. (18). Los datos de esta tabla constituyen, creo, la mejor aproximación que puede hacerse a la distribución de los fonemas en español: establecen el terreno en cuyo interior debe estar el porcentaje de cada fonema en el conjunto de la población. Aunque ello no implica que las frecuencias halladas en un texto concreto tengan que estar comprendidas entre esos límites, los márgenes calculados constituyen un punto de referencia de indudable utilidad para situar cualquier texto en relación a los valores generales de la población.

REFERENCIAS BIBLIOGRAFICAS

Alarcos Llorach, E. (1971), *Fonología española*, Madrid, Gredos, 1971[4].

Butler, Ch., (1985), *Statistics in Linguistics*, Oxford, Blackwell, 1985.

Delattre, P. (1965), *Comparing the phonetic features of English, German, Spanish and French*, Heidelberg, Groos, 1965.

Guirao, M. y A. M. Borzone de Manrique (1972), "Fonemas y sílabas y palabras en el español de Buenos Aires", *Filología*, 16, 1972, págs. 135-165.

Haber, A. y R. P. Runyon (1973), *General Statistics*, Reading (Mass.), Addison-Wesley. Trad. esp. de R. Lassala Mozo, Bogotá, 1973.

Lloyd, P. M. y R. D. Schnitzer (1967), "A Statistical Study of the Structure of the Spanish Syllabe", *Linguistics*, 37, 1967, págs. 58-72.

Mosterín, J. (1981), *La ortografía fonémica del español*, Madrid, Alianza, 1981.

Muller, Ch. (1968), *Initiation a la Statistique linguistique*, París, Larousse, 1968. Versión esp. de A. Quilis, *Estadística lingüística*, Madrid, Gredos, 1973.

Navarro Tomás, T. (1946), "Escala de frecuencias de los fonemas españoles", en *Estudios de fonología española*, Nueva York, Las Américas, 1966[2], págs. 15-30.

Quilis, A. y M. Esgueva (1980), "Frecuencia de fonemas en el español hablado", *Lingüística española actual*, 2/1, 1980, págs. 1-25.

Woods, A., P. Fletcher y A. Hughes (1986), *Statistics in Language Studies*, Cambridge, Cambridge Univ. Press, 1986.

(17) La asimetría es producida por la acumulación de redondeos al segundo decimal.

(18) La proximidad de los límites se debe, claro, al enorme tamaño de la muestra, que es lo que da fiabilidad al recuento. Si hubiéramos obtenido el mismo porcentaje de /a/ en una muestra de 100.000 elementos, la proyección de ese resultado a la población daría un error estándar de 0,00105 y, como consecuencia de ello, los límites estarían situados entre un mínimo de 13,13% y un máximo de 13,67%. Como se ve, hemos pasado a una oscilación de $\pm$ 0,27.

Zipf, G. K. y F. 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 experimentale*, 15, 1939, 111-147.

Apéndice: Textos utilizados

[*Bunge*]	Bunge, Mario: *Lingüística y filosofía*, Barcelona, Ariel, 1983.
[*Caimán*]	Buero Vallejo, A.: *Caimán*, Madrid, Espasa-Calpe, 1981.
[*Carta*]	Colinas, Antonio: *Larga carta a Francesca*, Barcelona, Seix Barral, 1986.
[*Crónica*]	García Márquez, Gabriel: *Crónica de una muerte anunciada*, Madrid, Mondadori, 1987.
[*Diego*]	Poniatowska, Elena: *Querido Diego, te abraza Quiela y otros cuentos*, Madrid, Alianza / Era, 1987.
[*Glenda*]	Cortázar, Julio: *Queremos tanto a Glenda*, Madrid, Alfaguara, 1981, 4ª. edición.
[*Historias*]	Bioy Casares, A.: *Historias desaforadas*, Madrid, Alianza, 1986.
[*Jóvenes*]	Aldecoa, Josefina: *Porque éramos jóvenes*, Barcelona, Seix Barral, 1986.
[*Laberinto*]	Mendoza, Eduardo: *El laberinto de las aceitunas*, Barcelona, Seix Barral, 1982.
[*Mirada*]	Guelbenzu, José María: *La mirada*, Madrid, Alianza, 1987.
[*Paisajes*]	Goytisolo, Juan: *Paisajes después de la batalla*, Barcelona, Montesinos, 1982.
[*Ratón*]	Sánchez Ferlosio, Rafael: *La homilía del ratón*, Madrid, El País, 1986.
[*Sonrisa*]	Sampedro, Jose Luis: *La sonrisa etrusca*, Madrid, Alfaguara, 1985.
[*Sur*]	García Morales, Adelaida: *El sur (seguido de Bene)*, Barcelona, Anagrama, 1985.
[*Ternura*]	Martínez de Pisón, I.: *La ternura del dragón*, Barcelona, Anagrama, 1988, 3ª. edición.
[*Tiempo*]	Paz, Octavio: *Tiempo nublado*, Barcelona, Seix Barral, 1983.
[*Usos*]	Martín Gaite, Carmen: *Usos amorosos de la postguerra española*, Barcelona, Anagrama, 8ª ed., 1988.

	Zipf-Rogers	Navarro	Alarcos	Delattre	Lloyd-Schnitzer	Guirao-Borzone	Quilis-Esgueva	Mosterín	Rojo	Media	Desv. estándar	Coefic. variación
/a/	14,06	13,55	13,70	13,06	15,21	12,45	12,19	12,13	13,46	13,22	0,95	7,14 %
/e/	12,20	12,92	12,60	14,06	9,73	14,51	14,67	13,89	13,46	13,23	1,48	11,32 %
/i/	4,20	6,81	8,60	4,63	8,53	7,27	7,38	7,97	7,51	7,34	1,18	16,82 %
/o/	9,32	9,11	10,30	9,21	10,27	9,85	9,98	9,23	9,55	9,69	0,45	4,66 %
/u/	1,76	2,82	2,10	1,87	2,77	3,08	3,33	3,27	3,15	2,80	0,51	18,91 %
diptongos	2,02											
tot. voc.	43,56 (sin recalcular)	45,21 (recalculados)	47,30	42,83 (recalculados)	46,51	47,16	47,55	46,49	47,13	46,27 (sin Zipf Rogers)	1,47	3,20 %
	(N=5.000)	(N= 20.000)	?	?	(N=aprox. 500.000)	(N= 62.980)	(N= 160.000)	(N= 58.620)	(N= 3.641.915)			

TABLA 1
Porcentaje de vocales según diferentes recuentos

	Zipf-Rogers	Navarro	Alarcos	Lloyd-Delattre	Schnitzer	Guirao-Borzone	Quilis-Esgueva	Mosterín	Rojo	Media	Desv. estándar	Coefic. variación
/b/	3,26	2,46	2,50	2,85	2,92	2,45	2,37	2,10	2,65	2,54	0,25	9,76 %
/B/			0,10		0,13		0,03		0,08	0,09	0,04	42,82 %
/c/	0,30	0,29	0,40	0,32	0,57	0,33	0,37	0,15	0,27	0,34	0,11	33,35 %
/d/	5,06	4,85	4,00	5,20	3,81	4,16	4,24	4,59	4,72	4,45	0,44	9,92 %
/D/			0,25		0,23		0,31		0,25	0,26	0,03	11,54 %
/f/	0,72	0,70	1,00	0,52	1,46	0,67	0,55	0,88	0,68	0,81	0,29	35,67 %
/g/	1,02	1,01	1,00	0,71	1,46	0,94	0,94	0,84	0,87	0,97	0,21	21,19 %
/G/			0,25		0,31		0,28		0,22	0,27	0,03	12,66 %
/j/	2,40	0,39	0,40	2,94	0,15		0,41	0,09	0,21	0,66	0,94	143,40 %
/k/	3,84	4,10	3,80	4,64	3,76	4,37	3,98	3,95	3,81	4,05	0,29	7,15 %
/l/	5,20	5,29	4,70	3,84	2,96	4,25	4,23	4,98	5,12	4,42	0,72	16,33 %
/m/	2,98	3,00	2,50	3,73	2,48	3,04	3,06	2,74	2,56	2,89	0,39	13,52 %
/n/	5,94	6,73	2,70	7,01	2,34	7,67	2,78	7,99	2,39	4,95	2,43	49,05 %
/N/			3,70		4,44		4,86		5,10	4,53	0,53	11,75 %
/ɲ/	0,36	0,35	0,20	0,30	0,28	0,28	0,25	0,00	0,19	0,23	0,10	43,27 %
/p/	2,92	2,97	2,10	2,29	2,36	2,76	2,77	2,40	2,59	2,53	0,27	10,79 %
/ɾ/	1,04	0,78	0,60		1,17	0,50	0,43	0,81	0,73	0,72	0,23	31,67 %
/r/	5,90	5,73	2,50	6,23	8,24	5,58	3,26	5,59	3,66	5,10	1,74	34,13 %
/R/			4,50				1,93		2,11	2,85	1,17	41,15 %
/s/	8,12	8,24	8,00	8,35	4,26	9,72	8,32	9,01	7,55	7,93	1,52	19,13 %
/t/	4,46	4,67	4,60	4,74	5,32	4,92	4,53	4,40	4,31	4,69	0,30	6,38 %
/x/	0,58	0,49	0,70	0,37	1,02	0,65	0,57	0,73	0,73	0,66	0,18	27,53 %
/ʎ/	0,60	0,58	0,50	0,47	0,60		0,38	0,31	0,38	0,46	0,10	21,89 %
/θ/	1,74	2,16	1,70	1,42	2,49		1,45	1,95	1,69	1,84	0,36	19,60 %
/w/				1,25				0,01		0,63	0,62	97,94 %
/ʒ/						0,55				0,55		
Tot. cons.	56,44 (sin diptongos)	54,79 (recalculado)	52,70	57,18 (recalculado)	52,76	52,84	52,30	53,52	52,87	53,62 (sin Zipf Rogers)	1,52	2,84 %
	(N=5.000)	(N= 20.000)	?	?	(N=aprox. 500.000)	(N= 62.980)	(N= 160.000)	(N= 58.620)	(N= 3.641.915)			

TABLA 2
Porcentajes de consonantes según diferentes recuentos

	Caimán	Crónica	Bunge	Mirada	Sur	Ternura	Glenda	Historias	Diego	Paisajes	Carta	Jóvenes	Laberinto	Tiempo	Usos	Sonrisa	Ratón
/b/	2,58	3,08	1,77	2,97	3,00	3,23	2,97	2,96	2,89	2,58	3,10	3,44	2,72	2,01	2,51	2,91	1,71
/B/	0,04	0,03	0,40	0,04	0,05	0,05	0,07	0,05	0,05	0,12	0,06	0,07	0,07	0,10	0,07	0,04	0,14
/c/	0,47	0,24	0,23	0,24	0,25	0,25	0,32	0,25	0,36	0,23	0,27	0,28	0,28	0,17	0,32	0,32	0,21
/d/	4,04	5,15	4,02	4,98	4,53	4,55	4,84	4,36	4,40	4,84	5,06	4,92	4,71	4,71	4,74	4,21	5,39
/D/	0,19	0,12	0,29	0,21	0,18	0,17	0,22	0,23	0,19	0,27	0,22	0,38	0,21	0,35	0,25	0,25	0,31
/f/	0,59	0,61	0,90	0,56	0,51	0,64	0,65	0,62	0,59	0,78	0,71	0,63	0,60	0,73	0,73	0,56	0,94
/g/	0,74	0,89	1,14	0,80	0,87	1,09	0,87	0,97	0,96	0,87	0,86	0,91	0,92	0,71	0,79	0,87	0,82
/G/	0,10	0,13	0,64	0,17	0,10	0,10	0,14	0,17	0,12	0,32	0,15	0,14	0,21	0,39	0,24	0,11	0,39
/j/	0,40	0,27	0,13	0,18	0,40	0,19	0,23	0,24	0,24	0,17	0,14	0,23	0,28	0,12	0,17	0,33	0,11
/k/	3,32	3,87	4,03	3,65	4,03	3,80	3,60	4,12	3,71	3,57	3,61	3,48	4,04	3,46	3,92	3,61	4,45
/l/	4,29	5,22	5,22	5,37	3,73	5,48	5,29	4,80	5,04	5,07	5,06	5,16	4,80	5,69	4,88	5,03	5,74
/m/	2,35	2,47	2,33	2,14	3,05	2,46	2,61	2,81	2,70	2,41	2,34	2,30	2,75	2,41	2,50	2,52	2,91
/n/	3,28	2,21	2,44	1,99	2,46	2,14	2,31	2,46	2,34	2,33	2,32	2,22	2,32	2,79	2,41	2,49	2,22
/N/	4,15	5,15	5,87	5,28	5,16	5,00	5,13	5,13	4,97	5,15	4,82	4,90	5,04	5,10	5,27	5,11	5,19
/ɲ/	0,20	0,16	0,06	0,18	0,20	0,25	0,14	0,19	0,23	0,14	0,16	0,20	0,20	0,11	0,23	0,35	0,15
/p/	2,28	2,78	2,60	2,57	2,51	2,39	2,45	2,83	2,49	2,44	2,39	2,59	2,74	2,35	2,59	2,55	3,01
/ɾ/	1,31	0,63	0,48	0,77	0,67	0,63	0,66	0,74	0,78	0,82	0,76	0,69	0,74	0,77	0,64	0,86	0,58
/r/	3,58	3,73	3,30	3,47	3,66	3,58	3,72	3,94	3,80	3,64	3,70	3,64	3,65	3,66	3,66	3,76	3,53
/R/	2,47	2,49	1,87	2,35	2,10	2,09	2,12	2,37	1,98	2,06	2,15	2,09	2,25	1,69	2,15	1,95	2,23
/s/	8,82	7,17	8,68	7,28	7,28	7,39	7,48	6,51	7,96	8,30	7,72	7,33	7,12	8,74	8,00	7,61	6,06
/t/	4,57	3,95	4,90	4,27	4,32	4,07	4,06	4,21	4,32	4,73	4,08	4,00	4,44	4,47	4,36	4,27	4,21
/x/	0,59	0,67	0,72	0,72	0,68	0,69	0,78	0,80	0,76	0,68	0,69	0,99	0,70	0,57	0,74	0,92	0,67
/ʃ/	0,48	0,40	0,13	0,45	0,63	0,50	0,41	0,31	0,43	0,36	0,56	0,36	0,41	0,15	0,34	0,50	0,27
/θ/	1,45	1,39	2,00	1,89	1,63	1,53	1,58	1,42	1,45	1,78	1,79	1,63	1,66	2,00	1,73	1,36	2,03
Tot. cons.	52,29	52,81	54,15	52,53	52,00	52,27	52,65	52,49	52,76	53,66	52,72	52,58	52,86	53,25	53,24	52,49	53,27
/a/	13,64	14,35	11,71	14,02	14,40	13,59	14,05	13,15	13,88	12,77	14,43	14,00	12,68	12,02	13,90	13,88	13,33
/e/	12,75	12,62	13,26	13,52	13,89	13,37	13,44	13,94	13,42	13,01	13,06	13,30	13,85	12,45	13,62	13,94	14,12
/i/	7,68	7,08	8,47	6,83	7,50	7,71	7,07	7,04	6,97	7,81	7,18	7,47	7,61	9,39	7,47	6,61	7,40
/o/	10,46	10,07	8,99	9,64	8,74	9,71	9,70	10,01	9,74	9,21	9,26	9,41	9,83	9,78	8,75	9,87	9,53
/u/	3,16	3,08	3,41	3,48	3,46	3,36	3,07	3,36	3,20	3,55	3,36	3,25	3,17	3,11	3,04	3,21	2,35
Tot. cons.	47,69	47,20	45,84	47,49	47,99	47,74	47,33	47,50	47,21	46,35	47,29	47,43	47,14	46,75	46,78	47,51	46,73
Total	99,98	100,01	99,99	100,02	99,99	100,01	99,98	99,99	99,97	100,01	100,01	100,01	100,00	100,00	100,02	100,00	100,00
N=	103.096	121.513	131.931	134.390	136.861	139.307	173.414	175.525	202.693	207.194	214.895	222.651	300.552	314.050	324.126	352.440	387.277

TABLA 3
Frecuencia de los fonemas en los distintos textos (en porcentajes)

	Valores medios	Valor mínimo	Valor máximo	Variación estándar	Coefic. variación
/b/	2,73	1,71	3,44	0,48	17,51 %
/B/	0,09	0,03	0,40	0,08	98,18 %
/c/	0,28	0,17	0,47	0,07	24,08 %
/d/	4,67	4,02	5,39	0,37	7,97 %
/D/	0,24	0,12	0,38	0,06	27,14 %
/f/	0,67	0,51	0,94	0,11	17,16 %
/g/	0,89	0,71	1,14	0,11	12,08 %
/G/	0,21	0,10	0,64	0,14	66,41 %
/j/	0,23	0,11	0,40	0,09	38,46 %
/k/	3,78	3,32	4,45	0,28	7,44 %
/l/	5,05	3,73	5,74	0,47	9,33 %
/m/	2,53	2,14	3,05	0,23	9,23 %
/n/	2,40	1,99	3,28	0,28	11,58 %
/N/	5,08	4,15	5,87	0,32	6,26 %
/ɲ/	0,19	0,06	0,35	0,06	33,17 %
/p/	2,56	2,28	3,01	0,18	7,17 %
/r̄/	0,74	0,48	1,31	0,17	23,05 %
/r/	3,65	3,30	3,94	0,14	3,71 %
/R/	2,14	1,69	2,49	0,20	9,47 %
/s/	7,61	6,06	8,82	0,73	9,58 %
/t/	4,31	3,95	4,90	0,25	5,79 %
/x/	0,73	0,57	0,99	0,10	13,89 %
/λ/	0,39	0,13	0,63	0,13	32,32 %
/θ/	1,67	1,36	2,03	0,22	12,94 %
Tot. cons.	52,82	52,00	54,15	0,53	1,00 %
/a/	13,52	11,71	14,43	0,78	5,80 %
/e/	13,39	12,45	14,12	0,48	3,55 %
/i/	7,49	6,61	9,39	0,64	8,51 %
/o/	9,57	8,74	10,46	0,45	4,71 %
/u/	3,21	2,35	3,55	0,26	8,20 %
Tot. voc.	47,17	45,84	47,99	0,52	1,11 %
Total	100,00				

TABLA 4
Valores medios (en %)

	FRECUENCIA	% sobre total	% sobre conson.	% sobre vocales
/b/	96.665	2,65	5,02	
/B/	3.080	0,08	0,16	
/c/	9.877	0,27	0,51	
/d/	171.799	4,72	8,92	
/D/	9.168	0,25	0,48	
/f/	24.913	0,68	1,29	
/g/	31.770	0,87	1,65	
/G/	8.194	0,22	0,43	
/j/	7.812	0,21	0,41	
/k/	138.681	3,81	7,20	
/l/	186.365	5,12	9,68	
/m/	93.199	2,56	4,84	
/n/	87.038	2,39	4,52	
/N/	185.636	5,10	9,64	
/ɲ/	6.977	0,19	0,36	
/p/	94.238	2,59	4,89	
/r̄/	26.413	0,73	1,37	
/r/	133.132	3,66	6,91	
/R/	76.939	2,11	3,99	
/s/	275.021	7,55	14,28	
/t/	156.835	4,31	8,14	
/x/	26.767	0,73	1,39	
/ʎ/	13.793	0,38	0,72	
/θ/	61.603	1,69	3,20	
Tot. cons.	1.925.915	52,88	100,00	
/a/	490.128	13,46		28,56
/e/	490.085	13,46		28,56
/i/	273.390	7,51		15,93
/o/	347.688	9,55		20,26
/u/	114.709	3,15		6,68
Tot. voc.	1.716.000	47,12		100,00
Totales	3.641.915	100,00		

TABLA 5
Resultados generales

	Porcentajes	% acumul.
/e/	13,46	13,46
/a/	13,46	26,92
/o/	9,55	36,47
/s/	7,55	44,02
/i/	7,51	51,53
/l/	5,12	56,65
/N/	5,10	61,75
/d/	4,72	66,47
/t/	4,31	70,78
/k/	3,81	74,59
/r/	3,66	78,25
/u/	3,15	81,40
/b/	2,65	84,05
/p/	2,59	86,64
/m/	2,56	89,20
/n/	2,39	91,59
/R/	2,11	93,70
/θ/	1,69	95,39
/g/	0,87	96,26
/r̄/	0,73	96,99
/x/	0,73	97,72
/f/	0,68	98,40
/ʎ/	0,38	98,78
/c/	0,27	99,05
/D/	0,25	99,30
/G/	0,22	99,52
/j/	0,21	99,73
/ɲ/	0,19	99,92
/B/	0,08	100,00
Total	100,00	

TABLA 6
Fonemas por orden descendente de frecuencia

	BILAB.	LABIOD.	DENT.	ALV.	PALAT.	VELARES	TOTAL
OCLUS. ORALES	10,07		17,54			9,28	36,89
OCLUS. NASALES	4,84			14,16	0,36		19,36
FRICATIVAS		1,29	3,20	14,28	0,41	1,39	20,57
AFRICADAS					0,51		0,51
LATERALES				9,68	0,72		10,40
VIBRANTES				12,27			12,27
TOTALES	14,91	1,29	20,74	50,39	2,00	10,67	100,00

TABLA 7
Frecuencia de los distintos
tipos de fonemas

	FRECUENCIAS					Limites al 99%	
	Rojo	Quilis-Esgueva	Total	%	Error estándar	Mínimo	Máximo
/b/	96.665	3.805	100.470	2,64	0,000082	2,62	2,66
/B/	3.080	62	3.142	0,08	0,000014	0,08	0,09
/c/	9.877	593	10.470	0,28	0,000026	0,27	0,28
/d/	171.799	6.785	178.584	4,70	0,000108	4,67	4,73
/D/	9.168	508	9.676	0,25	0,000025	0,25	0,26
/f/	24.913	895	25.808	0,68	0,000042	0,67	0,69
/g/	31.770	1.516	33.286	0,88	0,000047	0,86	0,89
/G/	8.194	462	8.656	0,23	0,000024	0,22	0,23
/j/	7.812	660	8.472	0,22	0,000024	0,22	0,23
/k/	138.681	6.383	145.064	3,82	0,000098	3,79	3,84
/l/	186.365	6.778	193.143	5,08	0,000112	5,05	5,11
/m/	93.199	4.901	98.100	2,58	0,000081	2,56	2,60
/n/	87.038	4.455	91.493	2,41	0,000078	2,39	2,43
/N/	185.636	7.784	193.420	5,09	0,000112	5,06	5,12
/ʃ/	6.977	403	7.380	0,19	0,000022	0,19	0,20
/p/	94.238	4.443	98.681	2,60	0,000081	2,57	2,62
/r̄/	26.413	699	27.112	0,71	0,000043	0,70	0,72
/r/	133.132	5.226	138.358	3,64	0,000096	3,61	3,66
/R/	76.939	3.093	80.032	2,11	0,000073	2,09	2,12
/s/	275.021	13.325	288.346	7,58	0,000135	7,55	7,62
/t/	156.835	7.256	164.091	4,32	0,000104	4,29	4,34
/x/	26.767	919	27.686	0,73	0,000043	0,72	0,74
/ʎ/	13.793	614	14.407	0,38	0,000031	0,37	0,39
/θ/	61.603	2.330	63.933	1,68	0,000065	1,66	1,70
Tot. cons.	1.925.915	83.895	2.009.810	52,86	0,000256	52,80	52,93
/a/	490.128	19.509	509.637	13,40	0,000174	13,36	13,45
/e/	490.085	23.476	513.561	13,51	0,000175	13,46	13,55
/i/	273.390	11.810	285.200	7,50	0,000135	7,47	7,54
/o/	347.688	15.976	363.664	9,57	0,000150	9,53	9,60
/u/	114.709	5.334	120.043	3,16	0,000089	3,13	3,18
Tot. voc.	1.716.000	76.105	1.792.105	47,14	0,000256	47,07	47,20
Total	3.641.915	160.000	3.801.915	100,00			

TABLA 8

Frecuencia y porcentajes de ambas muestras.
Estimación de valores mínimo y máximo en la población.